

Profitable Double-Spending Attacks

Jehyuk Jang and Heung-No Lee, *Senior Member, IEEE*

Abstract—Our aim in this paper is to investigate the profitability of double-spending (DS) attacks that manipulate a priori mined transaction in a blockchain. Up to date, it was understood that the requirement for successful DS attacks is to occupy a higher proportion of computing power than a target network’s proportion; i.e., to occupy more than 51% proportion of computing power. On the contrary, we show that DS attacks using less than 50% proportion of computing power can also be vulnerable. Namely, DS attacks using any proportion of computing power can occur as long as the chance to making a good profit is there; i.e., revenue of an attack is greater than the cost of launching it. We have novel probability theory based derivations for calculating time finite attack probability. This can be used to size up the resource needed to calculate the revenue and the cost. The results enable us to derive sufficient and necessary conditions on the value of a target transaction which make DS attacks for any proportion of computing power profitable. They can also be used to assess the risk of one’s transaction by checking whether or not the transaction value satisfies the conditions for profitable DS attacks. Two examples are provided in which we evaluate the attack resources and the conditions for profitable DS attacks given 35% proportion of computing power against *Syscoin* and *BitcoinCash* networks, and quantitatively shown how vulnerable they are.

Index Terms—Blockchain, Bitcoin, Double-Spending Attack, Profit, Gambler’s Ruin Theorem, Poisson Counting Process.

I. INTRODUCTION

A blockchain is a distributed ledger which has originated from a desire to find a novel alternative to centralized ledgers such as transactions through third parties [1]. Besides the role as a ledger, a blockchain has been applied to many areas, e.g., managing the access authority to shared data in the cloud network [2] and averting collusion in e-Auction [3]. In a blockchain network based on the proof-of-work (PoW) mechanism, each peer node who ever has downloaded and installed the pertinent full blockchain protocol suite can join as a full node for the network. Full nodes, or the so-called miners, verify transactions, put them into a block, and mold the block to a chain by solving a cryptographic puzzle. Specifically, a transaction is put into a block by a single full node which solves the cryptographic puzzle for the first time among all full nodes in competition. The reward of minting a certain amount of coins and paid to its own address is given to the first puzzle solver as motivation to join and remain in the network. As a result, transactions are verified by many decentralized full nodes in the network. A number of other researchers [4], [5], [6] have analyzed the winning of rewards under various specific assumptions using game theory.

A consensus mechanism is programmed for decentralized

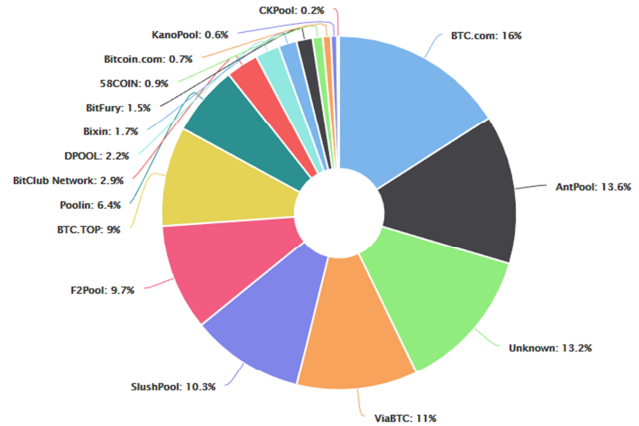


Fig. 1. Computation power distribution among the largest mining pools provided by *blockchain.com* (date accessed: 22 Oct. 2018).

peers in a network to share a common chain. If a full node succeeds in generating a new block, he/she has the latest version of the chain. All of the nodes in the network continuously communicate with each other to share the latest chain. If a node suffers from a conflict between two or more different chains, the consensus rule provides a rule that a single chain is selected. Satoshi Nakamoto suggested the *longest chain consensus* for *Bitcoin* protocol which conserves the longest chain among the conflicts [1]. There are also other consensus rules [7], e.g., GHOST [8].

Blockchains are motivated by the trust enabled by decentralized nodes. However, the decentralization mechanism is unfortunately prone to break down [9]. The PoW race is for a full node game of solving a cryptographic puzzle faster than others. As such, a node may form a pool of computing chips to increase the chance to win the PoW race. The problem is that a very limited number of pools occupy a major proportion of the computing power which operates the network. For example, the pie chart shown in Fig. 1 illustrates the proportion of computing power in the *Bitcoin* network as of October 2018. In the chart, five pools such as *BTC.com*, *AntPool*, *ViaBTC*, *F2Pool*, and *BTC.TOP* occupy a dominant proportion of the computing power. That is to say, they have recentralized the *Bitcoin* network [10].

Double-spending (DS) is one type of attacks made easily probable in a recentralized network. Since a few full nodes can easily occupy a sufficient proportion of computing power of the blockchain network, they are able to manipulate already confirmed transactions. Suppose that a public chain contains a *target transaction* which transfers the ownership of a certain amount of cryptocurrency from the attacker to a merchant for the price of a certain goods and service. Before shipping the goods, a careful merchant will wait until the transaction has been verified in a number of block confirmations by normal peers. We call this process *block confirmation*. At the same time, an attacker with a high computing power confidentially develops a fraudulent chain

[†]The authors are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Rep. of Korea. The asterisk * indicates the corresponding author. The e-mail addresses of authors are (jjh2014@gist.ac.kr, heungno@gist.ac.kr).

aimed at nullifying the target transaction in the public chain. After obtaining the block confirmation and making the fraudulent chain longer than the public one, the attacker then publicly announces the fraudulent chain. The consensus rule is to trust the longer chain, so the normal miners accept the fraudulent chain and discard the shorter public chain. Indeed, there have been a number of reports that cryptocurrencies such as *BitcoinGold*, *ZenCash*, *Zcash*, and *Litecoin Cash* suffered from DS attacks and millions of US dollars were lost in 2018 [11], [12], [13].

Recentralization is not the only concern for DS attacks. The advent of rental services which lend computing equipment for DS attacks can be a major concern as well [14]. Recently, rental services such as *nicehash.com* which provide a brokerage service between the suppliers and the consumers have indeed become available. The concern at hand is then to determine whether or not attacking with a rented computing power really returns a profit. The next concern is to find a strategy for such an attack.

Success by making DS attacks is possible but is believed to be difficult for a public blockchain with a large pool of mining network support. Nakamoto and Rosenfield provided probabilistic results of DS attack success (AS) in [1] and [15], respectively, using gambler's ruin analysis. They showed that the condition guaranteeing for making a successful DS attack is for the attacker to bring in a computing power more than the computing power which is already invested to operate the network; such an attack is thus called 51% attack. This result has been considered as the requirement for AS. This conclusion however shall be reconsidered given our result in the sequel that there are significant chances of making a good profit from DS attacks regardless of the proportion of computing power.

In this paper, our aim is to include profitability and find the requirements for DS attacks to be profitable. In our model, a DS attack succeeds if three conditions are achieved: *i*) block confirmation should be realized; *ii*) the fraudulent chain should be longer than the public chain; and *iii*) both conditions *i*) and *ii*) should be satisfied within a cut time.

A. Contributions

We show that attackers can expect a profitable DS attack not only in the super-50% proportion regime but also in the sub-50% proportion regime where computing power invested by the attacker is smaller than that invested by a target network. A DS attack is profitable if and only if the expectation of a profit function defined in (38) is positive.

To define a profit function, we introduce a novel set of mathematical tools. Specifically, we compute the probability distribution of the time spent for an AS. This AS time incorporates the probability of all possible AS within a *cut time*. The derivation of probability distribution enables us to draw results on expected revenue. Also, the expectation of AS time is used to compute expected expense spent for an attack attempt. As a result, the profit is the difference between the expected revenue and the expected expense.

We show that for a DS attack in the sub-50% proportion regime to be profitable, it is necessary to set the *cut time* to be finite. Otherwise, if an AS never be achieved, infinite deficit can happen. Under any finite *cut time*, we provide a condition on the value of target transaction which suffices a profitable DS attack.

Using these results, we provide examples of resources required for profitable attacks against *BitcoinCash* and *Syscoin*, as of December 2018 (see Section IV-B for details). Suppose that 35% proportion of computing powers is available, and the block confirmation number is 5. To compute the expected expense, we referred to the rental fee of computing power from *nicehash.com*. In the case of *Syscoin*, the expected expense is 1.810 BTC and the required value of the target transaction is 13.134 BTC. The expected AS time is around 9 minutes. In the case of *BitcoinCash*, the expected expense is 2.844 BTC and the required value of the target transaction is 20.639 BTC. The expected AS time is 1 hour 31 minutes.

B. Related Works

References [15] and [16] have analyzed the profitability of DS attacks in terms of revenue and opportunity cost. Opportunity cost is the expected rewards that could be paid out from normal mining and is generally a function of the time spent for an attack attempt. However, Rosenfield assumed the attack time to be a fixed number for the simple calculation of opportunity cost [15], while to simplify the estimation of attack time, Bissias *et al.* included an assumption that the attack stops if either the normal peers or the attacker achieves the block confirmation first [16]. On the contrary, in our model, an attack can be continued indefinitely if it brings a profit, even if the normal peers achieve block confirmation before the attacker does.

Budish conducted simulations on the profitability of DS attacks using more than 50% proportion of computing power [17]. He provided an empirical condition on the value of the target transaction that makes DS attacks not profitable. On the contrary, we consider not only the super-50% proportion regime but also the sub-50% proportion regime. We provide mathematical formulas for the required resources as functions of the computing power and block confirmation number. We also provide practical examples of profitable DS attacks against working blockchain networks.

The web-site *Crypto51.app* lists hourly rental fees for 50% proportion of computing power for the purpose of estimating the profit from DS attacks. However, there is no estimation of the AS time, and thus the estimation of the total cost is absent.

The probability distribution of AS time was analyzed in [18] and [19]. However, none of the results matched with our three conditions for AS. Specifically, neither analysis considered the first condition: *i*) block confirmation should be realized. We compare these results with ours in Section III-D in detail.

C. Organization of the Paper

Section II contains definitions of the three conditions required for a successful DS attack. DS attacks are modeled by the random walk of two independent Poisson counting processes (PCPs). Section III comprises the computation of the probabilities of DS AS and the stochastic behaviors of the first time when the DS attack is successful. In Section IV, we analyze the profitability of DS attacks, followed by providing the resources required to make them profitable. Finally, Section V concludes the paper with a summary.

II. THE ATTACK MODEL

Here, we define the conditions for a successful DS attack. DS attacks are modeled with two independent PCPs. The PCP events are carefully enumerated to account for the AS.

A. Attack Scenario

We consider blockchain networks which adopt the longest chain consensus. The longest one wins among all of the chains in competition. We assume there are two groups of miners, the normal group of miners and a single attacker. The normal group tends the *public chain*.

When the attacker decides to launch a DS attack, he/she issues a target transaction for the payment of goods or services to transfer the ownership of the cryptocurrency from the attacker him/herself to the victimized counterpart (VC). However, the attacker does not announce the target transaction to the normal group immediately but waits for a new block generation in the public chain. We denote the time at which this new block is generated as $t = 0$. At time $t = 0$, the attacker announces the target transaction to normal group so that normal group starts to put it into the public chain. At the same time, the attacker makes a fork of the public chain which stems from the newest block generated at $t = 0$ and builds it in secret. We refer to this secret fork as *fraudulent chain*. In the fraudulent chain, the target transaction is altered in a way that deceives the counterpart and benefits the attacker; one such an example is to get rid of any record of the target transaction after receiving the goods or services.

Before shipping goods or providing services to the attacker, the VC obviously chooses to wait for a few more blocks in addition to the block on which the target transaction has been entered. The number of blocks the VC chooses to wait for is referred to as the *block confirmation number* $N_{BC} \in \mathbb{Z}^+$ in this paper. Note that the number N_{BC} includes the block on which the target transaction is entered.

The attacker chooses to make the fraudulent chain public if his attack was successful. An attack is successful if the fraudulent chain is longer than the public chain after the moment the block confirmation is satisfied. This is possible because the public chain is always publicly open, while the fraudulent one is kept private by the attacker. However, the attacker will not wait for his success indefinitely since growing the attacker's chain incurs the expense per time spent for operating the computing power. The attack thus stops if the attack does not succeed within a cut time t_{cut} to cut loss.

To sum up, the AS of the DS attack is declared if all of the following conditions $\mathcal{G}^{(1)}$, $\mathcal{G}^{(2)}$, and $\mathcal{G}^{(3)}$ are satisfied.

Definition 1. A DS attack succeeds if

1. $\mathcal{G}^{(1)}$: the length of public chain counting from the moment $t = 0$ grows greater than or equal to N_{BC} ,
2. $\mathcal{G}^{(2)}$: the length of fraudulent chain counting from the moment $t = 0$ grows longer than the public chain, and
3. $\mathcal{G}^{(3)}$: starting from $t = 0$, the duration $T^{(1),(2)}$ at which both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are satisfied for the first time is less than t_{cut} .

When the cut time of attack is set to infinite $t_{cut} = \infty$, such an attack success is called attack success with the infinite cut time (AS-ICT).

B. The Stochastic Model

We model the length of the public chain and that of the fraudulent chain by two independent PCPs $H(t)$ with parameter λ_H (blocks per second) and $A(t)$ with parameter λ_A , respectively. Both processes start at the zero state $H(0) = A(0) = 0$ and independently increase by at most 1 at a time. An increment of 1 in the counting process occurs when the pertinent network adds a new block to its chain and the chain length is grown by 1 unit with each new mining success.

We rewrite the events AS and AS-ICT in terms of $H(t)$ and $A(t)$. In Definition 1, the first two conditions $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are expressed by $H(t) \geq N_{BC}$ and $A(t) > H(t)$, respectively. It is convenient to define the time $T^{(1),(2)}$ at which both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are satisfied first as follow:

$$T^{(1),(2)} := \inf \{ \{ t \in (0, \infty) : H(t) \geq N_{BC} ; A(t) > H(t) \} \cup \{ \infty \} \}. \quad (1)$$

From the last condition $\mathcal{G}^{(3)}$, the event AS-ICT is declared if $T^{(1),(2)} < \infty$. Similarly, for a finite $t_{cut} < \infty$, the event AS is declared if $T^{(1),(2)} < t_{cut}$.

To simplify $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, we form a random walk that represents the difference between $A(t)$ and $H(t)$. For this purpose, we first define two continuous stochastic processes $M(t)$ and $S(t)$, which are respectively defined as

$$M(t) := H(t) + A(t), \quad (2)$$

and

$$S(t) := H(t) - A(t). \quad (3)$$

The first process $M(t)$ is also a PCP [20] with the rate

$$\lambda_T := \lambda_A + \lambda_H \text{ (blocks per second)}, \quad (4)$$

which corresponds to the sum of computing powers of the two mining groups. The second process $S(t)$ is the continuous-time analog of a random walk. We define a random walk $S_i \in \mathbb{Z}$ for $i \in \mathbb{Z}^+$ as

$$S_i := S(T_i), \quad (5)$$

where T_i is the state progression time of S_i defined by

$$T_i := \inf \{ t \in \mathbb{R}^+ : M(t) = i \}. \quad (6)$$

Random walk S_i is a stationary Markov chain starting from $S_0 = 0$. The state transition probabilities from S_{i-1} to S_i equals the probabilities that a point arrival of $M(t)$ comes

from either $H(t)$ or $A(t)$. Specifically, the state transition probabilities are written as

$$p_A := \Pr(S_i = n+1 | S_{i-1} = n) = \frac{\lambda_A}{\lambda_T}, \quad (7)$$

and

$$p_H := \Pr(S_i = n-1 | S_{i-1} = n) = \frac{\lambda_H}{\lambda_T}, \quad (8)$$

for all $i \in \mathbb{Z}^+$ and $n \in \mathbb{Z}$. The state transition probabilities p_H and p_A are the proportions of computing power occupied by the normal miners and that by the attacker, respectively.

We define the independent and identically distributed (i.i.d.) state transition random variables $\Delta_i \in \{\pm 1\} \sim \text{Bernoulli}(p_H)$ as

$$\Delta_i := S_i - S_{i-1}, \quad (9)$$

for $i \in \mathbb{Z}^+$.

Using the random walk, we can rewrite $T^{(1)(2)}$ as

$$T^{(1)(2)} = \min\left\{\left\{T_i : H(T_i) \geq N_{BC}; S_i < 0, \forall i \in \mathbb{Z}^+\right\} \cup \{\infty\}\right\}. \quad (10)$$

C. Event sets of random walk

We aim to construct the event sets of state transitions Δ_i which imply the satisfaction of the two conditions in (10): $H(T_i) \geq N_{BC}$ (i.e., $\mathcal{G}^{(1)}$) and $S_i < 0$ (i.e., $\mathcal{G}^{(2)}$).

For the purpose, we define a DS attack as random experiment $\Delta_I = \mathcal{A}(p_A, t_{cut}; N_{BC})$ which produces random binary sequence $\Delta_I := (\Delta_1, \dots, \Delta_I) \in \{\pm 1\}^I$ of the state transitions in (9) with random length $I \in \mathbb{Z}^+$. The experimental output is an element of universal set \mathcal{U} of sequences, which is defined as

$$\mathcal{U} := \bigcup_{i=1}^{\infty} \mathcal{U}_i = \bigcup_{i=1}^{\infty} \{\pm 1\}^i, \quad (10)$$

where $\mathcal{U}_i := \{\pm 1\}^i$. We define $\delta_i := (\delta_1, \dots, \delta_i) \in \mathcal{U}_i$ as a binary sequence of length i , which is the realization of Δ_I .

We denote $s_k := \sum_{i=1}^k \delta_i$ for integer $k \in \mathbb{Z}^+$, which comprises observations of the state variables S_k of the random walk.

We denote the event sets $\mathcal{W}_i \subset \mathcal{U}_i$, for $i \in \mathbb{Z}^+$, each of which satisfies $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ at the i -th state for the first time. The time $T^{(1)(2)}$ in (10) then can be rewritten as

$$T^{(1)(2)} = \min\left\{\left\{T_i : \Delta_I \in \mathcal{W}_i, \forall i \in \mathbb{Z}^+\right\} \cup \{\infty\}\right\}. \quad (11)$$

To construct \mathcal{W}_i , we divide it into mutually exclusive sets $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_{j,i}^{(2)-(1)}$, for $j = 1, \dots, i$.

Set $\mathcal{D}_j^{(1)}$ is a subset of \mathcal{U}_i such that $\mathcal{G}^{(1)}$ is satisfied exactly at the j -th state S_j . One of the requirements on the

binary sequences of $\mathcal{D}_j^{(1)}$ is $s_j = 2N_{BC} - j$ since the first j transitions δ_k for $k = 1, \dots, j$ have N_{BC} number of $+1$'s and $j - N_{BC}$ number of -1 's.

Set $\mathcal{D}_{j,i}^{(2)-(1)}$ for $j \leq i$ is a subset of \mathcal{U}_i such that $\mathcal{G}^{(2)}$ is satisfied for the first time at the i -th state and $\mathcal{G}^{(1)}$ is satisfied already at state j prior or equal to state i . This set does not care about the first j transitions δ_k for $k = 1, \dots, j$, but only focuses on the interim transitions from the $j+1$ -th to the i -th, i.e. δ_m , for $m = j+1, \dots, i$. Recall that satisfying $\mathcal{G}^{(1)}$ at state j implies $s_j = 2N_{BC} - j$. Thus, the requirement for the elements of $\mathcal{D}_{j,i}^{(2)-(1)}$ is that the state changes from starting state $s_j = 2N_{BC} - j$ to state $s_i = -1$, while any interim state s_k remains non-negative; i.e., $s_k \geq 0$ for each $k = j+1, \dots, i-1$.

The elements of joint set $\mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)}$ for $j \leq i$ satisfy both $\mathcal{G}^{(1)}$ at state j and $\mathcal{G}^{(2)}$ at state i . When $j > i$, the elements of $\mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)}$ does not imply achieving $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ at the i -th state, since no confirmation has been obtained yet. Namely, achieving $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ is possible at a state only posterior to the state at which $\mathcal{G}^{(1)}$ is satisfied. When $j < N_{BC}$, $\mathcal{D}_j^{(1)} = \emptyset$ due to an insufficient number of state transitions for the block confirmation. Subsequently, \mathcal{W}_i is written as

$$\mathcal{W}_i = \bigcup_{j=N_{BC}}^i \mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)}. \quad (12)$$

We further explore (12). Remember that in the first j transitions of $\mathcal{D}_j^{(1)}$, the number of $+1$'s is fixed to N_{BC} and the rests of $j - N_{BC}$ transitions are -1 's. This implies that for $j > 2N_{BC}$, s_j in $\mathcal{D}_j^{(1)}$ are already negative. Equivalently, for $j > 2N_{BC}$, elements in $\mathcal{D}_j^{(1)}$ satisfy both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ at state j . Analogously, $\mathcal{D}_{j,i}^{(2)-(1)} = \emptyset$ for $j > 2N_{BC}$ and $j < i$, since the state $s_j = 2N_{BC} - j$ contradicts the requirement of $\mathcal{D}_{j,i}^{(2)-(1)}$: the interim transactions between s_j and s_i should be non-negative. For $j > 2N_{BC}$ and $j = i$, set $\mathcal{D}_{j,i}^{(2)-(1)}$ about interim states s_k , for $k = j+1, \dots, i-1$, equals set \mathcal{U}_i since there is no interim state to apply the requirement to. This implies $\mathcal{D}_j^{(1)} \cap (\mathcal{D}_{j,i}^{(2)-(1)}) = \mathcal{D}_i^{(1)}$ for $j > 2N_{BC}$ and $j = i$.

As a result, (12) becomes

$$\mathcal{W}_i = \begin{cases} \left(\bigcup_{j=N_{BC}}^{2N_{BC}} \mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)} \right) \cup \mathcal{D}_i^{(1)}, & \text{for } i > 2N_{BC}, \\ \phi, & \text{for } i \leq 2N_{BC}. \end{cases} \quad (13)$$

For example, suppose $N_{BC} = 2$, then a sequence $\Delta_5 = (-1, +1, +1, -1, -1)$ satisfies $\mathcal{G}^{(1)}$ at state index $j = 3$. After the 3-rd index, Δ_5 satisfies $\mathcal{G}^{(2)}$ at $i = 5$, thus $\Delta_5 \in (\mathcal{D}_3^{(1)} \cap \mathcal{D}_{3,5}^{(2)-(1)}) \subset \mathcal{W}_5$. The other example is a sequence $\Delta_5 = (-1, -1, +1, -1, +1)$, which satisfies $\mathcal{G}^{(1)}$ at $j = 5$ for the first time. In addition, at the same state index, the sequence Δ_5 satisfies $\mathcal{G}^{(2)}$ as well. Hence, $\Delta_5 \in \mathcal{D}_5^{(1)} \subset \mathcal{W}_5$. It is easy to check that for all $j > 2N_{BC}$, the sequences which satisfy $\mathcal{G}^{(1)}$ at the j -th state index for the first time satisfy $\mathcal{G}^{(2)}$ as well at the same state index, and thus are the elements of \mathcal{W}_j . As a counterexample, a sequence $\Delta_4 = (-1, -1, +1, +1)$ satisfies $\mathcal{G}^{(1)}$ at the 4-th state index, but never satisfies $\mathcal{G}^{(2)}$ due to the number of state transitions being insufficient.

The sets \mathcal{W}_i for $i \in \mathbb{Z}^+$ are mutually exclusive since the lengths of the sequences comprising these differ by i . Thus, for DS attack $\Delta_I = \mathcal{A}(p_A, t_{cut}; N_{BC})$, if i exists such that $\Delta_I \in \mathcal{W}_i$, for $i \in \mathbb{Z}^+$, then it is unique, which implies that the expression for $T^{(1),(2)}$ in (11) can be rewritten as

$$T^{(1),(2)} = \begin{cases} T_i, & \text{if } \exists i: \Delta_I \in \mathcal{W}_i, \forall i \in \mathbb{Z}^+, \\ \infty, & \text{otherwise.} \end{cases} \quad (14)$$

III. AS PROBABILITIES

For a DS attack task $\Delta_I = \mathcal{A}(p_A, t_{cut}; N_{BC})$, we aim to compute the probability of AS, which equals the probability that the AS conditions $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ have met within the time duration t_{cut} ; i.e., $T^{(1),(2)} < t_{cut}$. This probability requires the probability density function (PDF) of $T^{(1),(2)}$, which also enables to compute the expectation of the time at which a DS attack succeeds, i.e., expected AS time.

The probabilities and expectations in this section will be used to evaluate the profitability of DS attacks in Section IV.

A. AS-ICT Probability

We first compute the probability of AS-ICT with $t_{cut} = \infty$. The probability of AS-ICT is the probability that the state index i exists such that $\Delta_I \in \mathcal{W}_i$, and thus requires $\Pr(\Delta_I \in \mathcal{W}_i)$. Note that no occurrence of AS-ICT with infinite t_{cut} implies no occurrence of AS with a finite t_{cut} as well. That is to say, the probability of AS-ICT is also needed to compute the probability of AS.

In specific, from the mutual exclusiveness of \mathcal{W}_i for $i \in \mathbb{Z}^+$, the probability \mathbb{P}_{AS-ICT} of AS-ICT equals the sum of $\Pr(\Delta_I \in \mathcal{W}_i)$, for $i \in \mathbb{Z}^+$. Since $\mathcal{W}_i = \emptyset$, for $i \leq 2N_{BC}$, as given in (13), it can be computed as

$$\mathbb{P}_{AS-ICT}(p_A; N_{BC}) = \sum_{i=2N_{BC}+1}^{\infty} \Pr(\Delta_I \in \mathcal{W}_i). \quad (15)$$

The following Proposition 2 gives the probability $\Pr(\Delta_I \in \mathcal{W}_i)$ used in (15).

Proposition 2. Consider DS attack task $\Delta_I = \mathcal{A}(p_A, t_{cut}; N_{BC})$, then the probability that $\Pr(\Delta_I \in \mathcal{W}_i)$, for $i > 2N_{BC}$, can be computed as

$$\Pr(\Delta_I \in \mathcal{W}_i) = \sum_{j=N_{BC}}^{j=2N_{BC}} \binom{j-1}{N_{BC}-1} C_{\frac{i-1}{2}-N_{BC}, 2N_{BC}-j} P_A^{\frac{i+1}{2}} P_H^{\frac{i-1}{2}} + \binom{i-1}{N_{BC}-1} P_H^{N_{BC}} P_A^{i-N_{BC}}, \quad (16)$$

where

$$C_{n,m} = \begin{cases} \frac{m+1}{n+m+1} \binom{2n+m}{n}, & n, m \in \mathbb{Z}^+ \cup \{0\}, \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

and for $i \leq 2N_{BC}$, $\Pr(\Delta_I \in \mathcal{W}_i) = 0$.

Proof: As given in (13), set \mathcal{W}_i is the union of $\mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)}$, for $j = N_{BC}, \dots, 2N_{BC}$ and $\mathcal{D}_i^{(1)}$. As sets $\mathcal{D}_j^{(1)}$, for $j = N_{BC}, \dots, 2N_{BC}$, are mutually exclusive by definition, the probability of the union of $\mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)}$, for $j = N_{BC}, \dots, 2N_{BC}$, and $\mathcal{D}_i^{(1)}$ equals the sum of the respective probabilities. In addition, for every $i \geq 2N_{BC} + 1$ and j , $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_{j,i}^{(2)-(1)}$ are independent since the requirements for the two sets focus on the different indices of the state transitions. Thus, the probability of intersections $\mathcal{D}_j^{(1)} \cap \mathcal{D}_{j,i}^{(2)-(1)}$ equals the product of the respective probabilities. As a result, from (13), $\Pr(\Delta_I \in \mathcal{W}_i)$ for $i > 2N_{BC}$ can be computed as

$$\Pr(\Delta_I \in \mathcal{W}_i) = \sum_{j=N_{BC}}^{j=2N_{BC}} \Pr(\Delta_I \in \mathcal{D}_j^{(1)}) \Pr(\Delta_I \in \mathcal{D}_{j,i}^{(2)-(1)}) + \Pr(\Delta_I \in \mathcal{D}_i^{(1)}). \quad (18)$$

By definition, set $\mathcal{D}_j^{(1)}$ picks $N_{BC} - 1$ transitions among the first $j - 1$ transitions. The picked transitions are given $+1$ s and the rests are given -1 s. The j -th transition is $\Delta_j = 1$. The probability $\Pr(\mathcal{D}_j^{(1)})$ equals the point mass function of a negative binomial distribution:

$$\Pr(\mathcal{D}_j^{(1)}) = \binom{j-1}{N_{BC}-1} P_H^{N_{BC}} P_A^{j-N_{BC}}. \quad (19)$$

Computing the probability $\Pr(\mathcal{D}_{j,i}^{(2)-(1)})$ starts from counting the number of elements in $\mathcal{D}_{j,i}^{(2)-(1)}$. Remember the requirements on every element of $\mathcal{D}_{j,i}^{(2)-(1)}$, for $j = N_{BC}, \dots, 2N_{BC}$, are that the states change starting from state $s_j = 2N_{BC} - j$ and ending with state $s_i < 0$ while keeping $s_k \geq 0$, for $k = j + 1, \dots, i - 1$. The i -th transition should be $\Delta_i = -1$. The number of elements in $\mathcal{D}_{j,i}^{(2)-(1)}$ equals the ballot number [21], which is the number of random walks that consist of $i - j - 1$ steps and never become negative,

starting from point $2N_{BC}-j$ at the j -th state and ending at the origin with the $i-1$ -th state. This number is given as $C_{n,m}$ in (17) with relationships $2n+m=i-j-1$ and $m=2N_{BC}-j$. As a result, by multiplying the probabilities of state transitions, the probability $\Pr(\mathcal{D}_{j,i}^{(2)-(1)})$ is computed as

$$\Pr(\mathcal{D}_{j,i}^{(2)-(1)}) = C_{n,m} p_A^{(n+m+1)} p_H^n. \quad (20)$$

Finally, substituting (19) and (20) into (18) results in (16). \blacksquare

The following Corollary 3 gives an explicit formula of the probability \mathbb{P}_{AS-ICT} of AS-ICT given in (15).

Corollary 3. *The probability \mathbb{P}_{AS-ICT} has an algebraic expression*

$$\mathbb{P}_{AS-ICT}(p_A; N_{BC}) = \begin{cases} 1, & p_H \leq p_A, \\ 1 - p_A^{N_{BC}+1} p_H^{N_{BC}} \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} A_j, & p_H > p_A, \end{cases} \quad (21)$$

where

$$A_j \triangleq p_A^{j-2N_{BC}-1} - p_H^{j-2N_{BC}-1}. \quad (22)$$

Proof: See Appendix A

B. AS Probability

By Definition 1, the probability of AS equals

$$\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC}) := \Pr(T^{(1),(2)} < t_{cut}). \quad (23)$$

To compute \mathbb{P}_{AS} , we need the probability density function (PDF) of $T^{(1),(2)}$. From the mutual exclusiveness of \mathcal{W}_i for integers $i > 2N_{BC}$ and the relationship in (14), the PDF $f_{T^{(1),(2)}}$ can be computed as

$$f_{T^{(1),(2)}}(t) = \sum_{i=2N_{BC}+1}^{\infty} \Pr(\Delta_i \in \mathcal{W}_i) f_{T_i}(t) + (1 - \mathbb{P}_{AS-ICT}) \delta(t - \infty), \quad (24)$$

where $\delta(t)$ is a Dirac delta function and $f_{T_i}(t)$ is the PDF of T_i . The random variable T_i in (6) follows an Erlang distribution with shape parameter i and rate λ_T [20]. The PDF of T_i is thus given as

$$f_{T_i}(t) = \frac{\lambda_T (\lambda_T t)^{i-1} e^{-\lambda_T t}}{(i-1)!}. \quad (25)$$

Proposition 4. *The PDF of time $T^{(1),(2)}$ has a closed-form expression:*

$$f_{T^{(1),(2)}}(t) = \frac{p_A \lambda_T e^{-\lambda_T t} (p_A p_H (\lambda_T t)^2)^{N_{BC}}}{(2N_{BC})!} \cdot \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} {}_2F_3(\mathbf{a}; \mathbf{b}; p_A p_H (\lambda_T t)^2) + \frac{e^{-\lambda_T t} (p_H \lambda_T t)^{N_{BC}}}{t (N_{BC}-1)!} \left(e^{p_A \lambda_T t} - \sum_{i=0}^{N_{BC}} \frac{(p_A \lambda_T t)^i}{i!} \right) + (1 - \mathbb{P}_{AS-ICT}) \delta(t - \infty), \quad (26)$$

where ${}_pF_q(\mathbf{a}; \mathbf{b}; x)$ is the generalized hypergeometric function [22] defined in Appendix E with the parameter vectors

$$\mathbf{a} = \begin{bmatrix} N_{BC} + 1 - j/2 \\ N_{BC} + 1/2 - j/2 \end{bmatrix} \quad (27)$$

and

$$\mathbf{b} = \begin{bmatrix} 2N_{BC} + 2 - j \\ N_{BC} + 1 \\ N_{BC} + 1/2 \end{bmatrix}. \quad (28)$$

Proof: See Appendix B.

C. Expected AS Times

It will be shown to be convenient to define the AS time as

$$T_{AS} := \begin{cases} T^{(1),(2)}, & \text{if } T^{(1),(2)} < t_{cut}, \\ \text{not defined}, & \text{otherwise.} \end{cases} \quad (29)$$

The case for $T_{AS} > t_{cut}$ does not need to be defined since it is not useful.

The PDF of T_{AS} is just a scaled version of $f_{T^{(1),(2)}}(t)$ for $0 < t < t_{cut}$, which is given in (26), with a scaling factor of \mathbb{P}_{AS}^{-1} . Formally, the PDF $f_{T_{AS}}(t)$ equals

$$f_{T_{AS}}(t) = \begin{cases} \frac{f_{T^{(1),(2)}}(t)}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})}, & \text{for } 0 \leq t < t_{cut}, \\ 0, & \text{for } t \geq t_{cut}. \end{cases} \quad (30)$$

The expectation of AS time (EAST) is computed as

$$\mathbb{E}_{T_{AS}}(p_A, t_{cut}; N_{BC}) = \frac{\int_0^{t_{cut}} t f_{T^{(1),(2)}}(t) dt}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})}. \quad (31)$$

Similarly, we define AS-ICT time as the AS time with $t_{cut} = \infty$. From the results (24) and (31), if $t_{cut} = \infty$, the expectation of the time for AS-ICT is computed as follow

$$\begin{aligned}
\mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC}) &= \frac{\lim_{t_{cut} \rightarrow \infty} \int_0^{t_{cut}} t f_{T^{(1),(2)}}(t) dt}{\mathbb{P}_{AS-ICT}(p_A; N_{BC})} \\
&= \frac{\sum_{i=2N_{BC}+1}^{\infty} \mathbb{E}[T_i] \Pr(\Delta_i \in \mathcal{W}_i)}{\mathbb{P}_{AS-ICT}(p_A; N_{BC})} \\
&= \frac{\sum_{i=2N_{BC}+1}^{\infty} \frac{i}{\lambda_T} \Pr(\Delta_i \in \mathcal{W}_i)}{\mathbb{P}_{AS-ICT}(p_A; N_{BC})},
\end{aligned} \tag{32}$$

where $E[T_i] = i\lambda_T^{-1}$ [20].

The following Proposition 5 gives an explicit formula of

$\mathbb{E}_{T_{AS-ICT}}$.

Proposition 5. *Let $p_M := \max(p_A, p_H)$ and $p_m := \min(p_A, p_H)$, then the expectation $\mathbb{E}_{T_{AS-ICT}}$ has a closed-form expression:*

$$\mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC}) = \frac{\lambda_T^{-1} \left(\sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} Z_j + \frac{N_{BC}}{p_H} \right)}{\mathbb{P}_{AS-ICT}(p_A; N_{BC})}, \tag{33}$$

where

$$\begin{aligned}
Z_j := & p_A p_m^{N_{BC}} p_M^{-(N_{BC}-j+1)} \left(\frac{2N_{BC}-2j p_m + 1}{p_M - p_m} \right) \\
& - j p_A^{-(N_{BC}-j)} p_H^{N_{BC}}.
\end{aligned} \tag{34}$$

Proof: See Appendix A.

D. Comparison with Previous Works

The AS-ICT probability \mathbb{P}_{AS-ICT} (AS probability when indefinite cut time $t_{cut} = \infty$ is given) in Corollary 3 was computed by Nakamoto [1] and Rosenfield [15] using the gambler's ruin theorem [23]. In [1], Nakamoto suggested an additional assumption not in our scenario: the time spent for the first N_{BC} blocks mined by the normal group is not random and is determined as the average time $\lambda_H^{-1} N_{BC}$ instead. In other words, the block confirmation process was not treated as the stochastic processes. In [15], Rosenfield removed the assumption proposed by Nakamoto to derive the result in Corollary 3. However, the result was still based on the gambler's ruin theorem which only computes the asymptotical behavior of S_n as $n \rightarrow \infty$ by manipulating the recurrence relationship between two adjacent states. That is to say, he assumed that an indefinite number of attack chances are given to the attacker. There was no result related to the intermediate process such as Proposition 2.

In this paper, we introduce t_{cut} , which generalize the results by Nakamoto and Rosenfield, and compute the AS probability \mathbb{P}_{AS} using Proposition 2. In practice, attack chances are limited since the amount of resources such as time and cost are limited, and therefore a cut time t_{cut} is needed to cut loss.

Besides the probability $\mathbb{P}_{T_{AS-ICT}}$, the probability distribution of attack success time similar with $T^{(1),(2)}$ was

also analyzed in [16], [18], and [19]. However, none of the results matched with the AS conditions in Definition 1.

In [18], Goffard considered the race between two PCPs $H(t)$ and $A(t)$ with unfair initial states. Specifically, the initial states of the public chain $H(t)$ and the fraudulent chain were set to $H(0)=z$ and $A(0)=0$ for integer $z>0$, respectively, then an implicit expression of the probability distribution of first time τ_z at which $H(\tau_z)=A(\tau_z)$ was analyzed. Time τ_z can be interpreted as the interval spent for $\mathcal{D}_{j,i}^{(2)-(1)}$; i.e., the interval from the time at which $\mathcal{G}^{(1)}$ alone is satisfied to the time at which both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are satisfied. This analysis did not consider the time for $\mathcal{G}^{(1)}$.

In [16], Bissias *et al.* also considered the race between two PCPs. To derive an explicit formula of the probability distribution of AS time, they put in other conditions to end an attack attempt: the attack stops if either $H(t)$ or $A(t)$ reaches $N_{BC}+1$, whichever happens first. In other words, the only way to succeed in an attack is that the fraudulent chain should mine $N_{BC}+1$ blocks faster than the public chain. On the contrary, in our model and in reality, an attack can be continued even at a moment when the public chain is ahead of the fraudulent chain, if it will give any profit.

In [19], Karame *et al.* analyzed the first AS time under a fast-payment model which removed the block confirmation process by omitting condition $\mathcal{G}^{(1)}$.

IV. THE EXPECTED PROFIT OF A DS ATTACK

The previous probabilistic analyses in [1] and [15] show that the success of DS attacks is not guaranteed when $p_A < 0.5$. However, DS attacks with $p_A < 0.5$ might be pursued if they bring profit.

A. Profitable DS Attacks

Here, we analyze the *profitability* of DS attacks and to this end, we define profit function P of DS attack $\mathcal{A}(C, p_A, t_{cut}; N_{BC})$ in terms of value C of a fraudulent transaction, the block mining reward, and the operating expense (OPEX) of the computing power. We compute the expected profit function $\mathbb{E}_P(p_A, t_{cut}; N_{BC})$, which is the expectation of P .

Definition 6 (Profitable Attacks). *DS attack $\mathcal{A}(C, p_A, t_{cut}; N_{BC})$ is said to be profitable if and only if $\mathbb{E}_P > 0$.*

The OPEX (e.g. the rental fee for the computing power) and the block mining reward are increased by the average block mining speed λ_A by the attacker and the time t consumed during an attack. Thus, the OPEX and block mining rewards are expressed as functions $X(\lambda_A, t)$ and $R(\lambda_A, t)$, respectively, which can be any increasing function (e.g., linear, exponential, or log) with respect to λ_A and t . We define X and R , respectively, as follows:

$$X(\lambda_A, t) := \gamma \lambda_A t (\log_{x_1} x_2)^{\lambda_A} (\log_{x_3} x_4)^t \tag{35}$$

for real constants $\gamma > 0$, $x_1, x_2 > 1$, and $x_3, x_4 > 1$, and

$$R(\lambda_A, t) := \beta \lambda_A t (\log_{r_1} r_2)^{\lambda_A} (\log_{r_3} r_4)^t \quad (36)$$

for real constants β , $r_1, r_2 > 1$, and $r_3, r_4 > 1$.

To sum up, if an attack succeeds, the income from the AS is C as it is double-spent and the block mining reward R for time duration T_{AS} . In this case, the cost is the OPEX for duration T_{AS} . If the attack fails, the cost is the OPEX for duration t_{cut} without revenue. Hence, profit P of a DS attack is the random variable

$$P := \begin{cases} C + R(\lambda_A, T_{AS}) - X(\lambda_A, T_{AS}), & \text{if } T^{(1),(2)} < t_{cut}, \\ -X(\lambda_A, t_{cut}), & \text{otherwise.} \end{cases} \quad (37)$$

Subsequently, the expected profit function of a DS attack is

$$\begin{aligned} \mathbb{E}_P(p_A, t_{cut}; N_{BC}) &= \\ & \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC}) (C + \mathbb{E}[R(\lambda_A, T_{AS})] - \mathbb{E}[X(\lambda_A, T_{AS})]) \\ & - (1 - \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})) X(\lambda_A, t_{cut}) \\ & = \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC}) (C + \mathbb{E}[R(\lambda_A, T_{AS})]) - \mathbb{E}_X(p_A, t_{cut}; N_{BC}), \end{aligned} \quad (38)$$

where $\lambda_A = p_A \lambda_T$ and \mathbb{E}_X is the expected OPEX defined as

$$\begin{aligned} \mathbb{E}_X(p_A, t_{cut}; N_{BC}) &:= \\ & \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC}) \mathbb{E}[X(\lambda_A, T_{AS})] \\ & + (1 - \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})) X(\lambda_A, t_{cut}). \end{aligned} \quad (39)$$

Definition 6 and (38) imply that for fixed p_A and t_{cut} , DS attack $A(C, p_A, t_{cut}; N_{BC})$ is profitable if and only if $C > C_{Req}$, where the required value of target transaction C_{Req} is

$$C_{Req} = \frac{\mathbb{E}_X(p_A, t_{cut}; N_{BC})}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})} - \mathbb{E}[R(\lambda_A, T_{AS})]. \quad (40)$$

Theorem 7. Suppose $x_1 = x_2$ and $x_3 = x_4$ in (35), and

$r_1 = r_2$ and $r_3 = r_4$ in (36). DS attack $A(C, p_A, t_{cut}; N_{BC})$ for $p_A \in (0, 0.5)$ is profitable only if $t_{cut} < \infty$. In addition, let $t_{cut} = \infty$ and $p_A \in (0.5, 1)$, then the required value of target transaction in (40) becomes

$$C_{Req} = \max(0, (\gamma - \beta) \lambda_T p_A \mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC})). \quad (41)$$

Proof: See Appendix C.

By Theorem 7, setting $t_{cut} = \infty$ and $p_A \in (0, 0.5)$ makes DS attack non-profitable. The following Theorem 8 shows that by setting a finite t_{cut} , DS attacks $A(C, p_A, t_{cut}; N_{BC})$ can be profitable not only for $p_A \in (0.5, 1)$, but also for $p_A \in (0, 0.5)$.

Theorem 8. Let $x_1 = x_2$ and $x_3 = x_4$ in (35), and $r_1 = r_2$ and $r_3 = r_4$ in (36). Let $t_{cut} = c \mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC})$ for a positive real c , where $\mathbb{E}_{T_{AS-ICT}}$ is given in (33). A DS attack $A(C, p_A, t_{cut}; N_{BC})$ is profitable for any $p_A \in (0, 1)$, if $C > C_{Suf}$, where

$$C_{Suf} = \gamma'(p_A, c) \frac{\lambda_T p_A \mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC})}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})}, \quad (42)$$

and

$$\gamma'(p_A, c) := \gamma \cdot \left(\frac{\mathbb{P}_{AS-ICT}(p_A; N_{BC})}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})} + c(1 - \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})) \right). \quad (43)$$

Proof: See Appendix D.

B. Profitable DS Attacks against Working Blockchain Networks

As of 9th December 2018, we refer to blockchain explorers and *nicehash.com* (who provides the rental rates for borrowing computing power) to obtain block mining reward R and OPEX X . The parameters x_1, \dots, x_4 used in (35) are set to $x_1 = x_2$ and $x_3 = x_4$, which leads to a linear function

TABLE I
NUMERICAL COMPUTATIONS OF REQUIRED RESOURCES FOR PROFITABLE DS ATTACKS WHEN $t_{cut} = c \mathbb{E}_{T_{AS-ICT}}$, FOR $c=2$.

Block Confirmation Number (N_{BC})	1		3		5		7		9		
Computing Power (p_A)	0.35	0.4	0.35	0.4	0.35	0.4	0.35	0.4	0.35	0.4	
Cut Time (t_{cut})	Scaled by λ_H^{-1}	7.593	9.491	13.548	15.617	19.217	21.294	24.753	26.755	30.212	32.084
AS Probability (\mathbb{P}_{AS})		0.389	0.523	0.286	0.440	0.217	0.380	0.167	0.332	0.130	0.292
Expected AS Time ($\mathbb{E}_{T_{AS}}$)	Scaled by λ_H^{-1}	2.640	2.801	5.682	5.732	8.621	8.553	11.498	11.316	14.333	14.039
Expected OPEX (\mathbb{E}_X)		3.050	3.993	6.085	7.513	9.110	10.973	12.135	14.423	15.153	17.874
Required Value of Target Transaction (C_{Suf})	Scaled by γ	12.257	12.535	31.967	26.253	62.545	43.457	107.812	64.728	172.757	90.660

X with respect to λ_A and T_{AS} . Analogously, the parameters for R in (36) are set to $r_1 = r_2$ and $r_3 = r_4$, leading to a linear function R with respect to λ_A and T_{AS} . There are three more parameters: γ , β , and λ_H^{-1} . Parameter γ is the expected cost spent per generating a block and required for computing the expected OPEX. Parameter β is the reward per generating a block. Parameter λ_H^{-1} is the average block generation time of the public chain. They are different by blockchain networks.

We consider the *Syscoin* and *BitcoinCash* networks. The parameter γ is obtainable from *nicehash.com*. The two networks use the *SHA-256* cryptographic puzzle for which the unit of computation is *hash*. The rental fee for 1P hashes per second for a day is around 0.04 BTC, which is around $4.63 \cdot 10^{-7}$ BTC per second. In other words, the rental fee is approximately $4.63 \cdot 10^{-22}$ BTC per the computing of a hash.

Once parameters β , γ , and λ_H^{-1} are obtained, the required attack resources can be evaluated using Table I. Table I lists the required attack resources for each $p_A = 0.35$ and $p_A = 0.4$ when $t_{cut} = c \mathbb{E}_{T_{AS-ICT}}$, for $c = 2$.

1) The Syscoin Network Parameters

The average block generation time is fixed at $\lambda_H^{-1} = 60$ seconds. Referring to *poolexplorer.com*, the network's computing speed is 7.6E hashes per second; i.e., $7.6E \cdot 60 = 456E$ solutions are needed to mine one block on average. Then, the parameter γ is obtained as

$$\begin{aligned} \gamma &= 4.63 \cdot 10^{-22} \text{ [BTC/hash]} \\ &\quad \times 456E \text{ [hashes/block mining]} \\ &\approx 0.21 \text{ [BTC/block mining]}. \end{aligned} \quad (44)$$

The reward β per block mining is 38.5 SYS (without transaction fees), which is around $3.6 \cdot 10^{-4}$ BTC per block mining.

2) The BitcoinCash Network Parameters

The average block generation time is fixed at $\lambda_H^{-1} = 600$ seconds. Referring to *BTC.com*, the network's computing speed is 1.2E hashes per second; i.e., $1.2E \cdot 600 = 720E$ hashes are needed to generate one block on average. The parameter γ is obtained as

$$\begin{aligned} \gamma &= 4.63 \cdot 10^{-22} \text{ [BTC/hash]} \\ &\quad \times 720E \text{ [hashes/block mining]} \\ &\approx 0.33 \text{ [BTC/block mining]}. \end{aligned} \quad (45)$$

Table II

EVALUATION OF RESOURCES REQUIRED FOR PROFITABLE ATTACK AGAINST WORKING BLOCKCHAIN NETWORKS ($p_A = 0.35$, $N_{BC} = 5$).

Target Network	<i>Syscoin</i>	<i>BitcoinCash</i>
Cut Time (Seconds)	1153	11530
Required Value of Target Transaction (BTC)	13.134	20.639
Expected OPEX (BTC)	1.810	2.844
Expected AS Time (Seconds)	546	5466

The reward β per block mining is 12.5 BCH (without transaction fees), which is around 0.57 BTC per block mining. By Theorem 7, the relationship $\beta > \gamma$ implies that the required value $C_{Req} = 0$ for DS attack $\mathcal{A}(C, p_A, t_{cut}; N_{BC})$ with $p_A > 0.5$ and $t_{cut} = \infty$ to be profitable is 0; i.e., such DS attacks are always profitable regardless of the value C of target transaction.

3) DS Attack with a Finite Cut Time and $p_A < 0.5$

In Table II, we evaluate the resources required for profitable DS attacks using $p_A = 0.35$ against the two blockchain networks. The values in Table II are obtained from the values in Table I multiplied by scaling parameters γ and λ_H^{-1} . The results explain the importance of network parameter λ_H^{-1} . Remember that *Syscoin* has a greater network computing power (7.6E hashes per second) than *BitcoinCash* (1.2E hashes per second). This implies that *Syscoin* has a higher rental fee per unit time for a same proportion of computing power than *BitcoinCash*. Specifically, when $p_A = 0.35$, the rental fee for *Syscoin* is 163.69 BTC per day whereas that for *BitcoinCash* is 25.84 BTC per day. Nevertheless, *BitcoinCash* requires higher OPEX for profitable DS attacks than *Syscoin*, since the higher λ_H^{-1} of *BitcoinCash* implies the higher γ (the average cost per block mining). In addition, a high λ_H^{-1} proportionally delays the expected AS time.

V. CONCLUSIONS

We showed that DS attacks using 50% or a less proportion of computing power can be profitable. For both the super-50% and the sub-50% proportion regimes, we provided quantitative resources required for profitable DS attacks. Specifically, we provided the probability for an AS success as well as the operating time and expense of mining rigs. We summarized the results in Table I, which enable the easy calculation of the minimum resources required for a profitable attack against any blockchain network. We showed examples of the calculations against working networks.

Our results quantitatively show the importance of network policy. The less the average block mining period and block confirmation number, the less the minimum resources required for a profitable attack. That is to say, blockchain networks pursuing fast transaction speeds are risky. A way for developers of such networks to discourage DS attacks is, for example, to restrict the value of transactions depending on the network policy. If the value of the target transaction is limited below the minimum quantity we provided, attackers cannot expect to make a profit.

REFERENCES

- [1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System." [Online]. Available: <https://bitcoin.org/bitcoin.pdf>.
- [2] H. Ritzdorf, C. Soriente, G. O. Karame, S. Marinovic, D. Gruber, and S. Capkun, "Toward Shared Ownership in the Cloud," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 12, pp. 3019–3034, Dec. 2018.
- [3] S. Wu, Y. Chen, Q. Wang, M. Li, C. Wang, and X. Luo, "CREAM: A Smart Contract Enabled Collusion-Resistant e-Auction," *IEEE Trans. Inf. Forensics Secur.*, 2018.
- [4] I. Tsabary and I. Eyal, "The Gap Game," in *Proceedings of the 2018 ACM SIGSAC conference on Computer and Communications Security*, 2018, pp. 713–728.

- [5] J. A. Kroll, I. C. Davey, and E. W. Felten, "The Economics of Bitcoin Mining, or Bitcoin in the Presence of Adversaries," presented at the The Twelfth Workshop on Economics of Information Security (WEIS 2013), Washington, DC, 2013.
- [6] B. Biais, C. Bisiere, M. Bouvard, and C. Casamatta, "The Blockchain Folk Theorem," *TSE Work. Pap.*, Jan. 2018.
- [7] G.-T. Nguyen and K. Kim, "A Survey about Consensus Algorithms Used in Blockchain," *J. Inf. Process. Syst.*, vol. 14, no. 1, pp. 101–128, 2018.
- [8] Y. Sompolinsky and A. Zohar, "Secure High-Rate Transaction Processing in Bitcoin," presented at the Financial Cryptography and Data Security, Puerto Rico, 2015, pp. 507–527.
- [9] A. Beikverdi and J. Song, "Trend of centralization in Bitcoin's distributed network," presented at the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Takamatsu, Japan, 2015.
- [10] E. Homakov, "Stop. Calling. Bitcoin. Decentralized.," *Egor Homakov*, 03-Dec-2017. [Online]. Available: <https://medium.com/@homakov/stop-calling-bitcoin-decentralized-cb703d69dc27>.
- [11] C. Osborne, "Bitcoin Gold suffers double spend attacks, \$17.5 million lost," *ZDNet*, 25-May-2018. [Online]. Available: <https://www.zdnet.com/article/bitcoin-gold-hit-with-double-spend-attacks-18-million-lost/>.
- [12] "ZenCash Statement on Double Spend Attack," *Horizen*, 03-Jun-2018. [Online]. Available: <https://blog.zencash.com/zencash-statement-on-double-spend-attack/>.
- [13] A. Hertig, "Blockchain's Once-Fearful 51% Attack Is Now Becoming Regular," *CoinDesk*, 08-Jun-2018. [Online]. Available: <https://www.coindesk.com/blockchains-feared-51-attack-now-becoming-regular/>.
- [14] J. Bonneau, "Why buy when you can rent? Bribery attacks on Bitcoin consensus," presented at the The 3rd Workshop on Bitcoin and Blockchain Research (BITCOIN '16), Barbados, 2016.
- [15] M. Rosenfeld, "Analysis of Hashrate-Based Double Spending," *ArXiv14022009 Cs*, Feb. 2014.
- [16] G. Bissias, B. N. Levine, A. P. Ozisik, and G. Andresen, "An Analysis of Attacks on Blockchain Consensus," *ArXiv161007985 Cs*, Oct. 2016.
- [17] E. Budish, "The Economic Limits of Bitcoin and the Blockchain," Jun-2018. [Online]. Available: <http://www.nber.org/papers/w24717>.
- [18] P.-O. Goffard, "Fraud risk assessment within blockchain transactions," 2018. [Online]. Available: http://pierre-olivier.goffard.me/Publications/FraudRiskAssessmentWithinBlockchainTransaction_Goffard0218.pdf.
- [19] G. O. Karame, E. Androulaki, M. Roeschlin, A. Gervais, and S. Čapkun, "Misbehavior in Bitcoin: A Study of Double-Spending and Accountability," *ACM Trans Inf Syst Secur*, vol. 18, no. 1, pp. 2:1–2:32, May 2015.
- [20] A. Papoulis and S. U. Pillai, "Random walks and other applications," in *Probability, Random Variables and Stochastic Processes*, 4th edition., Boston, Mass.: McGraw-Hill Europe, 2002.
- [21] P. Flajolet and R. Sedgewick, "Combinatorial structures and ordinary generating functions," in *Analytic Combinatorics*, Cambridge University Press, 2009.
- [22] G. Gasper and M. Rahman, "Basic Hypergeometric series," in *Basic hypergeometric series*, Second., vol. 96, Cambridge University Press, Cambridge, 2004.
- [23] W. Feller, "Random walk and ruin problems," in *An introduction to probability theory and its applications*, New York: Wiley, 1968.
- [24] H. S. Wilf, "Analytic and asymptotic methods," in *generatingfunctionology: Third Edition*, 3 edition., Wellesley, Mass: A K Peters/CRC Press, 2005.
- [25] H. S. Wilf, "Introductory ideas and examples," in *generatingfunctionology: Third Edition*, 3 edition., Wellesley, Mass: A K Peters/CRC Press, 2005.
- [26] P. Flajolet and R. Sedgewick, "Labelled structures and exponential generating functions," in *Analytic Combinatorics*, Cambridge University Press, 2009.

APPENDIX A

PROOF OF COROLLARY 3 AND PROPOSITION 5

A. Proof of Corollary 3

We reduce the infinite summations in (15) into an algebraic form using generating functions.

By substituting (16) into (15), the probability \mathbb{P}_{AS-ICT} becomes

$$\mathbb{P}_{AS-ICT} = \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A \sum_{i=2N_{BC}+1}^{\infty} C_{i-1}^{N_{BC}, 2N_{BC}-j} (p_A p_H)^{\frac{i-1}{2}} + \left(\frac{p_H}{p_A}\right)^{N_{BC}} \sum_{i=2N_{BC}+1}^{\infty} \binom{i-1}{N_{BC}-1} p_A^i. \quad (46)$$

By rearranging the indices i in the summations, we can obtain

$$\mathbb{P}_{AS-ICT} = \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A \sum_{i=0}^{\infty} C_{i, 2N_{BC}-j} (p_A p_H)^{i+N_{BC}} + \left(\frac{p_H}{p_A}\right)^{N_{BC}} \left(\sum_{i=N_{BC}}^{\infty} \binom{i-1}{N_{BC}-1} p_A^i - \sum_{i=N_{BC}}^{2N_{BC}} \binom{i-1}{N_{BC}-1} p_A^i \right). \quad (47)$$

We define two generating functions as

$$M_k(x) := \sum_{i=0}^{\infty} C_{i,k} x^i, \quad (48)$$

and

$$G_k(x) := \sum_{i=k}^{\infty} \binom{i}{k} x^i. \quad (49)$$

By substituting M_k and G_k into (47), we can write

$$\mathbb{P}_{AS-ICT} = \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A (p_A p_H)^{N_{BC}} M_{2N_{BC}-j}(p_A p_H) + \left(\frac{p_H}{p_A}\right)^{N_{BC}} \left(p_A G_{N_{BC}-1}(p_A) - \sum_{i=N_{BC}}^{2N_{BC}} \binom{i-1}{N_{BC}-1} p_A^i \right). \quad (50)$$

The function $M_k(x)$ is a generating function of the ballot numbers $C_{i,k}$, for which the algebraic expression given in [24] is

$$M_k(x) = \left(\frac{2}{1 + \sqrt{1-4x}} \right)^{k+1}. \quad (51)$$

Putting $x = p_A p_H$ into $M_k(x)$ results in

$$M_k(p_A p_H) = \begin{cases} \left(\frac{2}{1 + \sqrt{1-4p_A p_H}} \right)^{k+1} \\ \left(\frac{2}{1 + \sqrt{1-4(1-p_A)p_A}} \right)^{k+1}, & p_A < p_H, \\ \left(\frac{2}{1 + \sqrt{1-4(1-p_H)p_H}} \right)^{k+1}, & p_A \geq p_H \end{cases} = \left(\frac{1}{p_M} \right)^{k+1}, \quad (52)$$

where $p_M := \max(p_H, p_A)$.

$G_k(x)$ is a generating function of binomial coefficients, and the algebraic expression for it is given in [25]:

$$G_k(x) = \frac{x^k}{(1-x)^{k+1}}. \quad (53)$$

Putting $x = p_A$ into $G_k(x)$ results in

$$G_k(p_A) = p_H^{-1} \left(\frac{p_A}{p_H} \right)^k. \quad (54)$$

Substituting (52) and (54) into (50) arrives at

$$\mathbb{P}_{AS-ICT} = \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A (p_A p_H)^{N_{BC}} p_M^{-(2N_{BC}-j+1)} + 1 - \left(\frac{p_H}{p_A}\right)^{N_{BC}} \sum_{i=N_{BC}}^{2N_{BC}} \binom{i-1}{N_{BC}-1} p_A^i. \quad (55)$$

We define $p_m := \min(p_A, p_H)$, then the relationship $p_A p_H = p_m p_M$ holds. By rearranging the order of operands, we can obtain

$$\mathbb{P}_{AS-ICT} = 1 - \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} \left(\left(\frac{p_H}{p_A}\right)^{N_{BC}} p_A^j - \frac{p_A}{p_M} \left(\frac{p_m}{p_M}\right)^{N_{BC}} p_M^j \right), \quad (56)$$

which is equal to (21). \blacksquare

B. Proof of Proposition 5

We use the generating functions and their derivatives to compute the infinite summations in (32).

By substituting (16) into (32) and rearranging the order of operands, we obtain

$$\begin{aligned} \lambda_T \mathbb{E}_{T_{AS-ICT}} &= \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} \sum_{i=2N_{BC}}^{\infty} (i+1) C_{i-2N_{BC}, 2N_{BC}-j} p_A^{\frac{i+2}{2}} p_H^{\frac{i}{2}} \\ &+ \sum_{i=N_{BC}-1}^{2N_{BC}-1} (i+1) \binom{i}{N_{BC}-1} p_A^{i+1-N_{BC}} p_H^{N_{BC}} \\ &- \sum_{i=N_{BC}-1}^{2N_{BC}-1} (i+1) \binom{i}{N_{BC}-1} p_A^{i+1-N_{BC}} p_H^{N_{BC}}. \end{aligned} \quad (57)$$

By rearranging the indices of summations, we arrive at

$$\begin{aligned} \lambda_T \mathbb{E}_{T_{AS-ICT}} &= \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A^{N_{BC}+1} p_H^{N_{BC}} \\ &\cdot \sum_{i=0}^{\infty} (2i+2N_{BC}+1) C_{i, 2N_{BC}-j} (p_A p_H)^i \\ &+ p_A \left(\frac{p_H}{p_A}\right)^{N_{BC}} \sum_{i=N_{BC}-1}^{\infty} (i+1) \binom{i}{N_{BC}-1} p_A^i \\ &- \sum_{i=N_{BC}}^{2N_{BC}} i \binom{i-1}{N_{BC}-1} p_A^{i-N_{BC}} p_H^{N_{BC}}. \end{aligned} \quad (58)$$

By substituting generating functions $M_k(x)$ and $G_k(x)$ defined respectively in (48) and (49), (58) becomes

$$\begin{aligned} \lambda_T \mathbb{E}_{T_{AS-ICT}} &= \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A^{N_{BC}+1} p_H^{N_{BC}} \\ &\cdot \left(2 \sum_{i=0}^{\infty} i C_{i, 2N_{BC}-j} (p_A p_H)^i \right. \\ &+ (2N_{BC}+1) M_{2N_{BC}-j}(p_A p_H) \\ &+ p_A \left(\frac{p_H}{p_A}\right)^{N_{BC}} \left(\sum_{i=N_{BC}-1}^{\infty} i \binom{i}{N_{BC}-1} p_A^i + G_{N_{BC}-1}(p_A) \right) \\ &\left. - \sum_{i=N_{BC}}^{2N_{BC}} i \binom{i-1}{N_{BC}-1} p_A^{i-N_{BC}} p_H^{N_{BC}} \right) \end{aligned} \quad (59)$$

We use the following relationships,

$$\sum_{i=0}^{\infty} i C_{i,k} x^i = x M'_k(x) \quad (60)$$

and

$$\sum_{i=k}^{\infty} i \binom{i}{k} x^i = x G'_k(x), \quad (61)$$

and their derivatives are given by

$$\begin{aligned} M'_k(x) &:= \frac{d}{dx} M_k(x) \\ &= \sum_{i=0}^{\infty} i C_{i,k} x^{i-1} \\ &= \frac{(k+1)}{\sqrt{1-4x}} \left(\frac{2}{1+\sqrt{1-4x}} \right)^{k+2} \end{aligned} \quad (62)$$

and

$$\begin{aligned} G'_k(x) &:= \frac{d}{dx} G_k(x) \\ &= \sum_{i=k}^{\infty} i \binom{i}{k} x^{i-1} \\ &= \frac{(kx^{k-1} + x^k)}{(1-x)^{k+2}}. \end{aligned} \quad (63)$$

By substituting (60) and (61) into (59), we obtain

$$\begin{aligned} \lambda_T \mathbb{E}_{T_{AS-ICT}} &= \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} p_A^{N_{BC}+1} p_H^{N_{BC}} \\ &\cdot \left(2 p_A p_H M'_{2N_{BC}-j}(p_A p_H) + (2N_{BC}+1) M_{2N_{BC}-j}(p_A p_H) \right) \\ &+ p_A \left(\frac{p_H}{p_A}\right)^{N_{BC}} \left(p_A G'_{N_{BC}-1}(p_A) + G_{N_{BC}-1}(p_A) \right) \\ &- \sum_{i=N_{BC}}^{2N_{BC}} i \binom{i-1}{N_{BC}-1} p_A^{i-N_{BC}} p_H^{N_{BC}} \end{aligned} \quad (64)$$

Putting $x = p_A p_H$ into $M'_k(x)$ results in

$$\begin{aligned} M'_k(p_A p_H) &= M'_k(p_m p_M) \\ &= \frac{(k+1)}{1-2p_m} \left(\frac{1}{p_M} \right)^{k+2}. \end{aligned} \quad (65)$$

Putting $x = p_A$ into $G'_k(x)$ gives

$$G'_k(p_A) = \frac{(k p_A^{k-1} + p_A^k)}{p_H^{k+2}}. \quad (66)$$

By substituting (52), (54), (65), and (66) into (64), we finally obtain (33). \blacksquare

APPENDIX B PROOF OF PROPOSITION 4

We use a generating function and generalized hypergeometric functions to compute the infinite summations in (24).

By substituting $\Pr(\Delta_i \in \mathcal{W}_i)$ in (16) and $f_{T_i}(t)$ in (25) into (24), we arrive at

$$\begin{aligned} f_{T^{(0,2)}}(t) - (1 - \mathbb{P}_{AS-ICT}) \delta(t - \infty) &= \\ &= \sum_{j=N_{BC}}^{2N_{BC}} \binom{j-1}{N_{BC}-1} \sum_{i=2N_{BC}+1}^{\infty} C_{i-1-N_{BC}, 2N_{BC}-j} p_A^{\frac{i+1}{2}} p_H^{\frac{i-1}{2}} \frac{\lambda_T i^{i-1} e^{-\lambda_T t}}{(i-1)!} \\ &+ \sum_{i=2N_{BC}+1}^{\infty} \binom{i-1}{N_{BC}-1} p_H^{N_{BC}} p_A^{i-N_{BC}} \frac{\lambda_T i^{i-1} e^{-\lambda_T t}}{(i-1)!}. \end{aligned} \quad (67)$$

By rearranging the indices of summations and the order of operands, we obtain

$$\begin{aligned}
f_{T^{(0,2)}}(t) - (1 - \mathbb{P}_{AS-ICT})\delta(t - \infty) = & \\
& \sum_{j=N_{BC}}^{j=2N_{BC}} \binom{j-1}{N_{BC}-1} \sum_{i=0}^{\infty} \left(D_{i,2N_{BC}-j} P_A^{N_{BC}+i+1} P_H^{N_{BC}+i} \right. \\
& \quad \left. \cdot \frac{\lambda_T^{2N_{BC}+2i+1} t^{2N_{BC}+2i} e^{-\lambda_T t}}{(2N_{BC}+2i)!} \right) \\
& + \left(\frac{P_H}{P_A} \right)^{N_{BC}} e^{-\lambda_T t} \left(\sum_{i=N_{BC}}^{\infty} \binom{i-1}{N_{BC}-1} P_A^i \frac{\lambda_T^i t^{i-1}}{(i-1)!} \right. \\
& \quad \left. - \sum_{i=N_{BC}}^{2N_{BC}} \binom{i-1}{N_{BC}-1} P_A^i \frac{\lambda_T^i t^{i-1}}{(i-1)!} \right).
\end{aligned}$$

We can define two generating functions as

$$\begin{aligned}
B(x) &:= \sum_{i=0}^{\infty} C_{i,2N_{BC}-j} \frac{x^i}{(2N_{BC}+2i)!} \\
&= (2N_{BC}-j+1) \sum_{i=0}^{\infty} \frac{(2i+2N_{BC}-j)!}{i!(i+2N_{BC}-j+1)!(2N_{BC}+2i)!} x^i,
\end{aligned} \tag{69}$$

and

$$\begin{aligned}
H(x) &:= \sum_{i=N_{BC}}^{\infty} \binom{i-1}{N_{BC}-1} \frac{x^{i-1}}{(i-1)!} \\
&= \sum_{i=N_{BC}-1}^{\infty} \binom{i}{N_{BC}-1} \frac{x^i}{i!}.
\end{aligned} \tag{70}$$

By substituting $B(x)$ and $H(x)$ into (68), we obtain

$$\begin{aligned}
f_{T^{(0,2)}}(t) - (1 - \mathbb{P}_{AS-ICT})\delta(t - \infty) = & \\
& \sum_{j=N_{BC}}^{j=2N_{BC}} \binom{j-1}{N_{BC}-1} P_A \lambda_T e^{-\lambda_T t} \left(P_A P_H (\lambda_T t)^2 \right)^{N_{BC}} B(P_A P_H (\lambda_T t)^2) \\
& + \left(\frac{P_H}{P_A} \right)^{N_{BC}} e^{-\lambda_T t} \left(P_A \lambda_T H(P_A \lambda_T t) - \sum_{i=N_{BC}}^{2N_{BC}} \binom{i-1}{N_{BC}-1} P_A^i \frac{\lambda_T^i t^{i-1}}{(i-1)!} \right).
\end{aligned} \tag{71}$$

We replace function $B(x)$ in (69) with the generalized hypergeometric functions defined in Appendix E. For this purpose, we first denote the sequences in $B(x)$ by

$$\beta_i = \frac{(2i+2N_{BC}-j)!}{i!(i+2N_{BC}-j+1)!(2N_{BC}+2i)!}, \tag{72}$$

and

$$\beta_0 = \frac{1}{(2N_{BC}-j+1)(2N_{BC})!}. \tag{73}$$

Next, the function $B(x)$ can be rewritten as

$$\begin{aligned}
B(x) &= (2N_{BC}-j+1) \sum_{i=0}^{\infty} \beta_i x^i \\
&= (2N_{BC}-j+1) \beta_0 \left(x^0 + \frac{\beta_1}{\beta_0} x^1 + \frac{\beta_2}{\beta_1} \frac{\beta_1}{\beta_0} x^2 + \dots \right).
\end{aligned} \tag{74}$$

The reformulated sequence in (74) is computed by

$$\frac{\beta_{i+1}}{\beta_i} = \frac{(i+1+N_{BC}-j/2)(i+1/2+N_{BC}-j/2)}{(i+2+2N_{BC}-j)(i+1+N_{BC})(i+1/2+N_{BC})(i+1)}, \tag{75}$$

which has 2 polynomials in i on the numerator and 3 polynomials in i on the denominator, except for $(i+1)$. By the definition of the generalized hypergeometric function [22], function $B(x)$ can be expressed as

$$\begin{aligned}
B(x) &= (2N_{BC}-j+1) \beta_0 {}_2F_3(\mathbf{a}_j; \mathbf{b}_j; x) \\
&= \frac{1}{(2N_{BC})!} {}_2F_3(\mathbf{a}_j; \mathbf{b}_j; x),
\end{aligned} \tag{76}$$

where vectors \mathbf{a}_j and \mathbf{b}_j respectively defined in (27) and (28) are the constants in the polynomials in i of the numerator and denominator in (75), respectively.

We use a closed-form expression of generating function $H(x)$ in (70) given by

$$\begin{aligned}
H(x) &= \sum_{i=N_{BC}-1}^{\infty} \binom{i}{N_{BC}-1} \frac{x^i}{i!} \\
&= \frac{1}{(N_{BC}-1)!} \sum_{i=N_{BC}-1}^{\infty} \frac{x^i}{(i-N_{BC}+1)!} \\
&= \frac{x^{N_{BC}-1}}{(N_{BC}-1)!} \sum_{i=0}^{\infty} \frac{x^i}{i!} \\
&= \frac{x^{N_{BC}-1}}{(N_{BC}-1)!} e^x,
\end{aligned} \tag{77}$$

where the following relationship is used [26]:

$$\sum_{i=0}^{\infty} \frac{x^i}{i!} = e^x. \tag{78}$$

By substituting (76) and (77) into (67), we arrive at

$$\begin{aligned}
f_{T^{(0,2)}}(t) - (1 - \mathbb{P}_{AS-ICT})\delta(t - \infty) &= \frac{P_A \lambda_T e^{-\lambda_T t} \left(P_A P_H (\lambda_T t)^2 \right)^{N_{BC}}}{(2N_{BC})!} \\
& \cdot \sum_{j=N_{BC}}^{j=2N_{BC}} \binom{j-1}{N_{BC}-1} {}_2F_3(\mathbf{a}_j; \mathbf{b}_j; P_A P_H (\lambda_T t)^2) \\
& + \left(\frac{P_H}{P_A} \right)^{N_{BC}} e^{-\lambda_T t} \left(P_A \lambda_T \frac{(P_A \lambda_T t)^{N_{BC}-1}}{(N_{BC}-1)!} e^{P_A \lambda_T t} \right. \\
& \quad \left. - \sum_{i=N_{BC}}^{2N_{BC}} \binom{i-1}{N_{BC}-1} P_A^i \frac{\lambda_T^i t^{i-1}}{(i-1)!} \right) \\
& = \frac{P_A \lambda_T e^{-\lambda_T t} \left(P_A P_H (\lambda_T t)^2 \right)^{N_{BC}}}{(2N_{BC})!} \\
& \cdot \sum_{j=N_{BC}}^{j=2N_{BC}} \binom{j-1}{N_{BC}-1} {}_2F_3(\mathbf{a}_j; \mathbf{b}_j; P_A P_H (\lambda_T t)^2) \\
& + \left(\frac{P_H}{P_A} \right)^{N_{BC}} e^{-\lambda_T t} \left(P_A \lambda_T \frac{(P_A \lambda_T t)^{N_{BC}-1}}{(N_{BC}-1)!} e^{P_A \lambda_T t} \right. \\
& \quad \left. - \frac{1}{(N_{BC}-1)!} \sum_{i=N_{BC}}^{2N_{BC}} P_A^i \frac{\lambda_T^i t^{i-1}}{(i-N_{BC})!} \right).
\end{aligned} \tag{79}$$

We obtain (26) by rearranging the indices of the summations and the order of operands. \blacksquare

APPENDIX C PROOF OF THEOREM 7

When $x_1 = x_2$, $x_3 = x_4$, $r_1 = r_2$, and $r_3 = r_4$, the OPEX and block mining reward respectively turn into

$$X(\lambda_A, T_{AS}) = \gamma \lambda_A T_{AS} \tag{80}$$

and

$$R(\lambda_A, T_{AS}) = \beta \lambda_A T_{AS}. \tag{81}$$

Combining these conditions implies that expected OPEX \mathbb{E}_X defined in (39) becomes

$$\begin{aligned}
\mathbb{E}_X(p_A, t_{cut}; N_{BC}) &= \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC}) \gamma \lambda_A \mathbb{E}_{T_{AS}}(p_A, t_{cut}; N_{BC}) \\
& + (1 - \mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})) \gamma \lambda_A t_{cut}.
\end{aligned} \tag{82}$$

For $p_A \in (0, 0.5)$, $\mathbb{P}_{AS} < 1$ always holds, since $\mathbb{P}_{AS} \leq \mathbb{P}_{AS-ICT}$ by the definition and $\mathbb{P}_{AS-ICT} < 1$ by (21). Thus, if $t_{cut} = \infty$, the expected OPEX in (82) diverges to minus infinity. In other words, the expected profit \mathbb{E}_p in (38) with $p_A \in (0, 0.5)$ is positive only if $t_{cut} < \infty$.

We next derive the sufficient and necessary condition (41) for $p_A \in (0.5, 1)$ and $t_{cut} = \infty$. When $t_{cut} = \infty$, $T_{AS} = T_{AS-ICT}$ and $\mathbb{P}_{AS} = \mathbb{P}_{AS-ICT}$ by the definition of AS-ICT. In addition, $p_A \in (0, 0.5)$ implies $\mathbb{P}_{AS-ICT} = 1$ by (21). In this case, by substituting (80) and (81), the expected profit \mathbb{E}_p in (38) becomes

$$\mathbb{E}_p(p_A, \infty; N_{BC}) = C + (\beta - \gamma) \lambda_A \mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC}). \quad (83)$$

Hence, $\mathbb{E}_p > 0$ if and only if the value C of target transaction is greater than $C_{Req.}$ given in (41). ■

APPENDIX D PROOF OF THEOREM 8

We obtain an upper bound $C_{Suf.}$ of $C_{Req.}$ in (40). If $C > C_{Suf.}$ then $C > C_{Req.}$, which implies that a DS attack $\mathcal{A}(C, p_A, c \mathbb{E}_{T_{AS-ICT}}; N_{BC})$ for a positive real number c is profitable.

As $\mathbb{E}[R(\lambda_A, T_{AS})]$ in (40) is nonnegative by the definition of function R , we arrive at the upper bound:

$$C_{Req.} \leq \frac{\mathbb{E}_X(p_A, t_{cut}; N_{BC})}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})}. \quad (84)$$

By substituting $t_{cut} = c \mathbb{E}_{T_{AS-ICT}}$, $x_1 = x_2$, and $x_3 = x_4$, the expected OPEX \mathbb{E}_X in (84) becomes

$$\begin{aligned} \mathbb{E}_X(p_A, c \mathbb{E}_{T_{AS-ICT}}; N_{BC}) &= \mathbb{P}_{AS}(p_A, c \mathbb{E}_{T_{AS-ICT}}; N_{BC}) \gamma \lambda_A \mathbb{E}_{T_{AS}} \\ &+ (1 - \mathbb{P}_{AS}(p_A, c \mathbb{E}_{T_{AS-ICT}}; N_{BC})) \gamma \lambda_A c \mathbb{E}_{T_{AS-ICT}}. \end{aligned} \quad (85)$$

We use the following relationship on the conditional expectation

$$\begin{aligned} \mathbb{E}_{T_{AS}} &= \frac{\int_0^{t_{cut}} f_{T^{(1)|(2)}}(t) dt}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})} \leq \frac{\lim_{x \rightarrow \infty} \int_0^x f_{T^{(1)|(2)}}(t) dt}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})} \\ &= \frac{\mathbb{P}_{AS-ICT}(p_A; N_{BC})}{\mathbb{P}_{AS}(p_A, t_{cut}; N_{BC})} \mathbb{E}_{T_{AS-ICT}}(p_A; N_{BC}). \end{aligned} \quad (86)$$

By substituting (85) and (86) into (84), we finally obtain the upper bound $C_{Suf.}$ given in (42). ■

APPENDIX E

THE GENERALIZED HYPERGEOMETRIC FUNCTION

A generalized hypergeometric series [22] is a power series $\sum_{n \geq 0} \beta_n z^n$, where the ratio of coefficients are expressed by polynomials $A(n)$ and $B(n)$ in n as follows:

$$\frac{\beta_{n+1}}{\beta_n} = \frac{A(n)}{B(n)(n+1)}, \quad (87)$$

for all integers $n \geq 0$. The polynomials are written by

$$A(n) = c(a_1 + n) \cdots (a_p + n) \quad (88)$$

and

$$B(n) = d(b_1 + n) \cdots (b_q + n). \quad (89)$$

The generalized hypergeometric series is denoted by

$${}_pF_q(\mathbf{a}; \mathbf{b}; z), \quad (90)$$

where \mathbf{a} and \mathbf{b} are vectors of a_1, \dots, a_p and b_1, \dots, b_q , respectively.

A generalized hypergeometric series defines a generalized hypergeometric function if it converges. If $p < q + 1$, then the ratio of coefficients (87) goes to zero as $n \rightarrow \infty$, which implies that the series converges for any finite value z and thus defines the function.

Time-Variant Proof-of-Work Using Error-Correction Codes

Sangjun Park, Haeung Choi, and Heung-No Lee, *Senior Member, IEEE*

Abstract— The protocol for cryptocurrencies is largely divided into three parts, namely consensus, wallet, and networking overlay. Consensus deals with coming to an agreement among the participating nodes to the current status of their blockchain. The status of the blockchain is updated only through valid transactions. This objective is achieved among trustless rational peer nodes. A proof-of-work (PoW) based consensus has been proven to be secure and robust owing to its simple rule and has served as a firm foundation for cryptocurrencies such as bitcoin and ethereum. However, the emergence of specialized mining devices for the existing PoW causes two problems: *i*) the re-centralization issue of the mining market and *ii*) the usage of a considerable amount of energy in mining. In this paper, we introduce a new PoW called Error-Correction Codes PoW in which the error-correction codes and their decoder can be utilized by concatenating the decoder with a hash function. In this new PoW, a puzzle can be intentionally made to change from block to block, leading to a time-variant property. This property is useful in repressing the emergence of specialized mining devices, which can be a solution to the problems that the existing PoWs face currently.

Index Terms— Consensus, Cryptocurrency, Blockchain, Proof-of-Work, Error-Correction Codes, Hash Functions

I. INTRODUCTION

In cryptocurrencies, the consensus mechanism is considered to be the most innovative part because it prevents the double spending attack [1] in a peer-to-peer network without trusted third parties. In bitcoin [2], for example, more than ten thousand of nodes randomly scattered across the world aim to reach a consensus in each block time. The Internet is the only way to connect them; communication packets are delayed and sometimes dropped though the Internet gives the best service. Also, cyberattacks frequently occur, which makes transactions over the Internet insecure. Nevertheless, bitcoin has shown secure peer-to-peer transactions over the past 10 years. This is possible with the help of proof-of-work (PoW) which is fundamental of the consensus mechanism.

In bitcoin, each node scattered in the world does competitive work, called *mining*, to forge a block. The node which wins this competition has the right to mint a specified number of coins as this mining reward. If a node was re-forging all the block alone, it could spend the total amount of works done to all the blocks when they were forged.

Without PoW, anybody with a computer can alter the content of the blockchain, which can make unauthorized changes in any mined blocks. If PoW is attached to each mined block, attackers cannot make any unauthorized modifications without redoing all the works. No one thus can alter any mined blocks, meaning an immutability property of blockchain.

In bitcoin, miners make a rational decision to maximize their profit. For this rational decision, the miners seek and extend the longest chain. To understand whether this decision is rational or not, we consider an example in which there are two chains. One chain is assumed to be longer than the other chain. The longer chain has to be adopted by miners because this has the maximum PoW, i.e., these miners select and extend the longer chain. Otherwise, the chance to get the mining reward can be probabilistically smaller when the miners select a shorter chain.

In bitcoin, for miners to get the chance, they spend computational resources to solve a puzzle which we generate using a secure hash algorithm 256 (SHA256) [3]. To solve this puzzle, the miners execute SHA256 many times by only varying nonce until they find a particular nonce. This particular nonce makes a hash of SHA256 begin with multiple zero bits. They attach the particular nonce into a mined block as a proof that the provided puzzle is correctly solved. Satoshi [2] intended for the miners to execute SHA256 using a central processing unit (CPU). But, as they competed to quickly do the mining work, there were needs to develop chips that can quickly run SHA256 rather than CPU can. In 2013, an Avalon company made the first mining device based on an application-specific integrated circuit (ASIC).

Nowadays, miners equipped with ASIC devices have begun to dominate the mining business [4]. However, they cause two problems. The first problem is that the mining markets have become re-centralized [5]. These miners have a large portion of the total hash power, implying that the blockchain is left with a handful of these miners. It can be possible for them to modify the mined blocks, meaning that the immutability property can be broken. The use of a considerable amount of electrical energy to mine blocks is the second problem [6]. New models of ASIC devices can outperform old models with respect to the hash power. Each miner buys the new models to win the mining competition. As the new models are widely used in the mining work, the total hash power inevitably increases, making a puzzle difficult to be solved. The miners have to spend more electrical energy to solve this puzzle whose difficulty gets larger as the new models are being used.

Error-correction codes are widely used in modern communication systems to combat errors occurring over noisy chan-

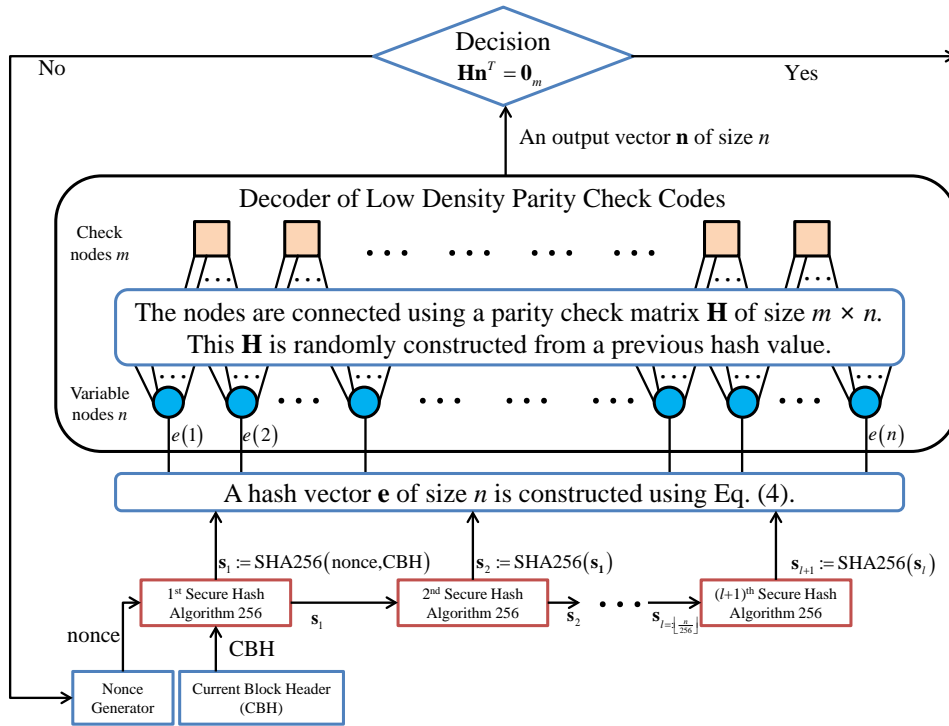


Fig. 1. A overall scheme of ECCPoW.

nels. Errors introduced over a noisy channel can be corrected by running a decoding algorithm. The codes have a plentiful history with numerous classes of codes having varying degrees of decoding complexity.

Error-correction codes have been used to make hash functions. This approach is categorized into a) the code-based one and b) the lattice-based one. We consider the code-based one because our solution for making ASIC devices useless is based on the usage of low-density parity check (LDPC) [7] codes and an error-correction decoding algorithm.

The first code-based approach is back to the late of 1970. In 1978, McEliece [8] proposed a McEliece cryptosystem where the private key consists of a generator matrix \mathbf{G} , a binary Goppa code, a t -error-correction decoding algorithm of that code, a permutation matrix \mathbf{P} and a nonsingular random matrix \mathbf{S} . The public key is then a randomly permuted generator matrix, i.e., $\mathbf{S} \times \mathbf{G} \times \mathbf{P}$. In this system, a hash of a given message is obtained as follows: a) a word is made by multiplying the given message with the public key and b) the hash is generated by adding this word to a binary random word whose number of ones is at most t . Peters et al. [9] used non-binary Goppa codes to extend the McEliece cryptosystem in 2010. By using the non-binary codes, the size of its public key was reduced. This cryptosystem with the non-binary codes could achieve the same security level as much as the McEliece cryptosystem could. Other codes such as low density generator matrix codes [10], LDPC codes [11][12], Reed-Solomon codes [13] and Reed-Muller codes [14] have been replaced with the Goppa codes in the McEliece cryptosystem, respectively. The aim for using these codes is to solve

a problem that the size of a given public key is too large, compared to that of an original message.

Aside from the applications of the error-correction codes into the McEliece cryptosystem, the codes are used to make secure hash functions. Preneel [15] proposed the construction methods of hash functions and proved that they can be collision resistant under assumptions. The codes used in [15] are either maximum distance separable (MDS) codes or hamming codes. Selman *et al.* [16] used LDPC codes to construct a hash function proven to be an average universal hash function defined in [17].

In this paper, we aim to propose a new proof-of-work system that is resistant to ASIC mining devices. This system named as error-correction codes proof-of-work (ECCPoW) can consist of a set of SHA256s and a decoder of LDPC codes. We give Fig. 1 to illustrate the structure of ECCPoW and address how it works.

First, we get n -bit hash values using SHA256s. They are used to construct a hash vector \mathbf{r} of size n . Second, we make a parity check matrix \mathbf{H} using a hash value of the previous block. Third, we input \mathbf{r} into the decoder and run a decoding algorithm to get an output \mathbf{n} of size n . The decision is made on the basis of \mathbf{n} , as we have shown in Fig. 1. Details on how to obtain \mathbf{r} and \mathbf{H} will be given in Section IV.

We define a random variable that represents the number of hash cycles required to solve a given puzzle. We obtain both the mean and the variance of this random variable. We investigate how this mean behaves in terms of either the number of miners or the length of a hash vector, leading to the following results:

- The mean value could be a decreasing function with respect to the number of miners.
- The mean value could be an increasing function with respect to the length n under assumptions given in Propo-

¹ For any Goppa code, there is a construction method which guarantees that the minimum distance d of that code is greater than a given number. Thus, the value of t can be known in advance using Theorem 1 [32].

sition 1.

Next, five properties of ECCPoW will be given in Section V. The unique property is time-variant, meaning that a puzzle can vary from block to block. This property prohibits the advent of ASIC mining devices, which can be a solution to the mentioned problems.

We organize the rest of this paper as follows. Section II gives literature surveys regarding SHAs and PoWs. Section III elucidates LDPC codes and a decoder. Section IV addresses how ECCPoW works, gives its pseudo codes, and presents the theoretical results of ECCPoW. Section V discusses the properties of ECCPoW and the reason for considering LDPC codes. Last, section VII presents the conclusions of this paper.

II. LITERATURE SURVEY ON BLOCKCHAIN

A. Secure hash standard and functions

The secure hash standard was formulated by NIST [3]. The purpose of this standard is to offer the specifications of SHAs that yield a hash of a given message. If the message is changed even slightly, the hash can come out completely different. Thus, a hash can be used to detect whether an original message was altered or not. SHAs can be used for the generation and verification of digital signatures as well as of message authentication. In addition, SHAs can be used for the generation of random numbers.

A secure hash function takes an arbitrarily sized message and yields a fixed-size hash. Let a function h be

$$h: \mathcal{X} \rightarrow \mathcal{Y}$$

which is said to be a cryptographically secure hash function if it satisfies the three requirements defined in [17].

(One-way function) Given any hash y which a corresponding message is not known, it is computationally infeasible to find a message x such that $h(x) = y$.

(Weak collision resistance) Given an arbitrary message x , it is computationally infeasible to find any message x' which has the same hashes $h(x) = h(x')$.

(Strong collision resistance) It is computationally infeasible to find any two different messages x and x' which make the same hashes $h(x) = h(x')$.

TABLE II. The routine of Ethash

Inputs: BH, L and DAG	
Step 1:	for nonce = 0, 1, 2, ... $2^{32} - 1$
Step 2:	mix0 = $f(\text{SHA3}(\text{nonce}, \text{BH}))$
Step 3:	for $i = 1, 2, \dots, 63$
Step 4:	data1 = Fetch(DAG, mix0)
Step 5:	tmp = Mixing(mix0, data1)
Step 6:	mix0 = $f(\text{tmp})$
Step 7:	end
Step 8:	If mix0 begins with L zero bits, then go to Step 10.
Step 9:	end
Step 10:	Block generation & broadcast

where f is a predefined function. Details on f is given in [18].

Based on the NIST standard [3], there are many SHAs such as SHA1, SHA224, SHA256, SHA384, and SHA512. A message of any size less than 2^{64} bits can be given as an input for SHA1, SHA224, and SHA256, while that of less than 2^{128} bits for SHA384 and SHA512. The size of a hash ranges from 160 to 512 bits, depending on the algorithm.

B. PoW of Bitcoin

In Table I, we provide routines to solve a puzzle in bitcoin. In Step 2, a miner puts a given set of the block header (BH) with a selected nonce through SHA256 to yield a hash of 256-bits. In Step 3, this puzzle is declared to be solved if this hash begins with L zero bits, where L is the difficulty of the given puzzle. Otherwise, the miner repeats the routines from Step 2 to Step 3 by only varying the nonce.

C. PoW of Ethereum

In [18], Ethash is used to prevent the advent of ASIC miners. Operations to fetch data from a memory called directed acyclic graph (DAG) are regularly required, where the data of DAG are randomly re-generated every 30,000 blocks. The usage of these operations makes Ethash resistant to ASIC devices.

Table II shows routines of Ethash. In Step 2, a given set of the current BH with a selected nonce is taken by SHA3 to get a hash. This hash is taken by a predefined function to yield mix0 that is random. In Step 4, mix0 is used to determine which data from DAG are fetched. Since mix0 is random, no one predicts which data will be fetched. The mixer takes both the fetched data and mix0 to get a random value in Step 5. In Step 6, mix0 is updated using this random value. The routines from Step 2 to Step 6 are repeated 63 times. The decision is conducted based upon this final mix0, as we have shown in Step 8.

The reason for prohibiting the advent of ASIC devices is that the operation time for the mixer is shorter than that of the fetch operation, leading to bottleneck between the mixing operation and the fetching operation. Thus, there is no advantage of using ASIC devices unless this bottleneck is solved.

Recently, in the Ethereum community, a Programmatic PoW (ProgPoW) has been proposed to improve the ASIC-resistance by randomly changing parameters of DAG from block to block. This modification makes the operation time related to the fetch operation increase, which intentionally generates the bottleneck. The foundation decided to replace ProgPoW with Ethash, but the development of ProgPoW is not completed.

D. PoW of Dash

X11 was proposed in 2014 by Duffield [19], the developer of Dash cryptocurrency. In Table III, we provide routines of X11, which consists of hash functions below:

Blake, Bmw, Groestl, Jh, Keccak, Skein, Luffa, Cubehash, Shavite, Simd and Echo.

Blake first takes a given set of the current BH with a selected nonce to get its hash. Next, Bmw takes this hash as its input to yield a hash. The same procedures are repeated until Echo, the last hash function, yields its hash. The decision is made on the basis of this last hash, as we have shown in Step 13.

TABLE IV. The map for X16r

Value	Hash	Value	Hash
0	Blake	8	Shavite
1	Bmw	9	Simd
2	Groestl	a	Echo
3	Jh	b	Hamsi
4	Keccak	c	Fugue
5	Skein	d	Shabal
6	Luffa	e	Whirlpool
7	Cubehash	f	Sha512

However, the order of the hash functions in X11 is fixed over time. This fixed order makes the development of ASIC devices possible. To develop such a device, we need to implement these 11 hash functions in this device. Logic gates to sequentially link these hash functions should be implemented as well. Until early 2016, the manufacturing cost of developing such a device was expensive. X11 was successful before early 2016. However, in 2016, an ASIC device called Antminer-D3 was developed.

The concept behind X11 has been extended to provide other PoWs such as X13, X15, and X17. As these names suggest, they use 13, 15, and 17 hash functions, respectively. Currently, ASIC devices for both X13 and X15 are available while there is no ASIC device for X17.

E. PoW of Raven

In 2018, a new extension of X11 was proposed in [20]. This is called X16r. It uses multiple hash functions given in Table IV to get the last hash like to other extensions of X11 that we have considered. But, unlike to the others, the sequence of the hash functions in X16r can vary from block to block. This variation prevents the development of ASIC devices for X16r.

We present an example to address how X16r works. First, the sequence is determined upon the last 16 bytes of the hash value of the previous block. Let the hash value of the previous block be 0000...04def2c3eff**6da11542ffcdabce**. The last 16 bytes are **6da11542ffcdabce**. Then, the sequence is decided on the basis of Table IV as follows:

Luffa -> Shabal -> Echo -> Bmw -> Bmw -> Skein -> Keccak -> Groestl -> Sha512 -> Sha512 -> Fugue -> Shabal -> Echo -> Hamsi -> Fugue -> Whirlpool.

A miner puts a given set of the BH with a selected nonce through Luffa to obtain the hash value taken by Shabal. This routine is repeated until the last hash is given. In this example, the last hash is provided by Whirlpool.

No one has succeeded in making ASIC devices for X16r. But, Black and Weight [20] stated that reordering the sequence of the hash functions cannot make the development of ASIC devices impossible.

F. Short summary from C to E

From subsection II.E to II.C, we have reviewed the existing ASIC-resistant PoWs which are categorized as follows:

- a. intentional memory access.
- b. multiple hash functions.

Ethash and ProgPoW can belong to the first class while X11

TABLE III. The routine of X11

Inputs: BH and L	
Step 1:	for nonce = 0, 1, 2, ... $2^{32} - 1$
Step 2:	$\mathbf{e} = \text{Blake}(\text{nonce}, \text{BH})$
Step 3:	$\mathbf{e} = \text{Bmw}(\mathbf{e})$

Step 12:	$\mathbf{e} = \text{Echo}(\mathbf{e})$
Step 13:	If \mathbf{e} begins with L zero bits, then go to Step 15.
Step 14:	end
Step 15:	Block generation & broadcast

and its variants such as X13, X15, X17 and X16r can belong to the second class. The basic idea of the first class is to use the bottleneck intentionally caused by randomly fetching data from a memory. The basic idea of the second one is to use the multiple hash functions, which can make the development costs of ASIC devices expensive.

At the present time of writing this manuscript, ASIC mining devices for Ethash, X11, X13 and X15 are available. The development of ProgPoW is not completed. X17 and X16r can be resistant to ASIC mining devices. But, as the ASIC-resistant property of the PoWs such as X11, X13 and X15 are broken, that of X17 will be cracked in the near future when the hardware development technology is improved. Also, according to the statements of the X16r developers [20], ASIC mining devices for X16r will be developed in the near future.

III. Literature Survey on LDPC

In ECCPoW, a decoder of LDPC codes is used to randomly generate an output which is used to determine whether a puzzle is solved or not. We present the summaries of LDPC codes and a decoder in this section.

In 1963, LDPC codes were proposed by Gallager [7]. But, the codes received no attention because the technology was not sufficiently mature to implement decoders. However, Mackay and Neal [22] reported that the codes could achieve the Shannon limit [21] when decoded using a belief propagation algorithm. Extensive studies on the codes have been done. They are categorized as follows: *i)* constructing LDPC codes to approach the Shannon limit [21] and *ii)* implementing decoders based on either ASIC [23]–[26] or field programmable gate array (FPGA) [27]–[28] for supporting real-time processes.

A. LDPC codes

LDPC codes can be generalized to a non-binary alphabet for improving its error-correction capability. But, the aim of using these codes in ECCPoW is not to correct errors. This makes us consider LDPC codes consisting of the binary alphabet. A (n, k) LDPC code is a linear code constructed by supplementing each message \mathbf{m} of size k with parity bits to get a codeword of size \mathbf{n} . This code is defined in terms of a parity check matrix \mathbf{H} of size $m \times n$ such that each element is either 0 or 1 and the number of 1s is small, where m is the number of parity bits and $m = n - k$.

The (n, k) LDPC code is either regular or irregular depending on the form of \mathbf{H} . If \mathbf{H} contains a constant number w_c of 1s in each column and a constant number w_r of 1s in each row, the

code is regular. For a given regular LDPC code, the parameters such as n , k , w_c , and w_r satisfy the following:

$$nw_c = (n - k)w_r. \quad (1)$$

If \mathbf{H} contains a different number of 1s in both each column and each row, \mathbf{H} is irregular. The error-correction performance of irregular codes is better than that of regular codes. However, we will consider regular codes because of the two reasons: *i*) it is much easier to implement a decoder of regular LDPC codes and *ii*) the aim of using this decoder is not to correct errors but to yield a random output.

A bipartite graph is often used to represent an LDPC code, as we have shown in Fig. 1. The lower and upper nodes are called the variable nodes and the check nodes, respectively. Each edge shows the adjacency of the i^{th} variable node and the j^{th} check node and corresponds to a nonzero $(i, j)^{\text{th}}$ element in \mathbf{H} .

The error-correction capability of a given LDPC code relies on the minimum (Hamming) distance d . This is given by considering any pair of 2^k codewords as follows:

$$d = \min_{\mathbf{u} \in \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{2^k}\} \setminus \{\mathbf{0}_n\}} \|\mathbf{u}\|_h \quad (2)$$

where $\|\mathbf{x}\|_h$ is the number of 1s in \mathbf{x} .

Studies on the computation of the minimum distance have been reported in the literature. Keha and Duman [29] proposed a branch and cut algorithm to compute the minimum distance of LDPC codes. But, it requires a considerable amount of time and memory for computing the minimum distance; it is only useful if n is small. Hashemi and Banihashemi [30] proposed an algorithm to find lower and upper bounds of the minimum distance of LDPC codes and obtained both of the bounds even when $n > 64,000$. For regular LDPC codes with a certain pair of w_c and w_r , the upper and lower bounds of a relative minimum distance which is the ratio between d and n are given in [31] and [7], respectively. We invoke the following theorem that states an error-correction capability of a given linear:

Theorem 1 [32]: Let a linear code have a minimum distance d . The number of correctable errors is

$$t = \lfloor (d - 1) / 2 \rfloor \quad (3)$$

where $\lfloor x \rfloor$ denotes the integer part of x .

We will use this theorem to compute an expected value of the number of hash cycles required to solve a given puzzle, i.e., mine a block, in ECCPoW.

Next, we consider how to encode a message \mathbf{m} for a given \mathbf{H} of size $m \times n$. To this end, we build a generator matrix \mathbf{G} of size $n \times k$ whose rows are orthogonal to the rows of \mathbf{H} as follows:

Step 1: Conduct the Gaussian elimination to rewrite \mathbf{H} as follows:

$$\mathbf{H} = \begin{bmatrix} -\mathbf{A}^T & \mathbf{I}_{n-k} \end{bmatrix}$$

where \mathbf{I}_{n-k} is the identity matrix of size $(n - k) \times (n - k)$.

Step 2: Form \mathbf{G} of size $n \times k$ as follows:

$$\mathbf{G} = \begin{bmatrix} \mathbf{I}_k & \mathbf{A} \end{bmatrix}.$$

The message \mathbf{m} is encoded to obtain a codeword \mathbf{c} via $\mathbf{c} = \mathbf{G}\mathbf{m}$.

A decoder takes both a parity check matrix \mathbf{H} and a corrupted codeword \mathbf{r} , which is $\mathbf{r} = \mathbf{c} + \mathbf{e}$, where \mathbf{e} is an error pattern. The decoder runs a message passing algorithm [32] known to be the standard decoding algorithm to remove \mathbf{e} .

The principle behind the algorithm is to iteratively propagate messages among the variable and check nodes. These iterations are terminated when either the number of iterations exceeds a given number or a decoded output is a codeword. Explanations on how it operates are given in [32], i.e., Algorithm 5.1 on page 220. The algorithm takes parameters such as \mathbf{H} , \mathbf{r} , maxIter, and a crossover error probability ε , where \mathbf{H} is a parity check matrix of size $m \times n$, \mathbf{r} is a vector of size n , and maxIter is the number of maximum iterations, and $0 < \varepsilon < 1$ is used to determine the initial value of the algorithm.

The error-correction performance of the algorithm depends on both maxIter and the crossover error probability. If maxIter is small, the algorithm fails to obtain a converged solution. If it is large, the algorithm may take a considerably long computational time to obtain its solution. In the literature, MaxIter is set from 10 to 20. Next, the crossover error probability is set if the transition probability of a binary symmetric channel is either given or estimated. If this is improperly set, two problems can occur: *i*) the performance degrades and *ii*) even if the performance does not degrade, it takes a considerably large number of iterations to obtain a solution. In ECCPoW, the aim of using the decoder is to intentionally spend a considerable amount of time for finding a codeword. Hence, there is no need to set MaxIter and ε strictly. One condition that we have to satisfy is that all of miners in ECCPoW have to use the same values of these parameters. We then will give details on how to make the other parameters, such as \mathbf{H} and \mathbf{r} , in Section IV.

B. FPGA and ASIC Implementation

LDPC decoders based on ASIC, a.k.a. ASIC-LDPC decoders aim to achieve low power consumption and fast process. In the decoders, the connections among the check and variable nodes have to be physically linked depending on a given parity check matrix. These connections cause the limited flexibility on the designs of the decoders. Thus, the decoders only support either a given set of parity check matrices or structured parity check matrices. Here, we provide a set of existing ASIC-LDPC decoders below:

The ASIC-LDPC decoder [23] supports quasi-cyclic parity check matrices decomposed into cyclic-shifted identity or zero matrices. The ASIC-LDPC decoder [24] supports parity check matrices included in the IEEE 802.16e system. In a survey paper regarding the state-of-the-art ASIC-LDPC [25], Hanzo *et al.*, stated that ASIC-LDPC decoders have to take a bank of hardware to support many parity check matrices. Thus, additional components such as memories, complex controllers and switchable interconnections are required, resulting in that they occupy the most area in the decoders. For example, let consider the ASIC-LDPC decoder discussed in [26]. This decoder supports about 100 parity check matrices, but its additional com-

ponents occupy 75% of the total area. This example shows that there is no practical implementation on ASIC-LDPC decoders that can support an infinite number of parity check matrices.

FPGA-LDPC decoders, which are LDPC decoders based on FPGA, consume more power and occupy more area rather than ASIC-LDPC decoders do. However, it is much easier for us to reprogram the FPGA-LDPC decoders. Namely, they provide the more flexibility in the design compared to the ASIC-LDPC decoders. The FPGA-LDPC decoder discussed in [27] supports parity check matrices up to a code length of $n = 65,000$. But, we need to manually load the parity check matrix onto this decoder when the matrix varies, requiring additional time. In a survey paper published by Hanzo *et al.* [28], FPGA-LDPC decoders require additional routing and processing units to support many parity check matrices. The use of these additional units can lead to complex designs, increasing the cost of the decoders.

IV. ERROR-CORRECTION CODES PROOF OF WORK

In this section, we will give details regarding ECCPoW. First, we will show how to build both the hash vector \mathbf{r} and the parity check matrix \mathbf{H} taken by a decoder as its inputs. Next, we will present the corresponding pseudo codes of ECCPoW and their explanations, and end this section with the theoretical results of ECCPoW.

Here, for simplicity, we refer to the parity check matrix, the current block header to be mined, and the previous block header as PCM, CBH, and PBH, respectively. The terminologies used are defined as follows.

Definition 1 – Hash Vector: A hash vector \mathbf{r} , which is a vector of concatenating outputs of SHA256s, of size n is as follows:

$$\mathbf{r} := \begin{cases} \mathbf{s}_1[1:n] & \text{if } n \leq 256 \\ \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_l & \mathbf{s}_{l+1}[1:j] \end{bmatrix} & \text{if } n > 256 \end{cases} \quad (4)$$

where $l = \lfloor n/256 \rfloor, j = n - 256 \times l$,

$$\mathbf{s}_1 := \text{SHA256}(\text{nonce}, \text{CBH}) \in \{0,1\}^{256} \quad (5)$$

and

$$\mathbf{s}_u := \text{SHA256}(\mathbf{s}_1) \in \{0,1\}^{256} \quad (6)$$

where $u = 2, 3, \dots, l + 1$.

Definition 2 – Decoder: A PCM \mathbf{H} of size $m \times n$ is constructed using the hash value of a PBH by following the pseudo codes given in Table VI. A hash vector \mathbf{r} of size n is constructed using (4). A decoder \mathcal{D}_{MP} takes both \mathbf{r} and \mathbf{H} as its inputs and runs the message passing algorithm given in [32] to yield a vector \mathbf{n} of size n :

$$\mathcal{D}_{MP} : \{\mathbf{r}, \mathbf{H}\} \mapsto \mathbf{n} \in \{0,1\}^{n \times 1}. \quad (7)$$

Definition 3 – Puzzle: A puzzle is defined using the parameters such as a PCM \mathbf{H} , a CBH, and a decoder defined in (7). Then, the puzzle is solved when a nonce makes the decoder yield a codeword:

$$\mathcal{P}_{\mathbf{H}, \mathcal{D}_{MP}, \text{CBH}} : \{\text{nonce}\} \mapsto \mathbf{n} \quad (8)$$

where \mathbf{n} is the output of the decoder that satisfies

$$\mathbf{H}\mathbf{n} = \mathbf{0}_m. \quad (9)$$

First, the construction of the hash vector is given in Definition 1. We input a nonce and a constant CBH into SHA256 to obtain its output \mathbf{s}_1 , as we have shown in (5). We then construct $\mathbf{s}_2, \mathbf{s}_3, \dots, \mathbf{s}_{l+1}$ using (6). Last, we construct a hash vector \mathbf{r} using all of these outputs, as we have shown in (4).

Next, we use a given PBH to construct a PCM, which has to satisfy the following conditions: *i)* any verifiers can reconstruct the PCM that the miner used, and *ii)* the PCM varies from block to block. To meet these conditions, we use a method proposed by Gallager [7] that constructs \mathbf{H} as follows:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} \\ \pi_1(\mathbf{A}) \\ \vdots \\ \pi_{w_c-1}(\mathbf{A}) \end{bmatrix} \in \{0,1\}^{\frac{mw_c}{w_r} \times n} \quad (10)$$

where $\pi_i(\mathbf{A})$ is the i^{th} matrix constructed by the column permutation of \mathbf{A} , π_i is the i^{th} permutation order, and

$$\mathbf{A} := \begin{bmatrix} \mathbf{1}_{w_r} & & & \\ & \mathbf{1}_{w_r} & & \\ & & \ddots & \\ & & & \mathbf{1}_{w_r} \end{bmatrix} \in \{0,1\}^{w_c \times n}$$

whose i^{th} row has 1s in rows from $(i-1) \times w_r$ to $i \times w_r$, and

$$\mathbf{1}_{w_r} := [1 \ 1 \ \cdots \ 1] \in \mathbf{1}^{1 \times w_r}.$$

Each permutation order requires a seed value. In ECCPoW, we first generate the initial seed value as follows:

$$S := \text{PBHV}[0] + \text{PBHV}[1] + \cdots + \text{PBHV}[31] \quad (11)$$

where PBHV is referred to as a hash value of a given PBH and the data type of PBHV is assumed to be 32bytes. The i^{th} permutation order is generated using $S - i + 1$. The pseudo code of the proposed method of PCM is given in Table VI.

Next, we address how this pseudo code given in Table VI can satisfy the conditions mentioned earlier. In Step 4, the i^{th} permutation order is constructed using $S - 1 + i$. As any verifier can know PBHV, a verifier needs to obtain the same seed value S without any communications. Thus, the verifiers will be able to easily reconstruct the PCM that the miner used, confirming that the proposed method satisfies the first condition. Next, PBHV is random; i.e., the initial seed values are random. Then, all of the permutation orders are provided using the initial seed value. Thus, the orders vary from block to block, confirming that the proposed method satisfies the second condition.

We show the pseudo codes of ECCPoW in Table V. In Step 1, a nonce is selected from 0 to $2^{32} - 1$. We construct a hash vector \mathbf{r} using (4), as we have shown in Step 2. In Step 3, we execute

the decoder to give an output by taking both \mathbf{e} and \mathbf{H} into it. The decision is obtained using (9) in Step 4. If the output is not a codeword, we repeat the routines from Step 1 to Step 4.

We have to try the routines from Step 1 to Step 4 many times until finding a particular nonce that makes the decoder gives a codeword. It is natural to study the number of trials for finding this nonce. For this end, we analyze ECCPoW to give answers to the following questions:

- Q1. How many the number of hash cycles is needed to solve a given puzzle?
- Q2. How does the number of miners affect the number of hash cycles?
- Q3. Which parameters affect the number of hash cycles?

We begin to define hash cycle, success event, and mining game as follows.

Definition 4 – Hash Cycle: A hash cycle is defined as the whole routines from Step 1 to Step 4, given in Table V.

Definition 5 – Success Event: A success event occurs if a nonce such that the LDPC decoder defined in (7) can yield a codeword is found.

Definition 6 – Mining Game: Let both a PCM \mathbf{H} of size $m \times n$ and a CBH be given. There are M miners, and they use the same single computer. Then, a mining game (MG)

$$\text{MG}\{\mathbf{H}, \text{CBH}, M, p\} \quad (12)$$

is defined that the miners struggle against each other in a race to get a success event first, where p is the success probability of the LDPC decoder for the given PCM \mathbf{H} :

$$p \triangleq \Pr\{\mathbf{H}\mathbf{n}^T = \mathbf{0}_m\} \quad (13)$$

where \mathbf{n} is the output of the decoder taking an arbitrary vector.

Definition 7: For a given MG, X_M is defined as a random variable that represents the number of hash cycles to end this given MG.

For a given PCM \mathbf{H} of size $m \times n$, there are 2^k codewords:

$$\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \dots, \mathbf{c}_2\}.$$

Under an assumption that the decoder defined in (7) is optimal, this decoder can find a success event with a probability:

$$p = \sum_{i=1}^{2^k} \Pr\{\mathbf{n} = \mathbf{c}_i\} = \sum_{i=1}^{2^k} \sum_{l=0}^t \Pr\{\|\mathbf{r} - \mathbf{c}_i\|_h = l\} \quad (14)$$

where \mathbf{r} is a hash vector constructed in (4) and \mathbf{n} is an output of the decoder. We define a sphere set for the given i^{th} codeword

$$\mathcal{A}(\mathbf{c}_i, t) \triangleq \{\mathbf{r} : \|\mathbf{r} - \mathbf{c}_i\|_h \leq t\} \quad (15)$$

whose cardinality is

$$|\mathcal{A}(\mathbf{c}_i, t)| = 1 + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{t} = \sum_{l=0}^t \binom{n}{l} \quad (16)$$

where t is a positive integer. Then, we assume that the decoder

TABLE V. The pseudo codes for ECCPoW

Inputs: CBH and PCM \mathbf{H}	
Step 1:	A nonce is uniformly chosen from $[0, 2^{32} - 1]$
Step 2:	Construct a HV \mathbf{r} using (4) with a chosen nonce and the given CBH.
Step 3:	Obtain a vector \mathbf{n} using (7) with the given PCM \mathbf{H} .
Step 4:	If \mathbf{n} can satisfy (9), then go to Step 5.
Step 5:	Block generation & broadcast

TABLE VI. The pseudo codes to construct PCM

Inputs: n, w_c, w_r and BHV	
Output: \mathbf{H}	
Step 1:	Construct S using (11).
Step 2:	Construct \mathbf{A} by following the statements below (10) and $\mathbf{H} = \mathbf{A}$.
Step 3:	for $i = 2$ to $w_c - 1$
Step 4:	Construct π_i with the seed value $S - i + 1$.
Step 5:	$\mathbf{H} = [\mathbf{H} \quad \pi_i(\mathbf{A})^T]$.
Step 6:	end
Step 7:	$\mathbf{H} = \mathbf{H}^T$.

is optimal, implying that this decoder can yield the i^{th} codeword when it takes a word belonging to the i^{th} sphere. We then have

$$p = 2^{k-n} |\mathcal{A}(\mathbf{c}_i, t)| = 2^{k-n} \sum_{l=0}^{\lfloor \frac{d-1}{2} \rfloor} \binom{n}{l} \leq 2^{k-n} \sum_{l=0}^{\lfloor \frac{d}{2} \rfloor} \binom{n}{l} \quad (17)$$

where the second equality comes from (3). Now, we investigate both upper and lower bounds on this probability p with respect to n . This investigation leads to Proposition 1, showing that p increases with an increase in n . Then if the ratio between w_c and w_r satisfies (21), an increase in n decreases p .

Proposition 1 – Let $w_c \geq 3$, $w_r > w_c$ be constant and their ratio be

$$w_c/w_r =: \alpha \in (0, 1) \quad (18)$$

which is also constant. Let the size of a PCM \mathbf{H} be $m \times n$. Then, for any $0 < \delta < 1/2$, we have

$$2^{-n\alpha} \leq p \leq 2^{-n(\alpha - H(\delta/2))} \quad (19)$$

where $H(x)$ is the binary entropy defined as follows:

$$H(x) = -x \log_2 x - (1-x) \log_2 (1-x). \quad (20)$$

Indeed, let the ratio further satisfy the following:

$$\alpha \in (H(0.25), 1). \quad (21)$$

Then, the success probability p can vanish with an increase in n .

Proof: From (1) with (18), we infer that

$$k - n = -n\alpha. \quad (22)$$

The results of [7] state that the minimum distance of a regular LDPC code with constant w_r and $w_c \geq 3$ can linearly increase with an increase in n , implying that for $0 < \delta < 1/2$, the distance is $d = \lfloor \delta n \rfloor$. Substituting (22) into (17) leads to

$$p = 2^{k-n} \sum_{l=0}^{\lfloor \frac{\delta n}{2} \rfloor} \binom{n}{l} \leq 2^{k-n} 2^{nH(\frac{\lfloor \delta n \rfloor}{2n})} \leq 2^{-n(\alpha - H(\delta/2))}$$

where the second inequality comes from the fact that any integers $n \geq k \geq 1$ with $k/n \leq 1/2$,

$$\sum_{l=0}^k \binom{n}{l} \leq 2^{nH(k/n)}.$$

The lower bound is obtained by assuming $t = 0$ in (16); i.e.,

$$p = 2^{k-n} |\mathcal{A}(\mathbf{c}_i, t)| \geq 2^{k-n} |\mathcal{A}(\mathbf{c}_i, 0)| = 2^{k-n} = 2^{-n\alpha}$$

which completes the proof.

Next, for a given MG $\{\mathbf{H}, \text{CBH}, M, p\}$, we let X_M be j . Then, we ensure that all of the miners fail to find a success event from the 1th hash cycle to the $(j-1)$ th hash cycle, but one of them at least succeeds to find this event at the j th hash cycle. Thus, X_M follows a geometric distribution as follows.

Theorem 2: For a given MG $\{\mathbf{H}, \text{CBH}, M, p\}$, we have

$$\Pr\{X_M = j\} = p_{\text{failure,all}}^{j-1} \times (1 - p_{\text{failure,all}})$$

which is a geometric distribution with the following parameter:

$$p_{\text{failure,all}} := (1 - p)^M. \quad (23)$$

Then, we have

$$\mathbb{E}[X_M] = (1 - p_{\text{failure,all}})^{-1} \quad (24)$$

and

$$\mathbb{V}[X_M] = p_{\text{failure,all}} (1 - p_{\text{failure,all}})^{-2}.$$

Proof: The proof is clear, as the random variable X_M follows a geometric distribution with (23). We thus omit it.

We remind that for constant w_c and w_r , Proposition 1 shows that p decreases with an increase in n as long as (21) is satisfied. We then notice that the expected value given in (24) decreases with respect to n , which implies that a puzzle becomes difficult to be solved with an increase in n .

We invoke the results of [31] and [7]. For a certain pair of w_c and w_r , the results show upper and lower bounds of a relative minimum distance δ , the ratio between the minimum distance d and the code length n . As an example, for $w_c = 4$ and $w_r = 5$, the upper and lower bounds are 0.3238 and 0.2111, respectively. For $w_c = 4$ and $w_r = 8$, they are 0.1765 and 0.0627, respectively. The results are given in an asymptotic analysis; i.e., n goes to infinity. By utilizing these results, we can obtain the lower and upper bounds on the expected value of the random variable as follows:

$$\frac{1}{1 - (1 - g(n, k, \delta_1))^M} \leq \mathbb{E}[X_M] \leq \frac{1}{1 - (1 - g(n, k, \delta_2))^M} \quad (25)$$

where δ_1 and δ_2 are the upper and lower bounds of the relative

minimum distance and

$$g(n, k, \delta) := 2^{k-n} \sum_{l=0}^{\lfloor \frac{n\delta-1}{2} \rfloor} \binom{n}{l}.$$

Then, we consider the trends of (24) with respect to n and M by manipulating both of the bounds in (25).

In Table VII, we provide the upper and lower bounds of (24) for constant $w_c = 4$ and $w_r = 5$. Although the ratio between w_c and w_r does not satisfy (21), we can see that the lower bound increases with an increase in n . That is, it increases from 1.58×10^4 to 2.46×10^4 when n is increased from 80 to 160. This result implies that increasing n makes a puzzle more difficult to solve when (21) is not satisfied. For the other pairs of w_c and w_r given in [31], the same result is observed.

Indeed, increasing the number of miners M can decrease the upper bound. As an example, we consider the upper bound for $n = 120$ and $k = 24$. This bound decreases from 6.69×10^{12} to 3.34×10^{12} as M is increased from 1 to 20. Intuitively, this result is valid because a given MG ends early, as more miners are involved in solving a puzzle. Then, there is another intuition that an MG ends at the 1st hash cycle if an infinite number of miners work. We confirm these intuitions by establishing Corollary 1.

Corollary 1: Let MG $\{\mathbf{H}, \text{CBH}, M, p\}$ be given. The expected value given in (24) decreases with an increase in the number of miners M . In particular, this value can converge to 1 as M goes to infinity.

Proof: It is immediately seen that

$$\frac{d\mathbb{E}[X_M]}{dM} = -\frac{\log p_{\text{failure,all}}}{p_{\text{failure,all}}} \leq 0$$

implying that the expected value given in (24) is a decreasing function of M . As M goes to infinity, the parameter defined in (23) goes to zero. Thus, the expected value converges to one.

The decoding process, i.e., Step 3 in Table V, can occupy the most computational time in a single hash cycle. In this decoding process, matrix-vector products are required, implying that the computational cost to run a single hash cycle can be modeled as $O(mn)$. We remind that each miner uses the same single computer in a given MG. Thus, we can assume that each miner only runs τ operations per second. This assumption makes us define an expected value of a block generation time as follows.

Definition 10 – Block Generation Time: A MG $\{\mathbf{H}, \text{CBH}, M, p\}$ is given. Each miner is assumed to run τ operations per second. Then, the block generation time T can be defined as

$$T := \tau^{-1} \mathbb{E}[X_M] O(mn). \quad (26)$$

We remind that proposition 1 states that *i*) the upper bound of the success probability p decreases with an increase in n and *ii*) the lower bound of p increases with a decrease in n . Then, the decrease in this upper bound makes a puzzle more difficult to be solved. In other words, more hash cycles are needed, leading to an increase in T . Similarly, it is seen that a puzzle becomes easy to be solved when we decrease n . As a result, we can vary

T by controlling n .

Corollary 2: Let $w_c \geq 3$ and w_r be constant, \mathbf{H} of size $m \times n$ be given, and the ratio α between m and n satisfy (21). Then, for any $0 < \delta < 1/2$, we have

$$\frac{\tau^{-1}O(\alpha n^2)}{1 - (1 - 2^{-n\alpha})^M} \leq T \leq \frac{\tau^{-1}O(\alpha n^2)}{1 - (1 - 2^{-n(\alpha - H(\delta/2))})^M}$$

where $H(x)$ is defined in (20) and M is the number of miners.

Proof: It is seen that

$$O(mn) = O(\alpha n^2)$$

because of both (18) and (1). Substituting (19) into (24) leads to

$$\frac{1}{1 - (1 - 2^{-n\alpha})^M} \leq \mathbb{E}[X_M] \leq \frac{1}{1 - (1 - 2^{-n(\alpha - H(\delta/2))})^M}$$

The proof is completed by combining this equation with (24).

V. DISCUSSIONS ON ECCPoW

In this section, we will consider the worth of ECCPoW as a new PoW. To this end, we begin to define general properties for PoWs and show how ECCPoW can satisfy these properties. We then define a property only belonging to ECCPoW as the most innovative part. This property makes ECCPoW a solution to the problems caused by ASIC mining devices, as we have stated in Section I. Here, we define the following properties as follows:

- P1. A puzzle has to be time-consuming, but it is easy to check whether a given solution is correct or not.
- P2. Any previous solution cannot be used to find a current solution.
- P3. The difficulty of a puzzle can be changed.
- P4. A puzzle can only be solved when miners follow the routines of PoW.
- P5. A puzzle can be time-variant from block to block.

The existing PoWs can have the properties from P1 to P4. As an example, let us begin to consider bitcoin. First, each puzzle is designed to be solved per roughly 10 minutes. On the other hands, validation of a given solution can be immediately done. Second, the BH at the current block being mined are different to that of previously mined blocks. SHA256 takes both a nonce and a BH to yield a hash. Any modification on these inputs thus makes the hash completely different. Thus, it is extremely rare that the hash can begin with L zero bits even we use a solution to the previously mined blocks. Third, whenever 2016 blocks are mined, the difficulty is adjusted based on the ratio between 20160 and the total spent time to mine these 2016 blocks. Last, the number of possible hash values is 2^{256} , while the number of solutions is $2^{(256-L)}$, i.e., solutions beginning with L zero bits. A possibility that a randomly given nonce is a solution is 2^L . In the 567,657th block of bitcoin, L is 72. Thus, it is probabilistically impossible to obtain a solution without following the routines in Table I.

We now use the results obtained in the previous section to

show how ECCPoW has the properties from P1 to P5.

Corollary 3: Let a puzzle defined in (8) be given. It is easy to verify whether a given solution is correct or not. But, solving this puzzle is time-consuming as compared to its verification.

Proof: The verification of whether this solution is correct is done by only operating the procedures from Step 1 to Step 4 given in Table V. Thus, it requires a single construction of a hash vector by following (4) and a single execution of the decoder in (7). In contrast, to solve the puzzle, the routines from Step 1 to Step 4 given in Table V have to be repeated several times, as we have shown in Theorem 2. Thus, the puzzle can be time-consuming as compared to the verification.

Malicious miners can conduct cheats to make more profits as compared to honest miners who follow the routines in Table V. We consider two possible cheats as follows. First, the malicious miners report one of the previous solutions as a current solution. Second, the malicious miners report a randomly selected nonce as a current solution. We prepare corollaries to show that these cheats cannot be allowed.

Corollary 4: Any previous solution cannot be used to find a current solution.

Proof: We remind that a solution to a puzzle in ECCPoW is the value of nonce. The decoder takes a hash vector \mathbf{r} and a PCM \mathbf{H} to yield its output. This \mathbf{H} varies from block to block. Then, for a fixed \mathbf{r} , the decoder yields a different output if \mathbf{H} varies. Even we succeed to construct the same hash vector, which we used in solving a previous puzzle, using the previous nonce, the decoder gives a different output. This is the reason that \mathbf{H} taken by the decoder at the current puzzle is different to that taken by the decoder at the previous puzzle.

Corollary 5: Miners have to follow the routines given in Table V to solve a puzzle defined in (8).

Proof: We assume that the decision can be done without running the decoder. If this can be possible, two assumptions are required. The first one is that a hash vector \mathbf{r} taken by the decoder is a codeword and the second one is that there is a table that shows a pair of inputs and outputs of the decoder. A possibility of the first assumption is zero because we declare a hash vector which is a codeword itself as an incorrect solution. Next, to construct the table for a given PCM \mathbf{H} , we consider all of the possible hash vectors. Since the number of hash vectors is 2^n , it is unrealistic to construct this table whenever n is large. Besides, even we succeed to make the table at a given \mathbf{H} , this table becomes useless because \mathbf{H} is intentionally designed to vary from block to block.

Now, we consider examples by assuming that all the puzzles in ECCPoW have the same difficulty. If the number of miners M increases, the block generation time T defined in (26) is very small. On the other hands, if M decreases, T is very large. The considerable amount of variation of T makes ECCPoW unstable, as we have stated in Section IV. Thus, the difficulty must increase or decrease based on the number of miners. We remind Theorem 2 meaning that the difficulty can properly vary using the variation of n , leading to the following corollary.

Corollary 6: The difficulty of a puzzle is changeable.

Proof: It is clear; we thus omit it.

We show that a puzzle can be time-variant, which is useful in repressing the advent of ASIC devices.

Corollary 7: A puzzle defined in (8) can be time-variant.

Proof: A PCM H is used to define the puzzle and is constructed using a previous hash. That is, the i^{th} PCM is constructed using a hash of the $(i - 1)^{\text{th}}$ block. Hence, the constructed PCM varies from block to block, immediately implying that the puzzle can be time-variant from block to block.

An infinite number of PCMs must be constructed, as miners mine new blocks continuously. An example in [25] shows that there is no ASIC-LDPC decoder to support the infinite number of PCMs because of the limited flexibility. Hence, the decoder in ECCPoW has to be executed by either graphical processing units or central processing units. This is the fundamental reason that ECCPoW can preclude the development of ASIC mining devices, which can be a solution to the problems stated in Section I: *i*) the re-centralization of mining markets and *ii*) the considerably high usage of electrical energy for mining. This is the most valuable contribution of ECCPoW.

VI. CONCLUSIONS

PoW is fundamental to blockchain, as it is used to prohibit an unauthorized modification of mined blocks. However, the usage of ASIC devices can cause the problems stated in Section I, namely *i*) the re-centralization of the mining markets and *ii*) a considerable amount of electrical energy spent to mine blocks.

As a solution to the abovementioned problems, we proposed an ECCPoW, as we have shown in Fig. 1. To the best of our knowledge, this is the first study in which LDPCs are applied to PoW. We investigated the expected value of the number of hash cycles needed to solve a puzzle in ECCPoW. We showed that this value can be either increased or decreased as we vary the code length, the size of a hash vector taken by the decoder, and the number of miners. We also discussed how ECCPoW satisfies the five properties defined in Section V. The discussions show the value of ECCPoW as a general PoW.

As we have reviewed in Section III, there is no ASIC decoder that supports an infinite number of LDPC codes. This result motivated us to intentionally vary the LDPC codes from block to block, leading to the last property that a puzzle defined in (8) is time-variant. This is the most innovative aspect of ECCPoW in repressing the advent of ASICs, implying that the problems caused by ASICs can be solved using our ECCPoW.

REFERENCES

- [1] U. W. Chohan, "The double spending problem and cryptocurrencies", SSRN Electronic Journal, Jan. 2019
- [2] Nakamoto, S. *Bitcoin: A Peer-to-Peer Electronic Cash System*. Retrieved from: <https://bitcoin.org/bitcoin.pdf>
- [3] Quynh H. Dang, "Secure Hash Standard FIPS PUB 180-4," U.S. Dept. of Commerce and NIST, Aug. 4th 2015. <https://www.nist.gov/publications/secure-hash-standard>, Accessed Sept. 25th, 2018.
- [4] Jordan Tuwiner, "Bitcoin mining hardware ASICs," <https://www.buybitcoinworldwide.com/mining/hardware/>, June 30th, 2018. Accessed Sept. 26th, 2018.
- [5] A. Gervais, G. Karame, V. Capkun and S. Capkun, "Is bitcoin a decentralized currency?", *IEEE Security and Privacy Magazine*, Vol. 12, No. 3, pp. 54 – 60, May, 2014
- [6] Digiconomist. (2018). *Bitcoin Energy Consumption Index*. Retrieved April 3, 2018 from: <https://digiconomist.net/bitcoin-energy-consumption>
- [7] R. G. Gallager, *Low Density Parity Check Codes*, Monograph, M.I.T. Press, 1963
- [8] R. J. McEliece, "A public-key cryptosystem based on algebraic coding theory", *DSN Progress Report*, pp. 114-116, Jan.-Feb. 1978
- [9] D. J. Bernstein, T. Lange and C. Peters, "Wild McEliece",
- [10] M. Baldi, M. Bianchi, F. Chiaraluce, J. Rosenthal and D. Schipani, "Using LDGM codes and sparse syndromes to achieve digital signatures",
- [11] Marco Baldi, "LDPC codes in the McEliece cryptosystem: attacks and countermeasures", *NATO Science for Peace and Security – D: Information and Communication Security*, vol. 23, pp. 160 – 174,
- [12] M. Baldi, M. Bodrato and F. Chiaraluce, "A new analysis of the McEliece cryptosystem based on QC-LDPC codes", 2008.
- [13] T. P. Berger, P.-L. Cayerel, P. Gaborit and A. Otmani, "Reducing key length of the McEliece cryptosystem",
- [14] V. M. Sidelnikov, "A public-key cryptosystem based on binary Reed-Muller codes," *Discrete Mathematics and Applications*, vol. 4 no. 3, 1994.
- [15] L. R. Knudsen and B. Preneel, "Construction of secure and fast hash functions using nonbinary error correcting codes", *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2254 – 2539, Sept., 2002.
- [16] S. Ermon, C. P. Gomes, A. Sabharwal and B. Selman, "Low-Density Parity Constraints for Hasing-Based Discrete Integration", *Proceedings of the 31st International Conference on Machine Learning, PMLR 32(1):271-279*, 2014.
- [17] Vadhan, S. Pseudorandomness. *Foundations and Trends in Theoretical Computer Science*, 2011.
- [18] Vitalik Buterin. Dagger: A memory-hard to compute, memory-easy to verify script alternative. Tech Report, [hashcash.org website](https://github.com/dashpay/dash/wiki/Whitepaper), 2013.
- [19] <https://github.com/dashpay/dash/wiki/Whitepaper>
- [20] T. Black and J. Weight, X16R ASIC resistant by design, 2018. <https://ravencoin.org/wp-content/uploads/2018/03/X16R-Whitepaper.pdf>
- [21] C. E. Shannon, "A mathematical theory of communication", *The Bell System Technical Journal*, Vol. 27, 99. 379 – 423, 623 – 656, Jul., Oct., 1948
- [22] D. J. C. Mackay and R. M. Neal, "Near Shannon limit performance of low density parity check codes", *Electronics Letters*, Vol. 33, No. 6, pp. 457 – 458, 1997
- [23] M. Awais and C. Condo, "Flexible LPDC decoder architectures", *VLSI Design*, Vol. 2012, Doi: 10.1155/2012/730835
- [24] C-H. Liu, S-W. Yen, C-L. Chen, H-C. Chang, C-Y. Lee and S-J. Jou, "An LDPC decoder chip based on self-routing network for IEEE 802.16e Applications" *IEEE Journal of Solid-State Circuits*, Vol. 43, No. 3, pp. 684 – 694, Mar. 2008
- [25] S. Shao, P. Hailes, T-Y. Wang, J-Y. Wu, R. G. Maunder, B. M Al-Hashimi and L. Hanzo, "Survey of Turbo, LDPC and Polar Decoder ASIC Implementation", *IEEE Communications Surveys & Tutorials* 2019
- [26] Y. L. Ueng, B. J. Yang, C. J. Yang, H. C. Lee, and J. D. Yang, "An Efficient Multi-Standard LDPC Decoder Design Using Hardware-Friendly Shuffled Decoding," *IEEE Trans. Circuits Syst. I*, vol. 60, no. 3, pp. 743–756, March 2013.
- [27] C. Beuschel and H-J. Pfleiderer, "FPGA implantation of a flexible decoder for long LDPC codes", *Proc. Int. Conf. Field Program. Logic Appl.*, Sep. 2008, pp. 185–190
- [28] P. Hailes, R. G. Maunder and L. Hanzo, "A survey of FPGA-based LDPC decoders", *IEEE Communications Surveys & Tutorials*, Dec. 2015.
- [29] A. B. Keha and T. M. Duman, "Minimum distance computation of LDPC codes using a branch and cut algorithm", *IEEE Transactions on Communication*, Vol. 58, No. 4, pp. 1072 – 1079, Apr., 2010
- [30] Y. Hashemi and A. H. Banihashemi, "Tight Lower and Upper Bounds on the Minimum Distance of LDPC Codes", *IEEE Communications Letters*, Vol. 22, No. 1, pp. 33 – 36, Jan., 2018
- [31] Y. Ben-Haim and S. Litsyn, "Upper bounds on the rate of LDPC codes as a function of minimum distance", *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2092 – 2100, May., 2006

- [32] W. E. Ryan and S. Lin, Channel Codes Classical and modern, Cambridge

OSIA S&TR Journal

ISSN 1738-9887 Vol. 32, No. 1, March 2019

· 발행인 : 이혁준 회장/OSIA · 편집위원 : 김형식 교수/성균관대학교

02 Editorial

블록체인의 보안 및 상호운용성 연구 동향
김형식/성균관대학교

Article

04 50%미만 이중지불 공격
장재혁, 이흥노/광주과학기술원

11 Inter-Chain 기술 동향
김준희, 김중헌/중앙대학교

16 블록체인 기반의 ID 관리 기술 동향
김석현, 조영섭, 김수형/ETRI

23 IoT 환경에 적합한 블록체인 및
스마트 컨트랙트 기술 연구
김베드로, 이대화, 지우중, 김형식/성균관대학교



블록체인의 보안 및 상호운용성 연구 동향



성균관대학교

김형식

이미 “블록체인”이라는 용어는 더 이상 새롭거나 낯설지 않고, 인터넷과 웹처럼 친숙한 용어가 되고 있습니다. 2년 전의 암호화폐의 열기를 거쳐 학계와 산업계에서는 블록체인 기술을 둘러싼 많은 토론이 진행되고 있습니다. 그 중 일부는 블록체인 기술에 대한 무용론부터 시작하여 차세대 인터넷 플랫폼에 대한 가능성까지 논의하고 있습니다. 중요한 것은 블록체인 기술은 아직 현재 진행 중이며, 성능 및 보안성에 대한 많은 기술적인 난제들을 풀어야 하는 상황이라는 점입니다.

본 호에서는 현재 논의되고 있는 가장 중요한 블록체인 이슈인 보안성과 상호운용성에 대한 연구 결과를 소개하고 있습니다. 블록체인의 특성상 발생할 수밖에 없는 이중 지불 문제와 이기종 간의 블록체인 네트워크를 통합하기 위한 인터 체인 기술을 논의합니다. 블록체인의 중요한 응용 분야인 ID 관리 기술 및 IoT 관리 기술에 대한 최신 연구 동향을 소개하고 있습니다.

첫 번째 논문은 이중 지불이라는 블록체인의 고전적인 보안 문제에 대한 논문으로서 이중 지불에 대한 기존의 명제인 51% 이상의 컴퓨터 자원이 필요하다는 가정에 대한 의문을 제기하고 있습니다. 현실 블록체인 네트워크에서는 공격자의 컴퓨터 자원이 50% 미만인 경우에도 여전히 효과적인 이중 지불 공격이 가능하다는 것을 공격 성공률 및 공격자의 이윤을 모델링하여 몬테카를로 시뮬레이션을 통하여 증명하고 있습니다.

두 번째 논문은 이기종 간 블록체인 트랜잭션 및 데이터를 상호 교환 관리할 수 있도록 제안되고 있는 Inter-Chain 기술에 대해서 소개하고 있습니다. 현존하는 다양한 블록체인 네트워크 기술을 고려했을 때, 상호 운용성을 위하여 서로 다른 블록체인간 데이터 교환을 가능하게 해주는 Inter-Chain 기술은 향후 블록체인 기반 생태계를 구축하는데 중요한 이슈가 될 것이라 판단됩니다.

세 번째 논문은 블록체인을 이용한 사용자 계정 관리 기술을 소개하고 있습니다. 기존의 ID 관리 기술은 사용자의 개인 정보가 중앙의 특정 기관에서 관리되기 때문에 사용자의 자기 정보 통제 기능이 미흡하고, 개인 정보가 유출되는 사고 등이 발생할 수 있는 문제를 가집니다. 따라서 이러한 문제를 해결하기 위해서 사용자가 스스로 자신의 계정을 관리할 수 있는 방안이 필요합니다. 본 논문에서는 블록체인을 이용하여 어떻게 이러한 문제를 해결하고 탈중앙화된 사용자 계정 관리 시스템을 구축할 수 있는지를 소개하고 있습니다.

네번째 논문은 IoT 환경에서의 블록체인 기술에 대한 요구 사항을 분석하고, 요구 사항에 적합한 다양한 프로젝트를 소개하고 있습니다.

본 특집호 발간을 위해 소중한 시간을 내어 원고를 집필해 주신 집필자분들과 편집에 수고해 주신 학회지 편집 위원회 여러분께 깊은 감사를 드립니다.

50%미만 이중지불 공격

장재혁, 이흥노
광주과학기술원

Abstract

블록체인은 전 세계에 분포된 수많은 네트워크 노드들이 하나의 거래장부를 공동으로 기록, 관리하는 분산화 장부 시스템이다. 블록체인 기술은 합의와 분산화를 바탕으로 거래 기록의 위조가 불가능하도록 설계되었으나, 가상화폐 시장의 성장과 함께 이중지불 공격 (double-spending attack)을 통해 장부를 위조하여 부당이익을 취하려는 시도가 끊임없이 존재해 왔다. 본 논문에서는 이중지불 공격의 위험성을 분석한다. Satoshi Nakamoto는 2008년 비트코인에 대한 이중지불 공격의 성공을 위해서는 전 세계의 노드들이 보유한 컴퓨터 자원보다 더 많은 컴퓨터 자원, 즉 51% 이상의 컴퓨터 자원이 필요하다는 결론을 내렸다. 반면, 본 논문에서는 50% 미만의 적은 컴퓨터 자원을 사용하는 이중지불공격도 위협적임을 보였다. 이러한 결론은 이중지불 공격의 성공 확률이 아닌 기대 이윤을 분석함으로써 얻을 수 있었다. 구체적으로는, 대규모 시뮬레이션을 통해 실제 작동중인 블록체인 네트워크에 50% 미만의 이중지불 공격을 행할 경우에 대한 공격자가 얻는 기대 이윤을 측정하였으며, 이러한 공격을 방지 할 수 있는 방안을 제시한다.

I. 서론

블록체인은 비트코인과 이더리움 등 현존 가상화폐의 핵심 기술이며, 세계에 분포된 수많은 노드들이 하나의 거래장부를 공동으로 기록, 관리하는 분산화 장부 시스템이다. 분산화 장부는 은행, 국가 혹은 중개사가 거래를 관리 및 기록하는 중앙화 장부 시스템과 대조적이며, 중앙화 장부에 비해 해킹 혹은 위조 등에 강인하다. 그러나 거래내역을 전세계의 노드들에게 검증 받아야 하기 때문에 거래속도가 상대적으로 느리다는 단점이 있다. 블록체인과 비트코인은 2008년 Satoshi Nakamoto의 백서에서 소개되었으며 [1], 2019년 기준 약 60억 달러의 유통 규모를 보유한 거대 가상화폐로 성장하였다.

블록체인 기술은 합의와 분산화의 개념을 바탕으로 거래 기록의 위 변조가 불가능하도록 설계되었다. 그러나 가상화폐시

장의 성장과 함께 이중지불 공격 (double-spending attack)과 같은 보안성 공격을 통해 부당이익을 취하려는 시도가 끊임없이 존재해 왔다. 이중지불공격은 공격자가 서비스 혹은 재화의 대가로 지불한 가상화폐의 거래기록을 무효화하여 재 사용하는 공격이다. 예를 들어, 가상화폐 거래소에게 가상화폐를 지불한 대가로 현금을 출금한 후, 이러한 거래 기록을 블록체인상에서 지워버리는 것이다. 실제로 지난 2018년에는 BitcoinCash, Zcash, ZenCash, LitecoinCash와 같은 대규모 가상화폐들이 이중지불 공격의 피해를 받았으며, 그 피해액은 수백만 달러에 달한다 [2], [3], [4].

이중지불공격은 블록체인 합의 알고리즘을 역이용하면 가능하다. 블록체인 합의 알고리즘은 통신지연 등의 이유로 노드들이 보유한 블록체인이 서로 다를 때, 어느 것을 유지하고 어느 것을 버릴지를 결정하는 알고리즘이다. 비트코인의 합의 알고리즘인 longest chain consensus는 길이가 더 긴 블록체인이 더 많은 노드들에게 검증 받았으며, 따라서 더 신뢰할 수 있다고 결정한다 [1]. 결과적으로, 이중지불공격이 성공하려면 전 세계에 분포된 노드들이 공동으로 생성한 블록체인보다 길이가 더 긴 사기 블록체인을 생성하여 합의 알고리즘을 속여야 한다.

Nakamoto [1]와 Rosenfield [5]는 이러한 합의 알고리즘의 의미를 수학적으로 분석하였고, 이중지불 공격의 성공률이 100%가 되기 위해서는 전 세계의 노드들이 보유한 컴퓨터 자원보다 더 많은 컴퓨터 자원(51%)이 필요하다는 결론을 내렸다. 이중지불 공격이 51% 공격으로 불리는 이유이다. 개인 혹은 하나의 집단이 전 세계의 컴퓨터 자원보다 더 많은 자원 (51%)을 보유하는 것은 현실적으로 매우 어렵고, 따라서 비트코인이 이중지불 공격으로부터 안전하다는 주장이다. 그러나 만약 50% 미만의 컴퓨팅 자원을 사용하는 이중지불 공격, 즉 50%미만 이중지불 공격이 공격자에게 큰 이윤을 가져다 줄 수 있다면, Nakamoto와 Rosenfield의 결론은 재검토되어야 한다.

본 논문은 IEEE Transactions on Information Forensics and Security 에 제출된 논문의 수학적 정리[6]를 바탕으로 50%미만 이중지불 공격의 가능성을 시뮬레이션을 통해 보여준다. 구체적으로는 이중지불 공격의 성공률이 아닌, 이중지

불 공격이 가져다 주는 기대 이윤을 분석하였다. 이러한 관점은 Nakamoto와 Rosenfield의 결론과는 전혀 다른 결론을 가져다 주었다. 시뮬레이션을 통해 50% 미만 이중 지불 공격의 성공이 보장되지 않더라도, 공격의 대상이 되는 거래의 가치가 충분히 크면 공격자 입장에서 이중 지불 공격의 기대 이윤은 크다는 것을 보였다. 만약 한번의 공격이 실패하더라도, 계속 시도하면 결국 그 동안의 지출을 모두 상쇄하고도 남는 이윤이 돌아온다는 것이 기대 이윤의 의미다. 공격자가 이중 지불 공격을 수행하며 소요하는 지출은 공격시간 동안 소요한 컴퓨터 자원의 운용 비용뿐인데, 이를 상쇄할 만큼 가치가 큰 거래를 공격하면 결국 이윤을 기대할 수 있다는 것이다. 결론적으로, 상대적으로 적은 (50% 미만) 컴퓨터 자원을 이용하는 쉬운 이중 지불 공격도 위험하다는 것이다.

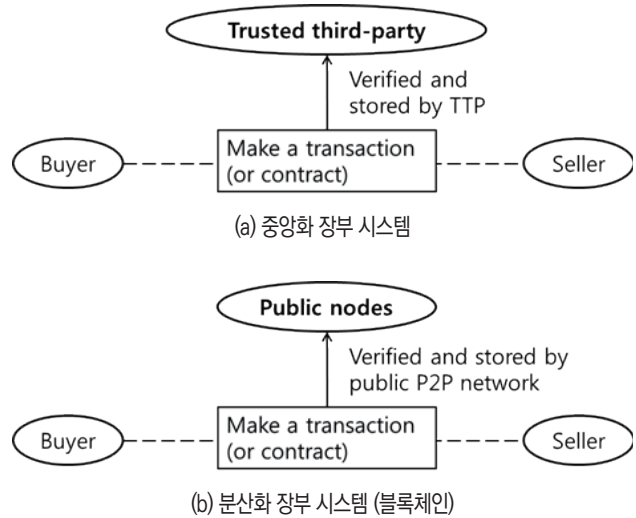
본 논문은 다음과 같이 구성되어 있다. 2장에서는 블록체인의 작동 원리를 소개한다. 3장에서는 이중 지불 공격을 소개하고 공격의 성공 조건을 정의한다. 4장에서는 이중 지불 공격의 성공률 분석에 관한 기존 연구문헌을 소개하고 분석의 한계점을 제시한다. 5장에서는 이중 지불 공격의 이윤을 분석하고 50% 미만의 컴퓨터 자원을 이용하는 이중 지불 공격도 위협적임을 실험을 통해 보인다. 마지막으로, 6장에서 요약과 함께 결론을 맺는다.

II. 블록체인

블록체인은 분산화 거래 시스템으로써, 기존 거래 시스템인 중앙화 거래 시스템과 대조적이다. 그림 1은 중앙화 거래 시스템과 분산화 거래 시스템을 비교하여 보여준다. 중앙화 거래 시스템은 공인된 3자인 trusted third-party (TTP)에 의해 거래내용이 검증 및 기록된다. 이러한 기존 방식은 TTP에 완전히 의존하고 있기 때문에, TTP가 부정한 행동을 취하거나 해킹당할 시 거래자는 금전적 피해를 받을 수 있다. 반면, 블록체인에 의한 분산화 거래 시스템은 전 세계에 분포한 풀 노드 (채굴자)에 의해 거래내용이 공동으로 검증 및 기록된다. 따라서 블록체인은 TTP에 의한 거래 시스템보다 더 신뢰할 수 있는 거래 시스템을 제안한다.

블록체인 네트워크를 구성하는 노드는 크게 거래자 노드와 채굴자 노드로 구분 할 수 있다. 거래자 노드는 거래 (transaction)을 생성하여 채굴자 노드에게 공표한다. 채굴자 노드는 검증되지 않은 거래들을 모아서 검증하고, 블록에 담는다. 이후 채굴자 노드는 블록에 담긴 거래들이 수정되지 못하게 하기 위해 작업증명을 수행한다.

작업증명의 방식은 블록체인 프로토콜마다 차이가 있으며, 본 논문에서는 비트코인 프로토콜의 작업증명 (proof-of-



〈그림 1. 중앙화 장부 시스템과 분산화 장부 시스템〉

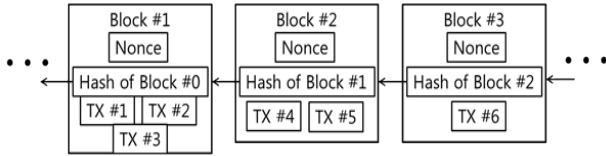
work)를 소개한다. 비트코인의 작업증명은 블록의 특별한 해쉬 (hash) 값을 찾는 것이다.

해쉬는 SHA-256 함수에 의해 1MByte 크기의 블록이 압축된 256bit 길이의 이진 문자열이다. SHA-256 해쉬 함수의 특성은 입력 블록의 이진 값 중 하나의 값이라도 수정되면, 출력되는 해쉬 값이 불규칙적으로 변한다는 것이다. 다시 말해, 누군가 악의적인 목적으로 블록의 거래내용을 수정하고 다시 해쉬 값을 계산하면, 이전에 계산된 해쉬 값과는 전혀 다른 결과를 얻는다. 이러한 SHA-256 해쉬 함수의 특성이 거래내용 위조 및 변조를 방지하기 위해 사용된다.

작업증명의 목적은 수 많은 채굴자가 하나의 블록을 함께 검증하였다는 사실을 증명하는 것이다. 이러한 목적을 달성하기 위해 비트코인 프로토콜은 채굴자에게 블록의 특별한 해쉬 값을 찾도록 요구한다. 구체적으로, 블록의 내용에 임시 값 (nonce)을 추가한 후, 프로토콜이 요구하는 조건을 만족하는 블록의 해쉬 값이 출력될 때까지 임시 값을 변경하도록 지시한다. 특별한 해쉬를 찾으면 하나의 블록을 완성하는 것이며, 블록을 완성한 노드에게는 암호화폐가 보상으로 주어진다. 특별한 해쉬 값의 조건이 어렵기 때문에, 조건을 만족시키려면 SHA-256 함수를 수없이 많이 실행하여야 한다. 즉, 조건을 만족시키기까지 오랜 시간이 소요되며, 그 시간 동안 더 많은 채굴자가 거래들을 검증 하는데 참여 할 수 있다.

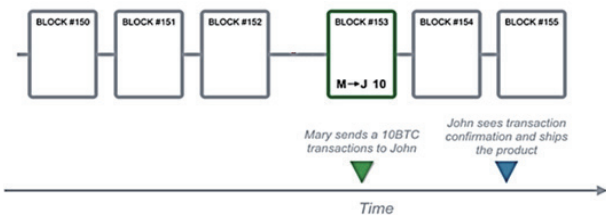
블록체인의 블록들은 서로 연결되어있다. 그림 2은 블록체인의 구조를 보여준다. 작업증명을 통해 하나의 블록이 완성되면, 그 다음 블록의 내용에는 이전 블록의 특별한 해쉬 값이 포함된다. 이러한 체인 구조는 누군가 악의적인 목적으로 이미 검증된 거래 내용을 위조 및 변조하는 것을 어렵게 만든다. 예를 들어 그림 2에서, 블록 #1에 들어있는 거래 #2를 수정하기 위해서는 먼저 블록 #1의 특별한 해쉬 값을 새롭게 찾아야

한다. 그리고 새롭게 찾은 블록 #1의 특별한 해쉬 값은 블록 #2의 해쉬 값에도 영향을 주기 때문에, 블록 #2의 특별한 해쉬 값도 새롭게 찾아야 한다. 이러한 일련의 과정을 가장 최신 블록까지 반복해야 한다. 특별한 해쉬 값을 찾는 작업증명 과정은 많은 시간을 소요한다. 따라서 혼자서 수 많은 작업증명을 완성하는 것은 매우 어렵다.



<그림 2. 블록체인의 구조>

거래가 포함된 블록 이후에 더 많은 블록이 연결된다면, 그 거래는 위-변조의 위험에 더욱 강인해진다. 이러한 이유로 거래자는 이체 확인 (block confirmation)라는 과정을 수행한다. 이체 확인은 거래를 완료하기 전에, 거래내용이 기록된 블록을 포함하여 몇 개의 블록이 더 생성되기까지 기다리는 것이다. 이때 생성을 기다리는 블록의 개수를 이체확인 수 (block confirmation number, N_{BC})라고 한다. 예를 들어, 그림 3은 Mary가 John에게 10 BTC의 암호화폐를 지불하고, John은 Mary에게 그 대가에 상응하는 제품을 제공하는 거래의 이체확인 과정을 보여준다. 이체확인 수 (N_{BC})가 2개이면, 거래가 포함된 블록과, 그 이후에 하나의 블록이 생성되는 것을 기다린다. 이체확인 수 N_{BC} 가 클수록 위-변조의 위험에 더 강인해지지만, 거래처리 속도는 느려진다.



<그림 3. 이체확인 과정의 예 (출처: Telemaximum.com)>

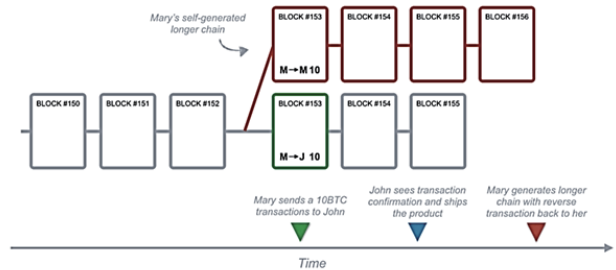
블록체인의 블록은 전 세계의 노드가 공동으로 형성하기 때문에, 네트워크 지연 등의 이유로 서로 연결되지 않는 블록들이 동시에 생성 될 수 있다. 구체적으로, 하나의 블록에 서로 다른 두 개 이상의 블록이 연결되어 체인의 갈래(fork)가 발생 할 수 있다. 비트코인 프로토콜은 여러 갈래가 존재하는 것을 허용하지 않는다. 다수의 갈래 중 하나의 갈래만 유지하는 과정을 합의 (consensus)라 한다. 합의는 여러 방식이 존재하며 [7], 비트코인의 경우 길이가 가장 긴 갈래를 유지시키는 longest chain 합의를 사용한다.

III. 이중지불 공격

이중지불 공격은 longest chain 합의를 악용하여 거래 내용을 변조하고 부당이익을 취하는 공격이다. 그림 4은 Mary (이하 M)가 John (이하 J)에게 이중지불 공격을 수행하는 과정을 보여준다. 공격에 앞서, M은 J에게 10 BTC의 암호화폐를 지급하고 재화를 공급받는 정상적인 거래를 작성 후 공표한다. 정상적인 거래는 채굴자들에 의해 검증된 후 정상블록 #153에 포함된다. 이후 채굴자들은 정상블록 #153을 잇는 또 다른 정상블록들을 지속적으로 생성한다. 한편, M은 정상적인 거래를 공표하자마자 정상적인 거래의 내용을 무효화하는 비정상적인 거래를 작성한다. 비정상적인 거래의 내용은, 예를 들어, "M이 M에게 10 BTC를 이체한다"가 될 수 있다. M은 비정상적인 거래를 채굴자들에게 공개하지 않는다. 채굴자들이 정상적인 거래가 포함된 정상블록을 생성하는 동안, M은 은밀하게 비정상적인 거래가 포함된 비공개블록 #153을 생성하고, 이를 잇는 또 다른 비공개블록들도 생성한다. M은 비정상적인 거래가 포함된 비공개 블록들의 갈래를 다음의 조건이 충족되면 공개한다.

- i. 채굴자들의 정상블록의 개수가 이체확인 수 (N_{BC}) 보다 크고,
- ii. 공격자의 비공개블록의 개수가 정상블록의 개수보다 많음

M이 비공개 블록들의 갈래를 공개하는 시점에서는, 첫 번째 조건에 의해 정상적인 거래는 완료되었고, 따라서 M은 J로부터 재화를 공급받았다. 그리고 두 번째 조건에 의해 네트워크 모든 노드의 longest chain 합의는 M이 은밀하게 개발한 갈래를 선택한다. 즉, 두 번째 조건에 의해 M이 J에게 10 BTC를 지급한다는 거래가 무효화되기 때문에 M은 10 BTC를 재사용 할 수 있다.



<그림 4. 이중지불 공격의 예 (출처: Telemaximum.com)>

IV. 이중지불 공격의 성공률 분석

이중지불 공격의 성공확률은 Nakamoto [1]와 Rosenfield [5]에 의해 계산되었다. 확률 분석을 위해, 채굴자와 공격자가 일정시간 동안 생성한 블록의 개수를 독립적인 Poisson 확률 분포[8]를 갖는 랜덤 변수들로 모델링 하였다. Poisson 확률 모델을 바탕으로, 이중지불공격의 성공을 위한 두 조건의 달성 확률 P_1 과 P_2 를 각각 계산 한 후 곱하였다.

먼저, 두 번째 조건인 ii) 비공개블록의 개수가 정상블록의 개수보다 많은 확률은 Gambler's ruin theorem [9]를 적용하여 계산하였다. 채굴자와 공격자가 보유한 컴퓨터 자원의 비를 각각 p 와 q 라 칭하겠다 ($p + q = 1$). 이중지불공격이 시작된 이후 어느 시점에서 채굴자 갈래의 정상블록 개수가 h 개이고 공격자 갈래의 비공개블록 개수가 a 개 ($h \geq a$)라 가정하겠다. 이후 무한정한 시간이 지났을 때, 공격자가 언젠가는 두 번째 조건을 달성할 확률은

$$P_2(h-a) = \begin{cases} \left(\frac{q}{p}\right)^{h-a+1}, & \text{if } p > q \text{ and } h \geq a, \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

이다.

Nakamoto와 Rosenfield의 분석 결과는 첫 번째 조건인 i) 정상블록의 개수가 이체확인 수 보다 클 확률의 계산 방법에서 차이가 있다. 이 확률을 계산하기 위해서는 채굴자가 생성한 정상블록의 개수가 이체확인 수 (N_{BC})와 같아지기까지 소요 되는 시간에 관한 확률모델이 필요하다. Nakamoto는 이 소요시간을 상수로 가정한 반면, Rosenfield는 이 소요시간을 랜덤 변수로 정의하여 보다 일반적인 방법으로 접근하였다.

구체적으로, 채굴자가 하나의 정상블록을 생성하는데 평균 D_H 의 시간이 소요될 때, 공격자의 비공개블록 생성 평균 소요시간 D_A 를 다음과 같이 계산 할 수 있다.

$$D_A = D_H \frac{p}{q}. \quad (2)$$

이는 블록생성 평균 소요시간이 공격자 혹은 채굴자가 보유한 컴퓨터 자원에 반비례한다는 가정을 바탕으로 계산되었다. Nakamoto는 공개블록의 개수가 N_{BC} 와 같아지기까지 소요된 시간을 $D_H N_{BC}$ 로 가정하였다. 따라서 $D_H N_{BC}$ 의 시간 동안 공격자가 생성한 비공개 블록의 평균 개수는 $D_H N_{BC} / D_A$ 이며, 식 (2)에 의해 이 값은 $q N_{BC} / p$ 와 같다. Poisson 확률 분포에 의해, Nakamoto가 계산한 정상

블록이 N_{BC} 개 생성되기까지 소요된 시간 동안 공격자가 생성한 비공개블록의 개수가 k 개일 확률은

$$G_1(k) = \left(\frac{q N_{BC}}{p}\right)^k \frac{e^{-\frac{q N_{BC}}{p}}}{k!} \quad (3)$$

이다. 반면 Rosenfield는 정상블록의 개수가 N_{BC} 와 같아지기까지 소요된 시간이 랜덤 변수 일 때, 그 시간 동안 공격자가 생성한 비공개블록의 개수가 k 인 확률이 negative binomial 확률 분포를 따른다는 사실을 적용하였다. 다시 말해, Rosenfield가 계산한 공개 블록이 N_{BC} 개 생성되기까지 소요된 시간 동안 공격자가 생성한 비공개 블록의 개수가 k 개일 확률은

$$P_1(k) = \binom{N_{BC} + k - 1}{N_{BC} - 1} p^{N_{BC}} q^k \quad (4)$$

이다. Rosenfield의 계산 결과를 바탕으로, 공격자가 언젠가는 이중지불공격에 성공할 확률 P_{AS} 는

$$P_{AS} = \sum_{k=0}^{\infty} P_1(k) P_2(N_{BC} - k) = \begin{cases} \sum_{k=0}^{N_{BC}} \binom{N_{BC} + k - 1}{N_{BC} - 1} p^{N_{BC}} q^k \left(\frac{q}{p}\right)^{N_{BC}-k+1} + \sum_{k=N_{BC}+1}^{\infty} \binom{N_{BC} + k - 1}{N_{BC} - 1} p^{N_{BC}} q^k, & \text{if } p < q, \\ 1, & \text{if } p \geq q \end{cases} \quad (5)$$

$$= \begin{cases} 1 - \sum_{k=0}^{N_{BC}} \binom{N_{BC} + k - 1}{N_{BC} - 1} p^{N_{BC}} q^k \left(1 - \left(\frac{q}{p}\right)^{N_{BC}-k+1}\right), & \text{if } p < q, \\ 1, & \text{if } p \geq q \end{cases}$$

이다.

식 (5)에 의해, $P_{AS} = 1$, 즉 이중지불 공격의 성공을 보장하기 위한 필요-충분 조건이 $p \leq q$ 이라는 것을 알 수 있다. 다시 말해, 공격자가 채굴자보다 더 많은 컴퓨터 자원을 보유하는 것이 공격 성공의 조건이다. 이러한 결론은 Nakamoto의 계산식에서도 마찬가지로 유도될 수 있다. Gambler's ruin theorem을 적용하여 얻은 이 결론은 공격자에게 무한정의 시간이 주어진다라는 가정이 내포되어 있다. 그러나, 공격을 시도하는 시간 동안 컴퓨터 자원을 운용하는 비용이 지속적으로 발생되기 때문에 이러한 가정은 비현실적이다. 뿐만 아니라, 공격 성공확률 100%가 아니어서 실패의 위험이 존재한다고 하더라도, 공격 성공 시 얻을 수 있는 이윤이 소요된 비용보다 훨씬 크다면 공격자의 입장에서는 공격을 시도해 볼 수 있다. 따라서 이중지불 공격의 성공률뿐만 아니라 이윤도 분석 할 필요가 있다.

V. 이중지불 공격의 이윤 분석

앞서 이중지불 공격의 성공률 분석을 통해 공격자의 컴퓨터 자원이 채굴자의 컴퓨터 자원보다 더 적을 때, 즉 $p > q$ 일 때는 이중지불 공격이 실패 할 수 있음을 확인하였다. 본 장에서는 $p > q$ 이더라도, 50% 미만 이중지불 공격이 수익성이 있으며 따라서 거래자에게는 위협적임을 Monte-Carlo 시뮬레이션을 통해 확인한다. 본 장의 내용은 [6]에서 수학적으로 증명되었다.

공격자의 컴퓨터 자원의 비율이 q 일 때, 이중지불 공격의 이윤 $F(q)$ 를 다음과 같이 정의 하였다.

$$F(q) = V(q) - C(q), \quad (6)$$

여기서 $V(q)$ 는 이중지불 공격으로부터 얻는 수익이며, $C(q)$ 는 q 만큼의 컴퓨터 자원을 운용하는데 소요된 지출이다.

지출 $C(q)$ 는 컴퓨터 자원을 운용한 시간과 컴퓨터 자원의 크기에 비례한다고 가정한다. $C(q)$ 를 계산하기에 앞서, $p > q$ 이기 때문에 이중지불 공격이 실패할 가능성이 존재함을 주의해야 한다. 만약 이중지불 공격이 실패하면, 컴퓨터 자원을 운용하는 시간이 무한정 늘어나며, 따라서 지출 $C(q)$ 도 무한대로 발산한다. $p > q$ 인 경우에서 이러한 무한대의 지출을 방지하기 위해, 중단 시간 (cut time) t_{cut} 을 정의한다 [참조문헌 6의 Theorem 7]. 공격자는 t_{cut} 의 시간 내에 이중지불 공격에 성공하지 못할 경우, 지출의 발산을 방지하기 위해 공격을 중단한다. 만약 t_{cut} 내에 공격이 성공할 경우, 지출 $C(q)$ 는 공격 성공 시간 동안 소요된 컴퓨터 자원 운용 비용이다. 공격 성공 시간은 불확정적이기 때문에 랜덤 변수 T_{AS} 로 모델링 될 수 있다. 컴퓨터 자원의 크기는 공격자의 시간당 평균 블록 생성량, 즉 D_A^{-1} 에 비례한다고 가정한다. 종합하면, 공격자의 컴퓨터 자원 비율이 q 이고 중단 시간이 t_{cut} 일 때의 지출 $C(q)$ 은

$$C(q) = \begin{cases} \gamma D_A^{-1} T_{AS}, & \text{if attack succeeds,} \\ \gamma D_A^{-1} t_{cut}, & \text{otherwise} \end{cases} \quad (7)$$

이며, 여기서 γ 는 하나의 블록을 생성하는데 소요되는 평균비용이다.

식 (7)의 지출을 계산하기 위해서는 파라미터 D_A 와 γ 가 필요하다. D_A 는 채굴자의 평균 블록생성 주기인 D_H 로부터 (2)번 식을 통해 계산이 가능하다. 파라미터 D_H 와 γ 는 공격 대상 블록체인 네트워크에 따라 다르며, 그 값은 인터넷에

공개된 정보로부터 얻을 수 있다. 본 논문에서는 BitcoinCash 네트워크를 예로 들겠다. BitcoinCash의 평균 블록생성 주기는 $D_H = 600$ 초로 네트워크 개발자에 의해 고정되어있다. 채굴자가 보유한 컴퓨터 자원의 크기가 변동하면, 채굴 알고리즘에 D_H 가 유지되도록 채굴 난이도가 변경된다. 또 다른 파라미터인 γ 의 값은 컴퓨터 자원 대 서비스 제공을 제공하는 업체인 nicehash.com에 의해 결정 될 수 있다. nicehash.com에 따르면, 2018년 12월을 기준으로 BitcoinCash에서 하나의 블록을 채굴하는데 소요되는 비용은 $\gamma = 0.33$ 비트 코인(BTC)이다.

수익 $V(q)$ 는 이중지불 공격이 성공 할 경우 이중지불 공격의 대상이 되는 거래의 가치 v 와 같으며, 공격이 실패할 경우 0이다. 공격이 성공 할 확률은 공격자의 컴퓨터 자원 q 에 영향을 받는다. 수식으로는,

$$V(q) = \begin{cases} v, & \text{if attack succeeds,} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

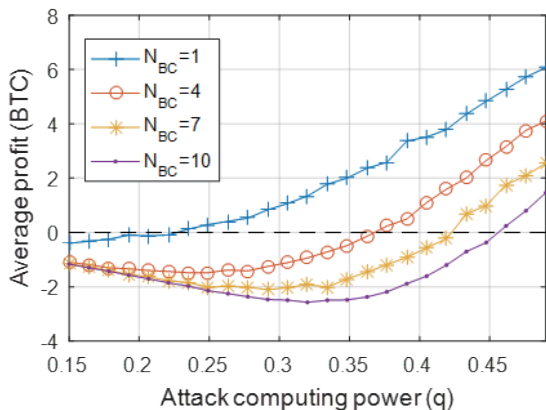
으로 표현 될 수 있다. 공격 대상 거래의 가치는 공격자가 상대 거래자 (피해자)와의 합의를 통해 함께 결정한다.

식 (6)의 이윤을 계산하기 위해 남은 변수는 랜덤 변수인 공격성공 소요시간 T_{AS} 이다. 랜덤 변수 T_{AS} 의 정확한 확률 분포는 [참조문헌 6의 Proposition 4]에서 계산되었으며, 본 논문에서는 T_{AS} 를 Monte-Carlo 실험으로 측정하였다. 실험을 위해, MATLAB상에서 두 개의 독립적인 Poisson counting process (PCP)를 구현하였다. 두 개의 PCP는 각각 공격자와 채굴자가 생성한 블록들의 생성 시간을 나타내며, 평균 블록생성 시간에 영향을 받는다. 공격자 PCP의 평균 블록생성 주기는 D_A 이며 채굴자 PCP의 평균 블록생성 주기는 D_H 이다. 중단 시간 t_{cut} 까지의 두 PCP의 블록생성 시간들을 실현 (realization) 한 후, 두 PCP를 비교하여 이중지불 공격의 두 가지 성공 조건이 달성유무를 판단하였다. 식 (7)의 지출을 계산하기 위해, 만약 이중지불 공격이 성공하였다면, PCP로부터 성공 시간 T_{AS} 를 추출하여 대입하였으며, 만약 공격이 실패하였다면, t_{cut} 을 대입하였다. 마찬가지로, 식 (8)의 수익을 계산하기 위해 만약 이중지불 공격이 성공하였다면, $V(q) = v$ 로 계산하였고, 그렇지 않을 경우 $V(q) = 0$ 로 계산하였다. 이러한 일련의 과정을 5000번씩 반복한 후 계산 결과들에 평균을 취하였다.

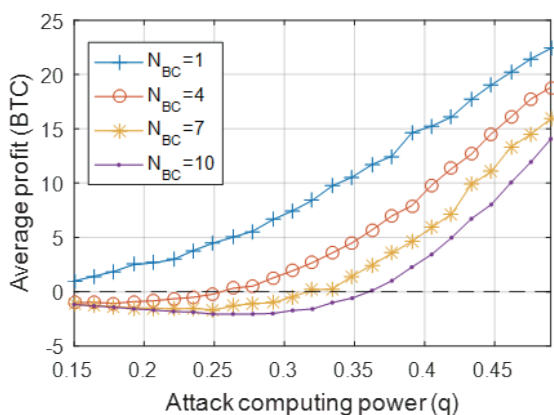
그림 5는 이중지불공격 실험 결과를 보여준다. 실험에 사용된 블록체인 네트워크 파라미터는 블록생성 평균비용 $\gamma = 0.33$ BTC와 블록생성 평균주기 $D_H = 600$ 초이며, 이는 2018년 12월 기준의 BitcoinCash 네트워크의 파라미터와

같다. 중단 시간 $t_{cut} = 12000$ 초로 설정하였다. 그래프는 다양한 이체확인 수 (N_{BC})와 다양한 공격대상 거래가치 (v)에 대해 이중지불 공격의 표본평균 이윤을 공격자의 컴퓨터 자원 (q)의 함수로 나타내었다.

그림 5로부터 공격자의 컴퓨터 자원이 네트워크 채굴자의 컴퓨터 자원보다 적은 경우, 즉 $q < p$ 인 경우에도 이중지불 공격에서 이윤을 기대할 수 있음을 알 수 있다. 다시 말해, 50% 이상의 컴퓨터 자원을 사용하는 이중지불공격뿐만 아니라 50% 미만의 컴퓨터 자원을 사용하는 이중지불공격도 위협적이다. 그래프는 이중지불 공격에 이윤을 가져다 주기 위한 공격대상의 거래가치 (v)의 요구조건이 이체확인 수 (N_{BC})가 커질수록 증가함을 보여준다. 즉, 가치가 큰 거래가 50% 미만 이중지불공격으로부터 안전하기 위해서는 충분히 큰 N_{BC} 가 필요함을 보여준다. 수익성이 있는 50%미만 이중지불공격을 위한 공격대상 거래가치에 대한 정확한 조건은 [참조문헌6의 Theorem 8]에서 계산되었다.



(a) 거래가치 BTC



(b) 거래가치 BTC

〈그림 5 다양한 거래가치와 이체확인 수에 대한 50% 미만 이중지불 공격의 평균 이윤〉

VI. 결론

이중지불공격에 관한 기존 문헌들은 공격자의 컴퓨터 자원이 네트워크 채굴자의 컴퓨터자원보다 더 클 때, 즉 공격자가 전체 컴퓨터 자원의 50% 이상을 보유해야만 이중지불 공격이 성공 할 수 있다는 것을 보였다. 이러한 이유로 이중지불 공격은 51% 공격으로 알려져 왔다. 반면 본 논문에서는 50% 미만의 컴퓨터 자원을 사용하는 이중지불공격, 즉 50%미만 공격의 위험성을 분석하였다. 대규모 시뮬레이션을 통해 50% 미만의 컴퓨터 자원을 사용하는 이중지불공격도 공격자에게 큰 이윤을 가져다 줄 수 있음을 보였다. 구체적으로는, 거래자가 설정하는 이체확인 수가 작을수록 이중지불공격의 이윤이 커짐을 보였다. 다시 말해, 이체확인 수가 작을수록 거래의 처리속도는 빠르지만 50%미만 공격에는 매우 취약하다. 본 논문의 실험 결과는 블록체인의 거래자가 거래 가치와 이체확인 수를 결정하는 것에 관한 가이드라인을 제공 할 수 있다.

Acknowledgement

이 논문은 2019년도 광주과학기술원의 재원으로 GRI(GIST연구원) 사업의 지원을 받아 수행된 연구임. 이 논문은 2019년도 광주과학기술원의 재원으로 “과학기술응용연구단의 실용화 연구개발사업”의 지원을 받아 수행된 연구임.

References

- [1] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System.” [Online]. Available: <https://bitcoin.org/bitcoin.pdf>.
- [2] C. Osborne, “Bitcoin Gold suffers double spend attacks, \$17.5 million lost,” ZDNet, 25-May-2018. [Online]. Available: <https://www.zdnet.com/article/bitcoin-gold-hit-with-double-spend-attacks-18-million-lost>.
- [3] “ZenCash Statement on Double Spend Attack,” Horizen, 03-Jun-2018. [Online]. Available: <https://blog.zencash.com/zencash-statement-on-double-spend-attack/>.
- [4] A. Hertig, “Blockchain’s Once-Feared 51% Attack Is Now Becoming Regular,” CoinDesk, 08-Jun-2018. [Online]. Available: <https://www.coindesk.com/blockchains-feared-51-attack-now-becoming-regular/>.
- [5] M. Rosenfeld, “Analysis of Hashrate-Based Double Spending,” arXiv:1402.2009 [cs], Feb. 2014.

- [6] J. Jang and H.-N. Lee, "Profitable Double-Spending Attacks," submitted to IEEE Transactions on Information Forensics and Security, Mar. 2019. Available: arXiv:1903.01711 [cs]
- [7] G.-T. Nguyen and K. Kim, "A Survey about Consensus Algorithms Used in Blockchain," Journal of Information Processing Systems, vol. 14, no. 1, pp. 101 - 128, 2018.
- [8] A. Papoulis and S. U. Pillai, "Random walks and other applications," in Probability, Random Variables and Stochastic Processes, 4th edition., Boston, Mass.: McGraw-Hill Europe, 2002.
- [9] W. Feller, "Random walk and ruin problems," in An introduction to probability theory and its applications, New York: Wiley, 1968.

Biographies



장재혁

2014년: 금오공과대학교 전자공학 학사

2016년: 광주과학기술원 정보통신공학 석사

2016년~현재: 광주과학기술원 전기전자컴퓨터공학 박사과정

관심분야: 블록체인, 신호 및 시스템, 압축센싱, 레이더
E-mail: jjh2014@gist.ac.kr



Heung-No Lee

Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He then worked at HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member from 1999 to 2002. From 2002 to 2008, he worked as an Assistant Professor at the University of Pittsburgh, PA, USA. In 2009, he then moved to the School of Electrical Engineering and Computer Science, GIST, Korea, where he is currently affiliated. His areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He has received several prestigious national awards, including the Top 100 National Research and Development Award in 2012, the Top 50 Achievements of Fundamental Researches Award in 2013, and the Science/Engineer of the Month (January 2014).
E-mail: heungno@gist.ac.kr

Introduction to Error-Correction Codes Proof-of-Work

박상준, 김형성,¹이흥노
{sjpark1, hyoungsung, heungno}@gist.ac.kr
광주과학기술원

1. Introduction

2008년 신원 미상의 사토시 나카토모는 백서[1]를 통해 탈중앙화된 - 즉 신뢰받는 기관이 없는 - 화폐 시스템을 소개하고, Peer-to-Peer (P2P) 네트워크를 통해 구현하였다. 이것이 최초의 암호화화폐인 비트코인이고, P2P 네트워크에서의 이중 지불을 막기 위해 블록체인을 사용하였다.

블록체인이란 블록들이 하나의 체인으로써 연결된 것을 뜻한다. 각 블록들은 거래 내역(데이터)을 담고 있고, P2P에 존재하는 불특정 다수 노드들에 의해 검증 받는다. 검증하는 노드들을 채굴자라 일컫고, 이 검증과정을 작업증명이라 한다.

작업증명의 목적은 다수의 채굴자들이 하나의 블록을 채굴 - 즉 검증 - 하기 위해 많은 노력을 했다는 것을 입증하기 위함이다. 비트코인의 경우, SHA256 (Secure Hash Algorithm) 함수의 특정 해시 값을 산출하게 하는 nonce를 찾음으로써 작업증명이 완료된다. SHA256의 출력 값을 통해 역으로 입력 값을 알아내는 것이 불가능하므로, 채굴자들은 무차별적으로 nonce들을 대입해야 한다. SHA256의 출력 값인 해시 값은 블록 간의 연결을 위해 사용된다. 이전 블록의 해시 값을 현재 블록 내역에 포함시킴으로써 인접한 블록들을 연결시킨다. 이와 같이 연결하여 채굴된 블록들의 위·변조를 어렵게 만들고, 이를 통해 이중 지불을 막는다.

초창기 비트코인의 작업증명은 CPU를 통해서 이뤄졌다. 이 시기는 사토시가 백서에서 언급한 것처럼, 어느 누구나 CPU만 있다면 공정한 채굴 경쟁이 가능하였다. 비트코인이 세상에 알려지고, 채굴 이윤이 발생함에 따라 채굴 경쟁이 시작되었고, 2010년, 2013년 각각 GPU와 ASIC (Application-Specific Integrated Circuit) 채굴 장비들이 등장하였

다. ASIC 채굴의 성능은 CPU/GPU 보다 월등하였기 때문에, 채굴 난이도의 급격한 상승을 야기하였다. 결국 CPU/GPU 채굴자들은 더 이상 이윤 창출이 불가능해졌고, 오늘날의 비트코인 채굴은 ASIC을 통해 이뤄진다.

ASIC 채굴로 전환됨에 따라 일반 사람들 혹은 자본이 적은 사람들은 채굴에서 배제되고, 막대한 자본력을 가지고 있는 소수의 집단들이 채굴을 독점하였다. 이 집단들이 전 세계의 채굴 능력을 많이 점유하면 (극단적인 예로써, 51%이상), 이중 지불 등을 수행하여 악의적인 용도로 블록들을 채굴하고, 채굴된 블록들을 위·변조할 가능성이 존재하게 되는 것이다. ASIC 채굴을 억제하기 위한 새로운 작업증명들 [3][4]이 제안됐지만, 결국 ASIC 채굴 장비들이 등장하였다.

오류-정정 부호 [6]는 무선 통신에서 발생하는 오류를 정정하기 위해 사용된다. 대표적인 부호들 중 하나로 LDPC [5] 부호가 있다. 문헌에 따르면, LDPC 디코더의 ASIC 구현은 구조적/비용적 문제로 인하여, 구현의 유연성이 떨어진다 [7]. LDPC 디코더와 해쉬 함수를 결합한 오류-정정 부호 기반의 작업증명 [2] (ECCPoW, Error-Correction Codes Proof-of-Work)을 제안하였다. 본 특집호의 목표는 ECCPoW의 동작과정과 ASIC 채굴 장비의 등장을 어떻게 억제하는지 설명하는 것이다.

본 특집호는 다음과 같이 구성되었다. 2장에서는 ASIC 장비 등장을 억제하기 위한 작업증명들을 소개하고, 이들의 한계점을 보여준다. LDPC 부호 및 디코더에 관한 문헌 결과를 보고한다. 3장에서는 ECCPoW의 동작과정과 ASIC 채굴 장비 등장 억제의 요인을 설명한다. 4장에서는 확률적 분석을 통해 ECCPoW의 작업증명 완료가 쉽지 않다는 것을 보여준다. 5장에서 본 특집호의 결론을 제시한다.

2. Background

본 섹션에서는 ASIC 채굴 장비 등장을 억제하기 위한 작업증명들 (Ethash와 X11)과 한계점들을 소개한다. LDPC 부호 및 디코더 소개와, ASIC 디코더에 대한 문헌들을 제공한다. 이 문헌들을 제공하는 이유는 ECCPoW의 ASIC 채굴 장비 억제 기능이 LDPC 디코더에서 기인하기 때문이다.

2.1. Ethash [3] and X11 [4]

¹ 교신저자

이더리움의 작업증명인 Ethash [3]는 비선형 그래프 (DAG, Directed Acyclic Graph)를 이용하여 ASIC 채굴 장비 등장을 억제하였다. 이 DAG는 30,000 블록 단위로 무작위로 생산되는 데이터들의 집합으로써, 2019년 5월 기준, DAG의 크기는 약 3기가 바이트이다.

표 1은 Ethash의 동작과정을 보여준다. Nonce와 BH (블록 헤더)를 활용하여 해쉬 값을 생산하고, 이 값은 mix0에 저장한다. 이 과정이 Step 2이다. Step 4에서는 DAG로부터 데이터 (data1)를 읽어온다. 읽어올 데이터 위치는 mix0에 의해 결정된다. Step 5에서는 data1과 mix0를 mixing 함수에 넣고 얻은 결과를 mix0에 저장한다. Step 4부터 5까지 총 63번 반복 수행하고, 최종적으로 얻은 mix0를 사용해 작업증명 완료 유무를 판단한다.

63번의 반복 수행에서의 mixing 함수는 ASIC을 통해 처리 가능하다. 데이터 읽기 연산은 ASIC과 무관하다. 더욱이, 어떤 데이터를 읽어오는지 미리 알 수 없고, DAG의 크기가 너무 크기 때문에 캐시를 이용해 빠르게 읽는 것 또한 불가능하다. 따라서, 데이터 읽기와 mixing 함수 실행 사이에서 병목현상이 발생한다. 이 병목으로 인해, ASIC 채굴 장비를 사용할 필요가 없던 것이었다. 하지만 병목이 해결되면, ASIC 채굴 장비를 활용해 좀 더 빠른 채굴이 가능해진다. 문헌 조사에 따르면, 2018년 7월 비트메인은 ASIC 장비를 공포하였다.

대쉬의 작업증명인 X11 [4]은 다수의 해쉬 함수들을 사용하여 ASIC 채굴 장비 등장을 억제하였다. 이 X11은 다음 함수들이 순차적으로 사용한다:

Blake, Bmw, Groestl, Jh, Keccak, Skein, Luffa, Cubehash, Shavite, Simd and Echo

표 2는 X11의 동작과정을 보여준다. Nonce와 BH를 이용하여 Blake의 해쉬 값을 얻는다. 얻어진 값을 Bmw의 입력 값으로 사용한다. 이 과정을 반복하여 최종적으로 Echo의 해쉬 값을 얻고, 이 해쉬 값을 통해 작업증명 완료 유무를 판단한다.

X11에서 사용되는 해쉬 함수들의 순서는 고정이다. 따라서 ASIC 채굴 장비를 개발하려면, 함수들을 구현하고, 연결 하면 된다. 2014년과 2016년 사이에는 하드웨어 생산 비용 문제로, ASIC 채굴 장비 개발이 억제되었다. 하지만, 공정 기술 발달로 저 비용 생산이 가능해짐에 따라, ASIC 채굴 장비

표 1. Ethash 의사 결정 코드

Inputs: BH, L and DAG	
Step 1:	for nonce = 0, 1, 2, ... $2^{32} - 1$
Step 2:	mix0 = (SHA3(nonce, BH))
Step 3:	for i = 1, 2, ..., 63
Step 4:	data1 = Fetch(DAG, mix0)
Step 5:	mix0 = Mixing(mix0, data1)
Step 6:	end
Step 7:	If mix0 begins with L zero bits, then break.
Step 8:	end

BH는 블록 헤더, L은 주어진 난이도. 해당 의사 결정, 코드는 본 연구팀의 논문인 [2]로부터 인용.

표 2. X11 의사 결정 코드

Inputs: BH and L	
Step 1:	for nonce = 0, 1, 2, ... $2^{32} - 1$
Step 2:	e = Blake(nonce, BH)
Step 3:	e = Bmw(e)

Step 12:	e = Echo(e)
Step 13:	If e begins with L zero bits, then break.
Step 14:	end

BH는 블록 헤더, L은 주어진 난이도. 해당 의사 결정, 코드는 본 연구팀의 논문인 [2]로부터 인용.

가 2016년부터 판매되고 있다.

X11을 확장하여 X13, X14, X15 그리고 X17 작업 증명들이 제안되었다. 이름에서 알 수 있듯이, 별도의 함수들을 추가적으로 사용하여 ASIC 채굴 장비 등장을 억제하는 것이다. 2019년, X17을 제외한 작업증명들의 ASIC 채굴 장비가 판매되고 있다.

2.2. LDPC 부호와 디코드

대표적인 오류 정정 부호 중 하나인 LDPC [5] 부호는 대부분의 원소 값이 1인 패리티 체크 행렬 $\mathbf{H} \in \{0,1\}^{m \times n}$ (PCM, parity check matrix)를 이용해 정의된다. 구체적으로, PCM이 주어졌을 때, 다음 조건을 만족시키는

$$\mathcal{C} := \{\mathbf{c} | \mathbf{H}\mathbf{c} = \mathbf{0} \cap \mathbf{c} \in \{0,1\}^{m \times 1}\}$$

벡터들 $\mathbf{c} \in \{0,1\}^{m \times 1}$ 의 집합이 LDPC 부호이다.

PCM을 이용해 LDPC 부호를 이분 그래프로 표현 할 수 있다. 이 그래프는, 변수 (variable) 및 체크 (check) 노드들과 이들을 연결하는 선으로 구성된다. 변수/체크 노드들은 PCM의 열/행에 각각 대

응한다. PCM의 (i, j) 번째 원소 값이 1이면 i 번째 변수 노드와 j 번째 체크 노드가 연결된 것을 뜻한다.

LDPC 부호의 성능 – 얼마나 많은 오류들을 고치는지 – 은 PCM의 최소 해밍 거리 d (minimum hamming distance)에 의해 결정된다. 이 값은 PCM을 통해 생성할 수 있는 $\mathbf{0}$ 벡터를 제외한 모든 부호들 중에 가장 적은 해밍 값이다:

$$d = \min_{\mathbf{u} \in \mathcal{C}, \mathbf{u} \neq \mathbf{0}} \|\mathbf{u}\|_h$$

여기서 $k = n - m$, \mathbf{c}_i 는 i 번째 부호, 부호의 개수는 총 2^k , 그리고 벡터 \mathbf{x} 의 해밍 값은 다음과 같다:

$$\|\mathbf{x}\|_h := \mathbf{x} \text{에 포함된 1의 개수.}$$

PCM의 최소 해밍 거리 d 가 주어지면, LDPC 부호를 활용해 정정할 수 있는 bits 오류들의 숫자는 다음과 같이 결정된다:

$$t = \lfloor (d-1)/2 \rfloor \quad (1)$$

여기서 $\lfloor x \rfloor$ 는 x 의 정수를 표시한다. 위의 결과의 유도 과정은 [6]에 있다.

LDPC 부호가 무선 채널을 통해 전송되면, 채널에 존재하는 잡음으로 인하여 오류가 발생한다. 잡음에 의해 왜곡된 부호 \mathbf{y} 는 다음과 같이 표현할 수 있다:

$$\mathbf{y} = \mathbf{c} + \mathbf{e}$$

여기서 $\mathbf{e} \in \{0,1\}^m$ 는 잡음에 의한 오류 벡터이고 그리고 \mathbf{c} 는 전송된 부호이다. LDPC 디코더의 목적은 오류를 정정하여, 원 부호 \mathbf{c} 를 찾는 것이다. 이 디코더는 일반적으로 메시지 전달 (message passing) [6] 알고리즘을 사용한다. 이 알고리즘은 변수/체크 노드들이 서로 메시지를 반복적으로 주고 받으며 원 부호를 찾기 위해 노력한다.

일반적으로, 알고리즘들을 빠른 속도 및 저전력으로 실행시키기 위해, ASIC 장치를 사용한다. 따라서, LDPC 디코더 또한 ASIC을 사용해 구현된다. ASIC-LDPC 디코더에서는 변수 및 체크노드들이 PCM에 따라 물리적으로 연결된다. 따라서, 부호의 길이의 변화에 따라 노드들을 늘리거나 혹은 PCM 변화에 따라 유동적으로 노드들의 연결을 재 설정

하는 것이 쉽지가 않기 때문에, 다수의 PCM들을 지원하는 ASIC-LDPC 디코더 구현은 어렵다.

최신 리뷰 논문[7]에 따르면, 추가적인 하드웨어 장치들을 이용하면 다수의 PCM들을 지원하는 ASIC-LDPC 디코더를 구현할 수 있다고 보고하였다. 하지만, 그로 인해 디코더 면적 혹은 생산 비용이 증가되는 문제가 발생한다고 보고했고, 그 사례로써 [8]에서 구현된 ASIC-LDPC 디코더를 소개했다. 이 디코더는 약 100 개의 PCM들을 지원하지만, 추가적인 장치들이 디코더 면적의 약 75%를 점유하는 문제를 가지고 있다. 더 많은 PCM들을 지원하려면 더 많은 장치들이 사용되고, 그로 인해 ASIC-LDPC 구현이 매우 비효율적이다.

마지막으로, [7]에는 ASIC-LDPC 디코더의 구현 사례들이 표로 제시되어있다. 이 표에서 부호의 길이가 가장 긴 경우는 $n = 64,800$ 로써, 해당 디코더는 [9]에 구현되어있다.

3. 오류-정정 부호 기반의 작업증명

본 섹션에서 ECCPoW 동작과정을 간단히 소개하고, ASIC 채굴 장비 등장 of 억제 요인을 설명한다. ECCPoW에 관한 자세한 설명 및 이론적 분석 결과들은 [2]에 있다. 원활한 설명을 위해 용어들을 다음과 같이 축약한다. 현재 블록 헤더와 이전 블록 헤더를 각각 CBH (current block header)와 PBH (previous block header)로 사용한다.

먼저, ECCPoW에 포함된 LDPC 디코더와 그 입력 값을 다음과 같이 각각 정의한다.

정의 1. 크기가 $m \times n$ 인 PCM \mathbf{H} 와 길이가 n 인 해쉬 벡터 \mathbf{r} 이 주어졌다 가정한다. LDPC 디코더는 \mathbf{H} 와 \mathbf{r} 을 취득하고, 메시지 전달 알고리즘을 사용해 길이가 n 인 벡터 $\hat{\mathbf{c}}$ 을 산출한다:

$$\mathcal{D}_{MP} : \{\mathbf{r}, \mathbf{H}\} \mapsto \hat{\mathbf{c}} \in \{0,1\}^{n \times 1}. \quad (2)$$

정의 2. 길이가 n 인 해쉬 벡터 \mathbf{r} 은 다음과 같이 정의된다:

$$\mathbf{r} := \begin{cases} \mathbf{s}_l[1:n] & \text{if } n \leq 256 \\ \begin{bmatrix} \mathbf{s}_l & \cdots & \mathbf{s}_l & \mathbf{s}_{l+1}[1:j] \end{bmatrix} & \text{if } n > 256 \end{cases} \quad (3)$$

여기서 $l = \lfloor n/256 \rfloor$, $j = n - 256 \times l$,

$$\mathbf{s}_1 := \text{SHA256}(\text{nonce}, \text{CBH}) \in \{0,1\}^{256} \quad (4)$$

그리고, $u = 2, 3, \dots, l+1$ 에 대해

$$\mathbf{s}_u := \text{SHA256}(\mathbf{s}_1) \in \{0,1\}^{256}. \quad (5)$$

하나의 블록을 채굴하기 위한 작업증명에서는, CBH와 PCM은 모두 상수로써 취급된다. Nonce가 변경되면, 정의 2에 나온 것처럼 해쉬 벡터가 재 생성되고, 그로 인해 디코더의 출력 값이 변경된다. Nonce와 디코더의 입력 값인 해쉬 벡터의 관계는 SHA256 함수들에 결정된다. 따라서 nonce만 보고 디코더의 출력 값이 무엇인지 미리 예측하는 것은 불가능하다. 특정 조건을 만족하는 디코더의 출력 값을 찾으려면 무수히 많은 nonce들을 대입해야 한다. 결론적으로 ECCPoW의 작업증명은 디코더의 출력 값 $\hat{\mathbf{c}}$ 이 다음 조건

$$\mathbf{H}\hat{\mathbf{c}} = \mathbf{0} \quad (6)$$

을 만족하게 하는 nonce를 찾으면 완료된다.

ECCPoW가 어떻게 동작하는지 살펴보았다. 이제 LDPC 디코더의 입력 값으로써 활용되는 PCM에 대해 논의한다. 블록들을 채굴 할 때, 하나의 PCM 사용을 가정한다. 섹션 2에서 이야기 한 것처럼, 이 가정에서는 ASIC-LDPC 디코더 구현에 아무런 문제가 없다. 하지만, 이 가정하에서는, ECCPoW를 위한 ASIC 채굴 장비가 등장 할 수 있다.

이제, 매 블록 채굴 할 때 마다 무작위로 생성된 PCM 사용을 가정한다. 섹션 2에서 언급한 것처럼, 다수의 PCM을 위한 ASIC-LDPC 디코더 구현에는 추가적인 하드웨어 장치들이 필요하다. 더욱이, 각 PCM들이 무작위로 생성된다. 따라서 어떤 형태로 생성될지 예측 할 수가 없다. 매 블록마다 변하는 PCM을 사용하면, 결국 ASIC-LDPC 디코더 구현을 억제할 수 있다. 이것이 ECCPoW에서 ASIC 채굴 장비 등장을 억제하는 이유이다.

이제 부호의 길이에 대해 논의를 해보자. 부호의 길이인 n 인 경우, ASIC-LDPC 구현을 위해 필요한 computing fabric의 총량은 n 의 제공에 비례한다. 가령 n 을 2배 키우면, 필요한 총량은 4배이다. 또한, n 은 가변적으로 변하는 값이다. 무수히 많은 PCM을 지원하는 ASIC-LDPC 구현에 성공하더라도, n 을 크게 키움으로써 구현된 ASIC-LDPC 디코더의 사용을 막을 수 있다.

어떻게 하면 매 블록마다 무작위로 변하는 PCM을 생성 할 수 있을까? 우리의 해답은 PBH의 해쉬 값과 Gallager의 PCM 생성 방법 [5]을 동시에 이용하는 것이다. 이 방법은 사용하면, 임의의 seed 값이 주어졌을 때, PCM을 결정적으로 생성할 수 있다. 즉, 여러 사람들이 동일한 seed 값을 가지고 있으면, 동일한 PCM을 생성 할 수 있는 것이다. 그리고, PBH의 해쉬 값은 매 블록 마다 바뀌고, 이미 채굴된 블록이므로, 알려진 값이다. 따라서, 이 값을 seed로 활용함으로써, 매 블록 마다 무작위로 PCM을 생성 할 수 있다. [2]에 우리의 PCM 생성 방법의 의사 결정 코드 및 좀 더 자세한 설명들을 기록하였다.

4. 해쉬 사이클 분석

이 섹션에서는 ECCPoW 작업증명을 푸는 것이 쉽지 않다는 것을 보인다. 이를 다음 아래에 해쉬 사이클을 정의한다.

정의 3. 하나의 nonce을 이용하여, LDPC 디코더의 출력 값을 산출하고, 이 값을 이용해 작업증명의 완료 유무를 검증하는 것까지 포함한 과정을 해쉬 사이클로 정의한다.

이제 작업증명을 완료하기 위해서 총 몇 번의 해쉬 사이클이 필요한지 고려한다. 이 섹션의 분석 결과들은 [2]에서 일부 밝혀진 것이다.

용이한 분석을 위하여 다음 2가지를 가정한다. 첫 번째로, 디코더는 이상적(optimal)이다. 즉, 섹션 2에서 나온 것처럼 t 개 이하의 오류가 발생시 항상 정정 가능한 것을 가정한다. 두 번째로, 해쉬 벡터와 nonce는 서로 1:1 관계로 가정한다. 즉, 다른 nonce들은 각각 다른 해쉬 벡터를 생성한다.

이제 i 번째 부호가 주어졌을 때, 해당 부호와의 해밍 거리가 t 보다 작은 벡터들의 집합을

$$\mathcal{A}(\mathbf{c}_i, t) := \{\mathbf{r} : \|\mathbf{r} - \mathbf{c}_i\|_h \leq t\} \quad (7)$$

로 정의한다. 이 집합의 크기는 다음과 같다

$$|\mathcal{A}(\mathbf{c}_i, t)| = 1 + \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{t} = \sum_{l=0}^t \binom{n}{l}$$

여기서 $i = 1, 2, \dots, 2^k$.

$$Iter \times O(n \log n) \times p^{-1}$$

첫 번째 가정으로 인하여, 해쉬 벡터가 $\mathcal{A}(\mathbf{c}_i, t)$ 의 원소라면, 출력 값이 항상 i 번째 부호이다. 즉,

$$\mathcal{D}_{MP} : \{\mathbf{r}, \mathbf{H}\} \mapsto \mathbf{c}_i.$$

따라서, i 번째 부호를 산출할 확률은 다음과 같다

$$\begin{aligned} \Pr\{\hat{\mathbf{c}} = \mathbf{c}_i\} &= 2^{-n} |\mathcal{A}(\mathbf{c}_i, t)| \\ &= 2^{-n} \sum_{l=0}^t \binom{n}{l} \\ &= 2^{-n} \sum_{l=0}^{\lfloor \frac{d-1}{2} \rfloor} \binom{n}{l} \end{aligned}$$

여기서 마지막 등호는 (1)에서 기인한다. 디코더의 출력 값이 임의의 부호일 확률은 다음과 같다:

$$\begin{aligned} p &:= \Pr\{\mathbf{H}\hat{\mathbf{c}} = \mathbf{0}_m\} \\ &= \sum_{i=1}^{2^k} \Pr\{\hat{\mathbf{c}} = \mathbf{c}_i\} \\ &= 2^{k-n} \sum_{l=0}^{\lfloor \frac{d-1}{2} \rfloor} \binom{n}{l}. \end{aligned} \quad (8)$$

LDPC 디코더가 부호를 출력할 확률이 p 이므로, 이것의 역수는 부호를 출력하기 위해 필요한 평균 해쉬 사이클이다. 따라서, p 를 알면 어느 정도 해쉬 사이클이 필요한지 계산 가능하다. p 를 계산하려면, d 를 알아야 한다. 임의의 주어진 PCM의 d 를 찾는 것은 어렵다. 우리는 편의를 위해 d 를 n 의 10%로 가정한다.

첫 번째로, $n = 64, m = 32$ 라 하자. 이 경우 p 는 약 2×10^{-10} 이다. 이 작업증명을 완료하기 위해 필요한 해쉬 사이클은 p 의 역수인 4×10^9 이다. 두 번째로, n 과 m 을 각각 128과 64로 하면, p 는 5×10^{-20} 이다. 따라서, 2×10^{19} 해쉬 사이클이 필요하다. 이것들은 작업증명 완료를 위해 많은 해쉬 사이클이 필요한 것을 보여준다.

마지막으로, LDPC 디코더의 총 연산량은

$$Iter \times O(n \log n)$$

여기서 n 은 부호의 길이이고, $Iter$ 은 메시지 전달 알고리즘의 반복 횟수이다. 각 nonce마다 디코더를 실행하므로, 평균적으로 작업증명을 완료하기 위해 사용되는 연산량은 다음과 같다:

이것을 ECCPoW의 채굴 난이도로 여길 수 있다.

4. 결론

본 특집호에서는 블록체인 커뮤니티에서 많이 주목 받고 있는 ASIC 채굴 장비 등장으로 인한 중앙화 문제를 고찰하였다. 이 문제를 해결하기 위해 제안된 오류-정정 부호 기반의 작업증명 [2] (ECCPoW, Error-Correction Codes Proof-of-Work)를 소개하였다. 이 방법의 핵심은 기존 SHA256함수와 LDPC 디코더를 연결한 것이다. SHA256의 출력 값이 디코더의 입력 값이 되고, 이 디코더의 출력 값을 이용해 작업증명의 완료 유무를 판단하였다.

ASIC 채굴 장비 등장의 역제는 LDPC 디코더의 사용에서 기인한다. 매 블록마다 새로운 패리티 체크 행렬을 무작위로 생성함으로써, ASIC-LDPC 디코더의 구현을 현실적으로 매우 어렵게 만든다. 그로 인해 디코더의 실행을 CPU/GPU에 의해서만 처리되도록 설계한 것이다.

Acknowledgement

이 논문은 2019년도 광주과학기술원의 재원으로 GRI(GIST 연구원) 사업의 지원을 받아 수행된 연구임. 이 논문은 2019년도 광주과학기술원의 재원으로 “과학기술응용연구 단의 실용화 연구개발사업”의 지원을 받아 수행된 연구임.

참고문헌

- [1] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System.” [Online]. Available: <https://bitcoin.org/bitcoin.pdf>.
- [2] Sangjun Park, Haeung Choi and Heung-No Lee, “Time-variant proof-of-work using error-correction codes,” in preparation.
- [3] V. Buterin. Dagger: A memory-hard to compute, memory-easy to verify scrypt alternative. Tech Report, hashcash.org website, 2013.
- [4] E. Duffield and D. Diaz, Dash: A payments-focused cryptocurrency, [Online]. Available: <https://github.com/dashpay/dash/wiki/Whitepaper>
- [5] R. G. Gallager, Low Density Parity Check Codes, Monograph, M.I.T. Press, 1963
- [6] W. E. Ryan and S. Lin, Channel Codes Classical and modern, Cambridge
- [7] S. Shao, P. Hailes, T-Y. Wang, J-Y. Wu, R. G. Maunder, B. M Al-Hashimi and L. Hanzo, “Sur-

vey of Turbo, LDPC and Polar Decoder ASIC Implementation”, IEEE Communications Surveys & Tutorials 2019

- [8] Y. L. Ueng, B. J. Yang, C. J. Yang, H. C. Lee, and J. D. Yang, “An Efficient Multi-Standard LDPC Decoder Design Using Hardware-Friendly Shuffled Decoding,” IEEE Trans. Circuits Syst. I, vol. 60, no. 3, pp. 743–756, March 2013.
- [9] T. Brack, M. Alles, T. Lehnigk-Emden et al., “Low complexity LDPC code decoders for next generation standards,” in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE '07), pp. 331–336, April 2007

Biography



박상준

2005년 3월 ~ 2009년 2월
충남대학교, 컴퓨터과 학사 졸업
2018년 5월 ~ 2019년 2월
광주과학기술원 블록체인 경제센터 연구원
2009년 ~ 현재
광주과학기술원 전기전자컴퓨터공학 석/박 통합과정

<관심분야>

정보이론, 신호처리, 블록체인, 수치최적화 이론, 압축센싱



김형성

2013년 3월 ~ 2019년 2월
전남대학교, 전자컴퓨터공학부 학사 졸업
2019년 3월 ~ 현재
광주과학기술원 전기전자컴퓨터공학 석사 과정

<관심분야>

블록체인



이홍노

1993년 University of California 전기공학과 졸업
1994년 University of California 전기공학과 석사
1999년 University of California 전기공학과 박사
1999년 ~ 2002년
HRL Laboratories Research Staff Member,
2002년 ~ 2008년
University of Pittsburgh Assistant Professor
2009년 ~ 현재
광주과학기술원 전기전자컴퓨터공학부 교수

<관심분야>

정보이론, 신호처리, 통신/네트워크, 압축센싱, 블록체인 경제, 센서지능화

암호화폐 투자와 규제현황

광주과학기술원 | 정현준·이흥노

1. 서론

2008년 10월 사토시 나카모토가 블록체인을 적용한 비트코인(P2P 전자 화폐 시스템) 논문을 세상에 공개하였다[1]. 블록체인은 지금까지 해결하지 못하고 있는 전자상거래상의 신뢰 문제를 해결할 수 있는 기술로 주목받고 있다. 우리는 전자 상거래에서 정보를 주고받을 때 신뢰를 제공하는 중재자에게 비용을 지급한다. 블록체인은 중재자 없는 거래를 가능하게 하고 있고 중재자가 없기 때문에 거래 비용을 줄일 수 있다.

블록체인은 저장된 데이터가 변경되지 않는다는 무결성을 확보할 수 있다[2]. 블록체인을 사용한 암호화폐는 암호학적 증명(cryptographic proof)에 기반을 두어 두 당사자가 제삼자(중재자)의 개입 없이 거래를 가능하게 한다. 블록체인은 블록에 거래 내용을 담고 블록과 블록 사이를 해시값을 이용하여 연결한다. 블록체인의 중간블록의 값을 변경할 시에는 해당 블록의 해시값이 바뀌고, 그 영향을 받아 뒤에 연결되어있는 블록의 해시값들도 모두 변경된다. 암호화폐는 블록체인을 공유하고 누군가 변경하지 못하게 서로 감시한다. 이를 통해 암호화폐는 중재자 없는 신뢰성을 확보할 수 있게 된다.

블록체인은 모두에게 오픈되어 있고 누구나 블록 생성에 참여할 수 있다. 해커는 블록체인의 일부 또는 전부를 변경하여 블록체인을 공격한다. 자신의 이득을 위해 해킹한 블록이 최신 블록이 되고 체인이 이어져 나갈 때 공격은 성공한다. 블록체인 해킹의 성공은 이론적으로 블록체인의 채굴파워를 51% 이상 점유하면 가능하다. 블록생성에 대한 권한을 독점하여 이중지급을 감행하거나 채굴된 암호화폐를 부정하게 다량으로 확보(selfish mining)하는 공격으로 블록체인

네트워크를 사실상 무력하게 할 수 있다. 새롭게 제안되는 암호화폐들은 알려진 공격을 어떻게 회피해야 할지에 대하여 면밀하게 검토하고 블록체인 네트워크를 설계해야 한다[3].

2017년 비트코인, 이더리움 등 암호화폐의 가격이 급격히 상승하면서 많은 사람의 관심이 집중되었다. 그 이후 현재까지 각종 암호화폐가 공개되고 투자를 받고 있다. 암호화폐 프로젝트들은 프로젝트의 사상이 담긴 백서(White Paper)를 발간하여 ICO(Initial Coin Offering)를 진행하고 투자금을 모집한다. 투자자들은 백서에 기술된 내용을 기반으로 판단하고 투자한다. 하지만 수 천개에 이르는 암호화폐의 공개 속에서 또한 검증된 사실이 아닌 가설과 계획만이 기술된 백서라는 것을 통해, 투자자들이 암호화폐 프로젝트를 이해하는 것은 매우 어렵다. 백서에 제안된 내용을 프로젝트팀이 성실하게 이행할 것인가도 보장받지 못한다. 또한 뛰어나다고 평가를 받는 백서에 기술된 블록체인 네트워크 설계를 실현할 수 있는 구성원이 팀 내에 확보되어 있는지 확인하기도 어렵다. 이러한 정보의 비대칭성으로 인해 암호화폐 시장은 ‘깜깜이 투자, 묻지마 투자’가 이루어지고 있다고 평가받고 있다[4]. 한편으로는 암호화폐의 객관적 가치를 평가하고 평가지표를 제공해주는 사이트들이 생겨나고 있다[5-7]. 이런 평가지표를 활용하여 투자위험을 줄일 수 있게 되는 것은 바람직하다.

암호화폐 시장에 접근하기 위해 일반적으로 투자자는 암호화폐 거래소라고 불리는 가상 자산 거래 플랫폼에 의존한다. 우리가 알고 있는 대부분의 암호화폐 피해 사건은 암호화폐 거래소와 연관되어 있다. 거래소는 증권 또는 상품 법률에 따라 등록되어 있지 않으며, 보안, 내부 통제, 시장 감시 프로토콜, 공개 또는 기타 투자자 및 소비자 보호에 대한 의무를 지고 있지 않다. 투자자들은 거래소에 대한 피해 손실을 떠안고 있다고 발표되었다[8-13].

이 논문의 기여점은 다음과 같다. 1) 한국을 포함한

† 이 논문은 2018년도 광주과학기술원의 재원으로 “과학기술응용 연구단의 실용화 연구 개발사업”의 지원을 받아 수행된 연구임.

국가별 최신의 암호화폐의 규제현황을 분석하여 규제 방향성을 보여준다. 그리고 암호화폐 투자현황 및 주요특징을 분석한다. 2) 건전한 암호화폐 투자 환경 조성을 위한 국가와 개인의 방향성을 제시한다. 블록체인과 암호화폐를 신산업으로 성장시키기 위해 투명하며 자유롭고 공정한 경쟁이 가능한 시장 환경 조성이 필요하다. 3) 대한민국의 규제를 전 세계적 규제현황과 대비해 보아, 자유롭고 창의적이며 투명한 시장 환경 조성을 위해 필요한 것들이 무엇인지 제시해 보고자 한다.

이 논문은 2장에서 세계적으로 암호화폐 규제와 육성 동향을 살펴본다. 3장에서는 주요 국가와 한국의 암호화폐 규제현황에 대해 알아본다. 4장에서는마켓캡 상위 암호화폐의 특징과 ICO 현황을 분석한다. 5장에서는 암호화폐 거래소의 투자 건전성에 대하여 알아보고 나아갈 방향을 제안한다. 6장에서는 암호화폐 투자현황을 분석하고, 건전한 시장환경 조성을 위하여 나아가야할 방향을 제시한다. 마지막으로 7장에서는 결론을 제시한다.

2. 전 세계국가의 암호화폐 규제와 육성 동향

암호화폐(Cryptocurrency)는 고유한 암호화를 사용하는 블록체인 기술을 기반으로 구성한다. 해당 기술을 설명하는데 사용되는 용어는 나라마다 다르다. 암호화폐를 참조하는 국가에서 사용되는 용어는 digital

currency(아르헨티나, 태국, 호주), virtual commodity(캐나다, 중국, 대만), cypto-token(독일), payment token(스위스), cyber currency(이탈리아, 레바논), electronic currency(콜롬비아, 레바논), virtual asset(온두라스, 멕시코) 등이 있다. 표 1은 암호화폐 시장의 적법성 인정관점에서 허용국가와 제한국가 분류를 보여준다[14]. 허용하는 국가는 암호화폐 시장을 인정하고 규율하는 특별법을 제정하였다. 브라질과 아르헨티나 같은 국가는 부분적으로는 시장이 존재하도록 허용은 하나 아직 산업 관련 특별법을 제정하지 않았다. 제한하는 국가는, 가령 알제리와 볼리비아, 암호화폐에 관한 모든 종류의 활동을 금지하였다. 카타르와 바레인 은 국경 내에서 암호화폐 거래를 금지하고 국경 밖에서 가능하도록 하였다. 그리고 중국과 방글라데시 같은 국가는 시민의 암호화폐 투자는 허용하지만, 국경 내의 금융기관을 통제하여 간접적으로 제한하였다.

암호화폐 프로젝트는 개발 자금 조달 계획으로 ICO를 사용할 수 있다. ICO는 사업자가 초기 자금 조달을 목적으로 가상화폐 코인을 발행하고 투자자들에게 선 판매하여 자금을 확보한다. 투자금은 비트코인, 이더리움으로 받기 때문에 전 세계 투자자들이 쉽게 참여할 수 있다. 하지만 ICO는 투자 위험성이 매우 크다. 첫째, ICO에 대한 내용과 실제 달성할 수 있는 구성원이 있는지에 대한 검증이 어렵다. 둘째, ICO가 제대로 진행되고 있는지 확인이 어렵거나 정해놓은

표 1 암호화폐 허용국가와 제한국가 비교

구분	특징	국가
암호화폐 허용 국가	- 국가에서 암호화폐 시장을 인정하고 규율하는 특별법을 제정하는 선제적 조치를 취함.	벨라루스, 지브롤터, 저지, 멕시코 등
	- 시장이 존재하도록 허용하기는 하나 아직 산업 관련 특별법을 제정하지 않음.	브라질, 아르헨티나, 프랑스 등
암호화폐 제한 국가	- 암호화폐의 모든 종류의 활동을 금지함.	알제리, 볼리비아, 모로코, 네팔, 파키스탄, 베트남
	- 국경 내에서는 암호화폐의 모든 활동을 금지하고 국경 밖에서는 허용함.	카타르, 바레인
	- 시민이 암호화폐에 투자를 허용하지만 국경 내의 금융 기관이 암호화폐와 관련된 거래를 용이하게 못하게 하여 간접적으로 제한함.	방글라데시, 이란, 태국, 리투아니아, 레소토, 중국 및 콜롬비아

표 2 ICO 규제 국가와 전면적 규제국가

구분	특징	국가
ICO 전면적 규제	- ICO와 관련된 어떠한 활동도 인정하지 않음.	중국, 마카오, 파키스탄 등
ICO 규제	- 토큰이 채무 보증, 주식 보안, 관리 투자 제품 또는 파생상품으로 분류되는지 여부에 따라 특정 의무를 적용함.	뉴질랜드, 네덜란드, 스위스 등

표 3 암호화폐 기술에 대한 육성정책 국가

구분	특징	국가
암호화폐 진화적 규제 체제 개발	- 우수한 암호화폐기술 기업에 대하여 투자를 유도함.	스페인, 벨로루시, 케이맨 제도, 룩셈부르크
암호화폐 통화 체제 개발	- 국가에서 암호화폐를 직접 설계하고 네트워크를 구축함.	마살 군도, 베네수엘라, 동 카브리 중앙 은행(ECCB)회원국, 리투아니아
암호화폐 규제 반대 판결	- 암호화폐 시장의 규모가 작은 시점에서 규제 및 금지를 할 충분한 우려가 없다고 판결함.	벨기에, 남아프리카, 영국

일정대로 진행되지 않는다. 셋째, ICO 계획이 이행되지 않았을 경우 투자자는 보상받을 수 없다. 2018년 3분기까지 ICO를 분석한 결과 부실하게 진행되고 있다고 발표했다[15].

국가들은 ICO의 위험성 때문에 ICO에 대한 규제를 발표하였다. 표 2는 국가 중 규제국가와 전면적 규제국가를 보여준다[14].

여러 국가에서 블록체인 기술과 암호화폐를 위협이 아니라 기회로 보고 육성하려고 한다. 일부 국가는 암호화폐를 합법적으로 인식하지는 않지만, 기술의 잠재력을 인식하여 투자하는 정책을 펼치고 있다. 대한민국은 2017년 9월 ICO 전면적 금지를 선언한 상태다. 그렇지만 블록체인 기술에 대해서는 정부 차원에서 육성 정책을 수립하고 공표했다[16]. 다음은 표 3은 암호화폐와 관련 기술에 대하여 육성정책을 펼치는 국가이다[14].

3. 주요국가의 암호화폐 규제현황

전 세계 국가들은 사법체계에서 암호화폐 정책 및 규제 체계를 결정하고 시행하고 있다. 중요한 쟁점은 암호화폐의 적법성, 암호화폐에 대한 조세 처우, 돈세탁방지법·조직범죄방지법·테러자금지원법 등의 적용가능성이다[17]. 표 4는 각 국가의 중앙은행과 세무부의 규제내용을 보여준다. 암호화폐를 금융자산, 기타 자산 등으로 분류하고 세금부과의 대상임을 명시하고 있다. 일본은 암호화폐를 재산신고 대상으로 확정하였다. 벨라루스는 암호화폐, ICO, 스마트계약을 합법화하고 관련 세금을 5년간 면제하는 시행령을 발효하였다.

표 5는 각 국가의 법무부와 국회의 암호화폐에 대한 규제내용이다. 각 국가는 법제화를 통해서 암호화폐에 대한 정의와 ICO의 가이드라인, 암호화폐 거래

표 4 각 국가의 중앙은행과 세무부의 규제현황

구분	국가	규제내용
중앙은행	벨라루스	(2018) 국립은행(NBRB)은 ICO에 참여하는 투자자 조건을 엄격하게 제안함[18].
	이란	(2018) 중앙은행(CBI)은 은행, 신용공여기관 등의 모든 금융기관의 암호화폐 취급을 금지함.
	이스라엘	(2018) 암호화폐를 금융자산으로 취급함.
세무부	아르헨티나	(2017) 암호화폐의 판매로부터 얻은 이익을 주식/채권에서 얻은 수입으로 취급하고 과세함.
	아이슬란드	(2017) 국세청은 암호화폐를 기타자산에 포함하는 소득세 신고 가이드라인을 발표함.
	벨라루스	(2018) 암호화폐, ICO, 스마트 계약을 합법화하고 관련 세금을 5년간 면제하는 시행령을 발효함.
	브라질	(2018) 암호화폐 조세 초안을 발표함. 거래소들은 월간 보고서를 의무적으로 제출하고 개인/법인 투자자가 외국 암호화폐 거래소에서 일정 금액 이상을 거래할 경우 의무적으로 신고해야 함.
	중국	(2018) 국가인터넷정보판공관실(CSA)은 블록체인 정보 서비스 관리 규정정책 초안을 발표함.
	이스라엘	(2018) 세무당국은 암호화폐의 사용을 가상지불 수단으로 취급하고 세금부과 대상으로 함.
일본	(2018) 암호화폐를 재산신고 대상으로 확정함[19].	

표 5 각 국가의 법무부와 국회의 규제현황

구분	국가	규제내용
법무부	캐나다	(2014) 암호화폐를 범죄수익(자금세탁) 및 테러리즘 자금지원 금지법에 포함함. (2018) 하원 재정위원회는 법정화폐로 암호화폐를 구매하는 모든 과정을 통화서비스 사업으로 규정함[20].
	프랑스	(2018) ICO를 허용하며 가이드 라인을 제시함[21].
국회	호주	(2017) 상원 위원회는 암호화폐 사업자 등록과 사업자는 의심스러운 거래를 신고해야 하는 의무를 적용함.
	프랑스	(2018) 국회는 암호화폐 양도소득세를 36%에서 30%로 낮추는 개정안을 발효함[22].

표 6 각 국가의 금융위원회의 규제현황

국가	규제내용
아르헨티나	(2014) 가상화폐로 수행되는 사업은 돈세탁 혹은 테러 자금지원을 포함한 수상한 거래에 대하여 보고 필요함.
미국	(2017) 증권거래위원회(SEC)는 ICO, 블록체인 기술, ICO 투자 평가, 권리에 대해 지침을 제공하는 가이드라인 발표함. (2018) SEC는 ICO에 증권법 적용을 강화하겠다고 발표함[23].
스위스	(2018) 금융시장감독청(FINMA)는 암호화폐 ICO 가이드라인 발표함[24].
호주	(2018) 금융감독원(Austrac)은 호주 암호화폐 거래소에 대한 돈세탁 방지와 테러단체 자금줄 차단을 위한 법률을 제정함.
캐나다	(2018) 증권 관리소(CSA)는 암호화폐 공개(ICO), 토큰 공개(ITO), 암호화폐 투자펀드, 거래소에 대해 증권법 적용요건을 제시함.
지브롤터	(2018) 금융 서비스 위원회(GFSC)와 정부는 개인/단체의 토큰 홍보, 판매, 배포에 대한 규제를 제정함. 토큰 규제 가이드라인을 발표함.
이스라엘	(2018) 증권 규제당국(ISA)는 메시지 시스템의 사이버 보안을 위해 블록체인을 도입함.
일본	(2018) 금융감독당국(FAS)은 일반가상화폐거래소 협회(JVCEA)를 거래소 자율 규제기관으로 공식적으로 선정함[25].
저지섬	(2018) 금융 서비스 위원회(JFSC)가 ICO 투자보호조치 가이드라인을 발표함[26].
스위스	(2018) 자본시장기술협회(민간)는 암호화폐에 적용되는 자금세탁 방지 표준안을 발표함[27].

에 대한 소득세를 규정한다. 기관들은 암호화폐가 자금세탁과 테러 자금으로 사용되지 못하도록 법적으로 금지한다. 호주는 자국 내에서 암호화폐 관련 사업자 등록과 의심스러운 거래에 대한 사업자의 신고 의무를 적용한다.

표 6은 각 국가의 금융위원회의 암호화폐에 대한 규제내용이다. 금융위원회는 미국의 증권거래위원회(SEC), 스위스의 금융시장감독청(FINMA), 호주의 금융감독원(Austrac), 캐나다 증권 관리소(CSA), 지브롤터 금융 서비스 위원회(GFSC), 이스라엘 증권규제당국(ISA), 저지섬 금융 서비스 위원회(JFSC), 일본 금융감독당국(FSA)을 포함한다. 기관들은 암호화폐의 ICO에 대한 가이드라인을 제시한다. 최근 미국 SEC에서 ICO에 대한 증권법 적용을 하겠다고 발표하였다. 투자자를 보호하기 위하여 취한 결정이며, ICO

규제는 더욱 강해질 전망이다.

3.1 한국의 규제현황

한국 정부는 암호화폐를 이용한 ICO를 전면적으로 금지를 선언하였다. 그 이후 정부는 법제화를 위한 구체적인 행동을 하지 않고 암호화폐 시장을 관망하고 있다. 각 부서와 협회들은 가이드라인을 제안하였지만, 실효성은 미비한 상태이다. 예를 들어, 금융위원회에서 가상화폐 관련 가이드라인을 발표하였다. 은행은 가이드라인을 기준으로 거래소에 대해 입금정지 조치를 취하려 하였다. 하지만 이에 반하여 거래소는 가처분신청을 법무부에 제출하였는데 거래소의 손을 들어주었다. 법원은 가이드라인으로는 법적 근거가 부족하다고 판단하였다. 다음 표 7은 우리나라의 암호화폐 규제현황을 시간순으로 보여준다.

표 7 한국의 암호화폐 규제현황

구분	규제내용
정부 - 암호화폐를 이용한 ICO 전면 금지 (2017.9.29.)[28]	- 증권발행 형식을 포함한 모든 형태의 ICO를 금지 하겠다고 발표함. - 금전 대여·코인마진거래 등 신용공여 금지 및 이와 관련한 금융회사의 영업·업무제휴 등 전면 차단하겠다고 발표함. - 고객정보 유출 사고 조사, 가상통화 취급업자에 대한 공동점검체계 구축 목표를 발표함.
블록체인협회 - 암호화폐 거래소 자율규제안 (2018.1)[29]	- 암호화폐 거래소 건전화를 위한 기준을 마련함. - 자금 세탁행위 방지(본인확인 절차, 거래기록 5년 보관 등), 이상 거래 감지시스템 도입, 암호화폐 상장 시 이용자 보호(백서, 해외 거래가격 등 이용자 보호에 필요한 정보 공개), 재무건전성 증명(자기자본 20억 원 이상, 지배구조와 재무자료 제출 등), 윤리 현장 제정을 의무화함. - 암호화폐 거래소 12곳(업비트, 빗썸, 코빗, 코인원, 고팍스, CPDAX, 한국디지털거래소, 네오프레임, 오케이코인코리아, 후오비코리아, 한빛코, 코인제스트) 자율규제 심사 결과 통과함[30].
금융위원회 - 가상통화 관련 자금세탁방지 가이드라인 개정안 (2018.6.27.)[31]	- 암호화폐 관련 금융거래에 관하여 ①특정 금융거래보고 및 이용 등에 관한 법률과 그 하위법령의 시행에 필요한 사항을 명확히 하고 ②자금세탁 및 공중협박자금 조달 행위를 효과적으로 방지하기 위해 금융회사 등의 준수가 필요한 사항을 규정함. - 비 집금제에 대한 모니터링 강화, 해외 가상통화 취급 업소 목록 공유, 거래거절을 할 수 있음.
법원 - 은행의 암호화폐 입금정지는 부당 판결 (2018.10.29.)[32]	- 암호화폐 거래소(코인이즈)가 '은행(NH농협)의 입금정지 조치를 막아달라'며 제기한 가치분신청에서 법원이 거래소 손을 들어줌. 은행이 거래소와 거래 정지를 하려면 가이드라인이 아닌 법적 근거가 있어야 한다는 취지임.
블록체인 스타트업 협회 - IEO 가이드라인 (2018.11.01.)[33]	- IEO(Initial Exchange Offering)를 "거래소 상장 직전, 토큰을 판매해 사업자금을 조달하는 행위"로 정의함. - 가이드라인에는 ①MVP(Minimal Visible Product) ②셀프 체크리스트 ③공시가를 포함함. - 블록체인 사업 자금을 모으려면 동작하는 댕(Dapp)을 개발하거나, 개념증명(Proof of Concept) 이상을 구현하는 메인넷이 정상 가동해야 MVP로 인정함. - 체크리스트에서 700점 이상을 얻으면 15억 이상 제안 가능함.
블록체인 분석평가 가이드 2.0 - 한국블록체인협회 (2018.11.16.) [34]	- 블록체인 비즈니스 평가 기준 가이드라인은 투자자들의 정확한 투자 판단에 도움을 주고, 블록체인 산업의 건전한 발전과 육성을 위해 제시함. - 블록체인 분석평가를 토큰 구조 평가, BM 평가, 조직 평가, 기술 평가의 4가지 영역으로 구분하여 평가를 위한 가이드 항목을 도출함.

4. 암호화폐 투자현황 분석

다음 표 8은 암호화폐 중 마켓캡 상위 20개에 대한 투자현황과 주요특징을 보여준다. CoinMarketCap[35]과 bitinfocharts[36]에서 제공하는 정보와 각 암호화폐에서 제공하는 정보를 참조하였다. 분석 항목은 다음과 같다. Total은 현재 암호화폐 공급량과 총생산 예정 공급량을 나타낸다. 암호화폐를 무한정으로 발행한다면 가치가 하락하기 때문에 많은 암호화폐는 총공급량을 제한하여 인플레이션을 막는다. Market Capitalization은 발행된 암호화폐 개수와 가격의 곱으로 측정한다. Price는 현재 거래되는 암호화폐의 가격이다. TPS(Transaction Per Second)는 초당 처리 가능한 트랜잭션 수이다. 암호화폐들은 실생활에 적용하기 위하여 VISA 카드 TPS(약 2만 4000 TPS)를 목표로 하여 속도를 개선하고 있다. Exchange volume 24h는 암호화폐가 하루 동안 거래되는 양을 나타낸다. 그리고 암호화폐 전체 발행량과 하루

동안 암호화폐 거래량의 비율을 표시하였다. 비율이 높다면 화폐의 기능을 하고 있으며, 비율이 낮다면 자산의 기능을 하고 있다고 유추할 수 있다. Block Time은 암호화폐의 블록이 생성되는 시간을 나타낸다. Hashrate는 블록체인을 보호하는 컴퓨팅 파워의 양을 나타낸다. 이 수치가 증가할수록 블록체인을 해킹하는데 들어가는 비용이 증가한다.

주요 암호화폐들의 특징을 분석한 결과, 정책적으로 전체 통화 공급량을 제한하는 부류(Bitcoin, XRP, Stellar 등등...)와 통화를 계속해서 공급하는 부류(Ethereum 등)가 있다. 암호화폐 NEM은 발매할 때 전체 코인을 고정 발행하여 제로 인플레이션을 지향한다. 마켓캡 상위 20개의 암호화폐는 사람들이 안전하다고 인정하고 있는 비트코인과 이더리움의 하드포크 형태가 많았다(Bitcoin, Ethereum, Bitcoin cash, Bitcoin gold, Ethereum classic).

표 8마켓캡 상위 20개 암호화폐의 투자현황 및 주요특징(2018.11.21. 기준)

Cryptocurrency	Total	Market Cap. (USD)	Price	TPS	Exchange volume 24h (Market Cap.)	Block Time	Hashrate
Bitcoin	17,386,917 BTC / 21,000,000 BTC	\$78.53 B	\$452	7	749,113 (4.3%)	12m6s	42,337 Ehash/s
	사토시 나카모토가 2008년 최초로 만든 최초의 암호화폐						
XRP	40,327,341,704 XRP / 100,000,000,000 XRP	\$17.63 B	\$0.47	1,500	923,526,214 (2.2%)	3-5	n/a
	블록체인 기반 송금 시스템 (기업간)						
Ethereum	103,337,483 ETH	\$13.84 B	\$134	50	4,363,839 (4.2%)	14s	219,335 Thash/s
	암호화폐 플랫폼, 어플리케이션운영 가능						
Bitcoin Cash	17,580,232 BCH / 21,000,000 BCH	\$4.4 B	\$251	25	85,354 (0.4%)	10m	3,881 Ehash/s
	비트코인의 하드포크 (블록크기 증가)						
Stellar	19,149,936,244 XLM / 104,463,177,385 XLM	\$3.81 B	\$0.20	2,000	290,049,600 (1.5%)	3-4	n/a
	리플(XRP)에서 하드포크 (개인간의 거래)						
EOS	896,148,480 EOS / 1,006,245,120 EOS	\$3.38 B	\$3.78	3,996	96,793,154 (10.8%)	<10s	n/a
	DPOS 방식의 트랜잭션처리, 수수료없는 dapp						
Litecoin	59,263,946 LTC / 84,000,000 LTC	\$1.99 B	\$33	56	6,478,554 (10.9%)	2m30s	190,247 Thash/s
	비트코인보다 빠른 블록생성과 코인유통성						
Tether	1,806,421,736 USDT / 2,580,109,502 USDT	\$1.77 B	\$0.98	n/a	n/a	n/a	n/a
	미국 달러(USD)와 1:1 비율로 교환						
Cardano	25,927,070,538 ADA / 31,112,483,745 ADA	\$1.20 B	\$0.046	7-10	621,341,926 (0.2%)	<1m	n/a
	금융 어플리케이션을 위한 스마트 컨트랙트 플랫폼						
Monero	16,586,977 XMR	\$1.13 B	\$68.7	1,000	306,749 (1.8%)	2m	449,947 Mhash/s
	거래의 익명성 중시						
TRON	65,835,863,744 TRX / 99,000,000,000 TRX	\$951 M	\$0.014	10,000	3,059,758,404 (4.6%)	n/a	n/a
	엔터테인먼트에 초점을 맞춘 블록체인						
Dash	8,459,572 DASH / 18,900,000 DASH	\$931 B	\$110	28	280,874 (3.3%)	2m38s	2,956 Phash/s
	거래의 익명성 중시, 코인조인방식 사용						
IOTA	2,779,530,283 MIOT	\$883 M	\$0.318	1500	37,260,835 (1.3%)	n/a	n/a
	M2M(Machine-to-Machine) 거래를 위한 플랫폼을 목표로함, 컨센서스로 탱글 사용함.						
Binance	130,798,842 BNB / 190,799,315 BNB	\$779 B	\$5.96	n/a	2,792,557 (2.1%)	n/a	n/a
	홍콩 암호화폐 거래소 바이낸스의 기축통화용코인						
NEM	8,999,999,999 XEM	\$716 M	\$0.08	4,000	71,101,479 (0.7%)	1m	n/a
	POI 알고리즘 사용						

Cryptocurrency	Total	Market Cap. (USD)	Price	TPS	Exchange volume 24h (Market Cap.)	Block Time	Hashrate
NEO	65,000,000 NEO / 100,000,000 NEO	\$604 B	\$12.6	1,000	4,390,328 (6.7%)	<1m	n/a
	중국최초의 블록체인, 이중 토큰 시스템						
Ethereum Classic	106,196,104 ETC	\$601 M	\$5.66	14	19,694,712 (18.5%)	14s	11.326 Thash/s
	이더리움의 DAO 해킹 사태 해결을 위한 하드포크 하기전 블록체인을 기반						
Zcash	5,283,563 ZEC	\$449 M	\$84	27	448,655 (8.4%)	2m30s	n/a
	사용자의 익명성 초점						
Tezos	607,489,041 XTZ / 763,306,930 XTZ	\$432 M	\$0.7	40	1,033,205 (0.1%)	1m	n/a
	DPOS 사용, 하드포크가 필요없는 블록체인						
Bitcoin Gold	17,442,113 BTG / 21,000,000 BTG	\$ 366 M	\$20	7	191,392 (1%)	10m	5,026 Mhash/s
	비트코인 하드포크 (GPU로 채굴가능)						

표 9 마켓캡 상위 20개 중 ICO 현황 분석(2018.11.21. 기준)

구분	USD Raised / Tokens sold	ICO date complete	Sale Price	Current Price	Return
IOTA	\$434,511.63 / 999,999,999	2015-11-25	\$0.00043	\$0.318	+73081%
Ethereum	\$15,571,000 / 50,000,000	2014-07-22	\$0.311	\$134.550	+43105%
NEO	\$556,500 / 17,500,000	2015-10-01	\$0.032	\$9.121	+28582%
Binance	\$15,427 / 100,000,000	2017-07-14	\$0.150	\$5.948	+3865%
Cardano	\$63,000,000 / 26,000,000,000	2015-09-01	\$0.002	\$0.046	+1811%
EOS	\$185,000,000 / 200,000,000	2017-06-26	\$0.925	\$3.746	+305%
Tezos	\$230,498,884 / 490,423,158	2017-07-01	\$0.470	\$0.707	+50%

표 9는 마켓캡 상위 20개 중 ICO를 진행한 암호화폐에 대한 분석이다. ICOSTATE[37]에서는 암호화폐 ICO의 상세한 정보를 제공한다. 상위 20개의 암호화폐 중 ICO를 실행한 것은 상위 20개 중 8개이다. ICO 단계에서 투자자들은 비상장 코인에 투자하는 것이므로 위험을 감수해야 한다. ICO의 최대 수익을 보여주고 있는 IOTA는 2015년 11월 ICO 기준으로 현재 73081%의 수익률을 보여준다. 이어서 이더리움은 2014년 7월 ICO 기준으로 현재 43105%의 수익률을 보여준다.

5. 암호화폐 거래소의 투자 건전성

암호화폐 투자자들은 거래를 위하여 암호화폐 거래소를 이용한다. 암호화폐 거래소는 투자자의 암호화폐 교환과 자체적으로 투자를 진행하고 있다. 하지만 정보 불균형으로 인하여 투자자에게 불리한 불투명하고 불공정한 시장이 되고 있다고 지적되고 있다[38].

미국 뉴욕주 검찰청은 암호화폐 거래소의 건전성을

확인하기 위하여 설문 조사하고 결과를 발표하였다[39]. 검찰청은 일반법에 근거하여 투자자를 기만하는 불공정한 행위를 보호하고 규율할 수 있다. 보고서의 결과에서는 현재 영업하고 있는 거래소의 영업행태가 세 가지 측면에서 우려할 수준이라고 발표했다.

첫째, 거래소는 암호화폐의 거래소 역할 뿐만 아니라 딜러와 투자자 역할을 동시에 수행하고 있다. 거래소는 투자자들보다 빠르게 암호화폐에 대한 정보를 얻을 수 있으며, 비공개 정보를 이용하여 투자할 수 있다. 비공개 정보를 이용하여 투자하기 때문에 투자자들과 공정한 경쟁이 어려우며, 고객의 이해에 반하는 행동을 할 수 있다. 둘째, 공격적인 투자자들은 Bot이나 알고리즘을 사용하여 거래소를 이용한다. 이런 도구들은 암호화폐 가격조정으로 이용할 수 있다. 거래소는 이상 거래가 발생할 경우를 대비하지 못하거나 감시할 의무로 생각하지 않고 있다. 셋째, 거래소가 고객의 자산을 안전하게 보유하고 있는지 여부

를 일관되고 투명하게 감시할 방법이 없다. 또한 사고 발생 시 보호를 받을 보험이나 정책적 지원이 없다.

보고서에서는 거래소를 이용할 때 정보 불균형으로 인한 불건전 거래를 경고한다. 한국에서도 파산한 거래소에 대한 피해가 발생하였으며 투자자들에 피해가 전가되었다고 보도되었다[11].

국가에서는 암호화폐 거래소가 투명한 거래를 담보하기 위한 가이드라인을 제시해야 한다. 그리고 거래소가 정보 불균형을 이용하여 투자자의 이익에 반하는 행동을 하는 경우 법적인 제재가 필요하다. 거래소의 투명성이 확보된 후에 암호화폐에 대한 공정한 경쟁이 가능한 시장질서가 확립될 수 있다.

6. 건전한 암호화폐 투자환경 조성

2장과 3장의 내용을 기반으로 암호화폐에 대한 전 세계적인 규제정책의 흐름을 파악할 수 있었다. 많은 국가는 암호화폐 시장의 투자 위험에 대하여 경고하고 있다.

투자자 보호를 위해서 국가가 발행하고 보증하는 실제 통화와 그렇지 않은 통화를 구분할 수 있도록 투자자 지도가 필요하다. 그리고 국가는 암호화폐가 높은 변동성과 거래를 촉진하는 많은 조직이 규제받고 있지 않다는 사실을 알려 주어야 한다. 모든 투자에는 위험이 따른다. 그러나 암호화폐 투자는 위험이 매우 크며, 손실이 발생할 경우 투자자가 보호받을 수 있는 법적 수단이 마련되어 있지 않다. 법적인 기준을 마련하여 투자자를 보호하는 조치가 필요한 상황이다. 또한 국가는 돈세탁 및 테러와 같은 불법 활동에 암호화폐가 이용되지 않도록 기업활동을 감독하고 시장을 규율해야 한다. 국가는 필요하다면 자금 세탁, 테러 및 범죄 조직에 대한 법률을 확대하여 암호화폐 시장을 포함하고 은행 및 기타 금융 기관이 부과된 의무를 수행하도록 요구해야 한다.

개인은 자신이 투자하는 암호화폐에 대해서 분석하고 평가해야 한다. 블록체인 분석 및 평가 가이드라인(표 7 참조) 등을 활용하여 투자하려는 암호화폐가 제대로 설계되었는지 점검해야 한다. 그리고 일정 점수 이상의 암호화폐에 투자하는 것이 안전할 것이다. 그리고 암호화폐 가치평가 사이트[5-7]를 이용하여 객관적인 평가 지표를 활용하는 것이 중요하다.

개인과 국가는 암호화폐 투자를 위해서 각자의 역할이 존재한다. 국가는 규제화·법제화를 통하여 자국민들의 피해를 최소화해야 한다. 다른 한편으로는 국가는 신산업을 육성하여 국가의 동력원을 만드는 데 신기술을 활용해야 한다. 그러므로, 국가는 산업 육성

과 규제 간에 균형을 맞추어야 한다. 최소한의 규제 및 법제화로 건강한 시장을 만들고 산업이 국제적 경쟁력을 갖고 지속 성장할 수 있는 환경을 만들어 주는데 목표를 두어야 한다. 이런 목표를 이루기 위한 규제 및 법제 연구가 절실히 필요하다.

7. 결 론

이 논문에서는 암호화폐 투자와 규제현황에 대해서 알아보았다. 암호화폐의 규제는 국가별로 매우 다르게 진행되고 있다. 암호화폐 규제현황의 방향성을 확인하기 위하여 암호화폐 허용국가와 제한국가를 비교하였다. ICO는 전면이나 부분적이거나의 구분일 뿐 대부분 국가가 규제하고 있었다. 블록체인기술은, 기술이 내재한 가능성을 믿고 육성정책을 펴는 국가들이 있었다. 주요 국가들의 암호화폐 규제현황의 쟁점은 암호화폐에 대한 조세 처우, 돈세탁방지법·조직범죄방지법·테러 자금지원법 등의 적용 가능성이다. 한국 정부는 2017년 9월 암호화폐를 이용한 ICO를 전면적 금지를 발표하였다. 각 부처에서는 후속 조치로 자금세탁방지 가이드라인을 만들어 발표하였다. 블록체인 기업 협회들은 암호화폐시장과 IEO에 관련한 가이드라인을 만들어 자율규제를 하려는 노력을 하고 있다. 하지만 발표된 가이드라인들은 법적 근거를 확보하고 있지 않기 때문에 실효성 측면에서 문제가 발생하고 있다. 암호화폐 거래소와 투자자 사이에는 정보 불균형으로 인한 불공정 경쟁이 발생한다. 암호화폐와 블록체인 신산업을 성장시키기 위해서는 투명하고 공정한 경쟁이 가능한 시장 질서를 확립해야 한다. 시장제 공자와 규제기관이 협의하여 성숙하고 건전한 시장을 만들기 위한 노력이 지속적으로 필요한 상황이다.

참고문헌

- [1] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008).
- [2] ZYSKIND, Guy, et al. Decentralizing privacy: Using blockchain to protect personal data. In: Security and Privacy Workshops (SPW), 2015 IEEE. IEEE, 2015. p. 180-184.
- [3] Eyal, Ittay, and Emin Gun Sirer. "Majority is not enough: Bitcoin mining is vulnerable." Communications of the ACM 61.7 (2018): 95-102.
- [4] 한정닷컴 게임톡, "[김태림 칼럼] 암호화폐 평가가 꼭 필요한 이유", 2018.11.22.
- [5] <https://www.weissratings.com/>

- [6] <https://cryptorated.com/>
- [7] <https://coincheckup.com/>
- [8] 중앙일보, "국내 비트코인 거래소 해킹...55억 피해 고객에 떠넘겨 논란", 2017.4.27.
- [9] 한겨레, "빗썸 직원 PC 해킹 ... 회원 3만명 개인정보 유출 피해", 2017.7.3.
- [10] 보안뉴스, "가상화폐 거래소 또 다시 해킹! '코인이즈' 21억 어치 털렸다", 2017.10.7.
- [11] 중앙일보, "가상화폐거래소 '유빗', 해킹으로 파산 절차 진행 "피해 최소화 할 수 있도록 노력할 것", 2017.12.19.
- [12] 한겨레, "가상통화 거래소 '코인레일' 해킹...400억원 상당 피해", 2018.6.11.
- [13] 연합뉴스, "가상화폐거래소 빗썸 해킹...회사보유 코인 350억원 털려", 2018.6.20.
- [14] The Law Library of Congress, Global Legal Research Center, "Regulation of Cryptocurrency Around the World", June, 2018.
- [15] ICORATING, "ICO Market Research Q3 2018", 2018.11.15.
- [16] 과학기술정보통신부/정보통신기술진흥센터, "2018 ICT R&D 기술로드맵 2023 발표회", 2018.11.2.
- [17] The Law Library of Congress, Global Legal Research Center, "Regulation of Cryptocurrency in Selected Jurisdictions", June, 2018.
- [18] EMERICs, "벨라루스, 암호화폐 투자 규제 강화", 이슈 & 트렌드, 2018.6.11.
- [19] 파이낸셜뉴스, "암호화폐도 재산...소득세-상속세 부과 ...日정부 과세 결정", 2018.11.22.
- [20] THE BCHAIN, "캐나다의회 - 거래소 허가제, 암호화폐 취득관리필요", 2018.11.19.
- [21] <https://www.gouvernement.fr/en/pacte-the-action-plan-for-business-growth-and-transformation>
- [22] 블록타임스TV닷컴, "암호화폐 세금 6% 낮추는 프랑스의 2019년 개정안", 2018.11.25.
- [23] cointelegraph, "US SEC Levies 'First' Civil Penalties Against Two ICOs for 'Unregistered' Securities", 2018.11.16.
- [24] SWISS finma, "FINMA publishes ICO guidelines", 2018.2.16.
- [25] CCN: Cryptocurrency News, "Japan Approves Self-Regulation for Cryptocurrency Industry", 2018.10.4.
- [26] Jersey Financial Services Commission, "The Application Process for Issuers of Initial Coin Offerings (ICOs)" Guidance Note, 2018.8.
- [27] cointelegraph, "Swiss Financial Association Publishes Anti-Money Laundering Standards for Digital Assets", 2018.10.23.
- [28] 금융위원회, "기관별 추진현황 점검을 위한 가상통화 관계기관 합동TF 개최" 보도자료, 2017.9.29.
- [29] 한국블록체인협회, "한국블록체인협회 자율규제안", 2017.12.15.
- [30] 한국블록체인협회, "한국블록체인협회, 제1차 자율규제심사 결과발표를 위한 기자간담회 개최", 2018.7.10.
- [31] 금융위원회, "가상통화 관련 자금세탁방지 가이드라인 개정안", 2018.6.27.
- [32] 코인데스크, "법원 - 은행의 암호화폐 거래소 입금정지는 부당", 2018.10.30., <https://coinsosik.com/punch?id=38388>
- [33] 블록체인 스타트업 협회, 블록체인산업진흥협회, 고려대 암호화폐연구센터, "IEO GUIDELINE", 2018.11.1.
- [34] 블록체인학회, "블록체인 분석평가 가이드", 2018.11.16.
- [35] <https://coinmarketcap.com/>
- [36] <https://bitinfocharts.com/>
- [37] <https://icostats.com/roi-since-ico>
- [38] 이흥노 교수, "암호화폐시장 투자 건전성 확보 시급!", 광주 MBC 라디오칼럼, 2018.11.8.
- [39] Attorney General, Barbara D. Underwood, "Virtual Markets Integrity Initiative: Report," Office of the New York State, 2018.11.8 <https://virtualmarkets.ag.ny.gov/#section6>.

약 력



정 현 준

2008 삼육대학교 컴퓨터과학과 졸업
 2010 숭실대학교 컴퓨터학과 석사
 2017 고려대학교 컴퓨터전파통신공학과 박사
 2017~현재 광주과학기술원 센서지능화연구센터 /블록체인인터넷경제연구센터 연구원
 관심분야: 사물인터넷, 인공지능, 블록체인, 센서 지능화
 Email: junhj85@gist.ac.kr



이 흥 노

1993 University of California 전기공학과 졸업
 1994 University of California 전기공학과 석사
 1999 University of California 전기공학과 박사
 1999-2002 HRL Laboratories Research Staff Member
 2002-2008 University of Pittsburgh Assistant Professor
 2009~현재 광주과학기술원 전기전자컴퓨터공학부 교수
 관심분야: 정보이론, 신호처리, 통신/네트워크, 압축센싱, 블록체인, 센서지능화
 Email: heungno@gist.ac.kr



블록체인개발 현황과 보안이슈 변화 동향

저자 (Authors)	정현준, 이흥노 Hyunjun Jung, Heung-No Lee
출처 (Source)	정보보호학회지 28(3) , 2018.6, 47-52 (6 pages) REVIEW OF KIISC 28(3) , 2018.6, 47-52 (6 pages)
발행처 (Publisher)	한국정보보호학회 Korea Institute Of Information Security And Cryptology
URL	http://www.dbpia.co.kr/Article/NODE07484331
APA Style	정현준, 이흥노 (2018). 블록체인개발 현황과 보안이슈 변화 동향. 정보보호학회지, 28(3), 47-52.
이용정보 (Accessed)	광주과학기술원 210.107.***.17 2019/02/18 15:01 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

블록체인개발 현황과 보안이슈 변화 동향

정 현 준*, 이 흥 노**

요 약

암호화폐가 세계적으로 주목받으며 핵심기술인 블록체인에 대한 관심이 증가하고 있다. 블록체인은 블록을 P2P 방식을 기반으로 생성된 체인 형태의 연결고리로 분산 저장되어 있으며 임의로 수정할 수 없고 누구나 변경의 결과를 열람가능하다. 블록체인의 공개 형태에 따라 공개형 블록체인과 허가형 블록체인으로 나뉘어 연구되고 있다. 이 논문에서는 세대별 블록체인의 개발 현황과 특징에 대해서 알아본다. 또한 블록체인 공개형태에 따른 특징과 보안성에 대하여 알아보고자 한다.

I. 서 론

블록체인(Blockchain)은 비트코인(Bitcoin)의 등장과 함께 세상에 알려졌다. 비트코인은 2009년 등장하여 최근 가격이 급격히 증가함에 따라 암호화폐(Cryptocurrency)가 세계의 주목을 받고 있다. 그리고 암호화폐의 기반기술인 블록체인에 대한 기술적 분석과 블록체인을 접목한 비즈니스가 양성되고 있다.

미국 애리조나주에서는 블록체인 기록 등의 법적 유효성 목적의 입법을 진행하였다. 이 입법에서 블록체인 기술이란 분산, 탈중앙화, 공유, 복제의 성질을 가진 분산화된 원장이라고 정의하였다. 미국 하와이주에서는 블록체인 산업의 진흥의 목적 입법을 진행하였다. 이 입법에서 블록체인이란 새로운 P2P 네트워킹 및 탈중앙화의 분산 데이터 저장 기술이라고 정의하였다. 한국은행은 블록체인을 거래정보를 기록한 원장을 특정 기관의 중앙서버가 아닌 P2P 네트워크에 분산하여 참여자가 공동으로 기록하고 관리하는 기술이라고 정의하였다 [1]. 블록체인이란 1) 통제에 대한 탈중앙화를 목적으로, 2) 분산화된 구조를 가지며, 3) 데이터의 저장할 수 있는 구조를 말한다.

블록체인 기술은 비트코인의 기반기술로 알려져 있다. 블록체인은 새로운 정보화 기술으로써 산업 전반에 걸쳐 영향을 미칠 기술이다. 일반적인 시스템에서는 클라

이언트 서버 모델(Client-server model)을 적용한다. 이 모델은 서비스 요청자인 클라이언트와 서비스 자원의 제공자인 서버 간에 작업을 분리해주는 네트워크 아키텍처이다. 블록체인은 중개자 없이도 개인(peer)간의 거래, 가치, 자산 등을 교환할 수 있는 신뢰 프로토콜을 제공한다. 기존의 서버는 신뢰 프로토콜을 유지하기 위하여 비용과 노력이 필요하다. 서버에 저장하고 있는 데이터(혹은 코인)는 공격자의 타깃이 된다. 시스템 관리자는 공격자의 공격을 막기 위하여 보안을 지속해서 관리해야 한다. 블록체인은 지켜야 하는 데이터를 모두에게 공개하여 서로를 감시하게 하여 무결성을 유지한다. 암호화폐에서 블록체인은 모든 거래 내용을 공개하고 인터넷에 분산 저장한다. 즉, 블록체인이란 공개 장부를 공정하게 만들고 관리하기 위한 기술이다. 공개 장부에 한 번 기록된 것은 변경할 수 없으며 위조할 수 없기 때문에 화폐에 응용 가능하다.

비트코인의 가격이 2017년 1월 1BTC 가격은 약 600만원에서 2018년 1월 약 2500만원까지 상승하였다. 한국 정부에서는 암호화폐의 가격의 변동성이 심해지자 신규계좌의 생성을 막는 등의 제재를 실행했다. 블록체인과 암호화폐에 대한 사회적 관심이 증가하였으며 급격한 가격변동으로 인해 우려도 증가하였다. 블록체인의 공개 형태에 따라 공개형 블록체인(Public Blockchain)과 허가형 블록체인(Private Blockchain)으

이 논문은 2018년도 광주과학기술원의 재원으로 “과학기술융용연구단의 실용화연구개발사업”의 지원을 받아 수행된 연구임.

* 광주과학기술원 센서지능화 연구센터 연구원 (junghj85@gist.ac.kr)

** 교신저자, 광주과학기술원 전기전자컴퓨터공학부 교수 (heungno@gist.ac.kr)

로 개발되고 있다.

이 논문은 블록체인 공개형태에 따른 블록체인의 특징과 보안 연관성에 대해서 알아본다. 2장은 세대별 암호화폐를 구분하여 특징을 정리한다. 3장은 공개형태에 따라 공개형 블록체인과 허가형 블록체인의 특징을 정의하고 비교한다. 4장은 블록체인에 적용할 수 있는 암호화폐를 분류한다. 5장은 결론으로 블록체인을 이용한 비즈니스가 나아가야 할 방향에 대하여 말한다.

II. 세대별 암호화폐

암호화폐는 비트코인을 시작으로 현재까지 수백개가 제안되었다. 암호화폐는 단순 결제 기능의 화폐성 1세대 암호화폐(표 1)와 스마트 계약이 가능한 2세대 암호화폐(표 2)로 구분한다. 이 논문에서는 1.5세대 암호화폐(표 3)를 1세대 암호화폐의 기본 기능에 추가기능을 넣은 암호화폐, 2.5세대 암호화폐(표 4)를 2세대의 한계를 극복하고자 나온 암호화폐로 구분하였다.

암호화폐는 세대별로 진화할수록 기능에 초점을 맞춰 발전되고 있다. 예를 들어, 새로 제안되는 암호화폐는 비트코인의 확장성, 거래 속도 향상을 위하여 기존의 구성요소를 변경 혹은 추가하여 구성한다. 하지만 신규 암호화폐의 새로운 기능과 속도를 위한 구조로 인하여 보안성이 약해지는 경향이 있다.

[표 1] 1세대 - 단순 결제 기능의 화폐성 암호화폐

이름	발행 or ICO	특징
비트코인 (BTC)	2009. 1.3	사토시 나카모토가 제안한 최초의 코인[2]. 전자 화폐를 디지털 서명의 체인으로 정의함. 코인 소유자는 거래 내역에 디지털 서명을 한 후 다음 사람에게 전달하고, 이를 받은 사람은 자신의 공개 키를 코인 맨 뒤에 붙임. 돈을 받은 사람은 앞사람이 유효한 소유자였다는 것을 확인가능.
비트코인 캐시 (BCH)	2017. 8.1	비트코인에서 하드포크 되어 생성된 알트코인이다[3]. 한계 속도를 극복하기 위해 블록 크기를 2~8MB까지 유동적으로 늘리는 편법 정책을 적용함.
비트코인 골드 (BCG)	2017. 10.25	비트코인에서 하드포크 되어 생성된 알트코인임[4]. 그래픽카드(GPU)로 채굴할 수 있음.

이름	발행 or ICO	특징
비트코인 다이아몬드 (BCD)	2017. 12	블록 크기 제한을 8MB로 변경하여 트랜잭션 용량이 향상되고 블록이 5배 빠르게 생성된다[5]. 거래 전송 시에 금액을 암호화하여 개인정보를 보호함.
리플 (XRP)	2012	블록체인 기반 송금 시스템임. 중앙 통제식(채굴이 존재하지 않음)이며 국제간 화폐 거래를 이용한 프로그램을 지원하여 수수료 및 환율 시세차익을 얻음[6]. 암호화폐의 기본 이념인 탈규제, 탈중앙화, 익명에 정면으로 반대되는 코인임.
스텔라 루멘 (XLM)	2014. 7	리플에서 하드포크 하여 개발된 암호화폐이다[7]. 비영리 기업 스텔라 재단에서 운용하는 화폐이다. 리플은 기업 간의 자금 송금을 목적으로 하고 스텔라루멘은 개인 간의 거래를 위하여 만들어짐.
라이트 코인 (LTC)	2011. 10.7	비트코인을 중심에 두고 개발되었음. 비트코인보다 약 4배 빠른 거래가 이루어진다[8]. 라이트닝 네트워크는 비트코인과 라이트코인에 복수 적용될 예정임. 이를 통해 아토믹스왑이 실현 가능해질 예정임.

[표 2] 1.5세대 - 1세대 암호화폐의 기본 기능과 추가 기능을 넣은 암호화폐

이름	발행 or ICO	특징
제트 캐시 (ZEC)	2016. 10.28	트랜잭션의 프라이버시와 선택적 투명성을 제공하는 분산형 오픈소스 코인임[9]. 제트캐시 지급은 영 지식 증명 기술(zero-knowledge proof) 기반으로 공개 블록체인에 개시되지만 거래의 보낸 사람, 받은 사람 및 금액은 사적으로 유지됨.
모네로 (XMR)	2014. 4.18	CryptoNight이라는 독자적 작업 증명 기법을 사용하여 채굴기와 이를 소유한 자원에 의한 탈중앙화(decentralization)적 가치가 훼손되는 것을 막음[10]. 거래내역이 비공개로 되어있어, 누가 누구에게 얼마를 보냈는지 알 수 없음.
대시 (DASH)	2014. 2.14	Dash 전송을 요청하면 마스터노드가 3개 이상의 거래 내역을 섞어서 보내는 코인조인(coinjoin) 방식을 사용함[11]. 마스터 노드는 Dash를 1,000개 이상 가진 사람이며 향후 대시의 개발 및 운영 방향에 대한 투표권을 가짐.

이름	발행 or ICO	특징
팩텀코인 (FCT)	2015. 9	데이터(문서)의 투명성과 지속성을 위하여 제안된 플랫폼[12]. 팩텀 블록에는 문서/기록의 고유값을 저장할 수 있음.
지코인 (XZC)	2016. 10.6	CPU와 GPU를 통해 채굴할 수 있음[13]. 1. 공용코인 소유자는 개인코인을 주조할 수 있음. 개인코인 사용 시 송금 이력 추적이 불가능함(영지식 증명 기반).
나브코인 (NAV)	2014. 6	비트코인 코어를 개량하여 만들어진 코인임[14]. 빠른 전송속도(블록타임 30초, 블록사이즈 20mb 확장)와 익명성(Navtech라는 이중블록체인 기술이용)이 특징이며 aDapp(익명화된 분산 어플리케이션)을 지원하는 플랫폼으로서 역할을 함.
시아코인 (SC)	2015. 6	클라우드 데이터 저장 서비스임. 블록체인을 이용한 스토리지 서비스를 제공함[15]. 컴퓨터의 저장공간을 다른 사람에게 임대하고 사용료를 받음. 기존 상용 클라우드서비스보다 평균 10배 이상 저렴함.
버스트 코인 (BURST)	2014. 8	PoC(Proof of Capacity)를 사용하여 채굴함[16]. PoC는 컴퓨터에 남아있는 잉여부분의 하드디스크를 사용하여 채굴하는 방식임.
스토리지 (STORJ)	2017. 7.2	이더리움 기반으로 만들어진 분산화된 클라우드 저장 플랫폼임 [17]. 하드디스크의 남은 용량을 클라우드 형태로 임대하고 코인을 획득함.
NEM코인 (XEM)	2015. 3.31	약 90억 개의 고정된 통화 발행으로 인플레이션이 제로인 코인임. 자바, 자바스크립트로 코드가 작성됨[18]. NEM 코인에 적용된 PoI (Proof of Importance) 알고리즘은 코인의 유용성과 거래참여 기여도를 측정하여 채굴자의 중요도를 결정하고 중요도에 비례하여 보상 함.
버트코인 (VTC)	2014. 1	NIST5기반의 Lyra2Re 체인 알고리즘을 제안하여 마이닝 중앙 집권화를 방지함[19]. 지금까지 2차례 PoW를 변경함.
디지털 바이트 (DGB)	2014. 1.10	다섯 개의 마이닝 알고리즘을 사용하여 마이닝 중앙 집권화를 방지함 [20]. 초당 280회의 트랜잭션을 처리할 수 있음. DGB 코인을 보상으로 지급하는 게이밍서비스를 제공함.

[표 3] 2세대 - 계약 기능을 포함한 암호화폐

이름	발행 or ICO	특징
이더리움 (ETH)	2015. 7.30	블록체인 기술을 기반으로 스마트 계약 기능을 구현하기 위한 분산 컴퓨팅 플랫폼과 탈중앙화된 앱 개발 환경 제공[21]. 비탈리크 부테린(Vitalik Buterin)이 개발함. P2P 컴퓨터 네트워크를 데이터 및 코인거래 내역을 블록체인에 저장하는 것은 물론, 스마트 계약이 설정된 코드도 실행할 수 있는 컴퓨팅 플랫폼 제공.
퀀텀 (QTUM)	2016	비트코인 블록을 사용하여 이더리움의 스마트 계약 엔진을 연결한 플랫폼임[22]. 비트코인의 디자인을 사용하고 이것을 블록체인에 코드로 비즈니스 규칙을 저장하고 EVM(이더리움 Virtual Machine)과 연결함. 채굴방식으로 PoS (Proof-of-Stake, 지분 합의 증명)을 택했음.

[표 4] 2.5세대 - 2세대의 한계를 극복하고자 나온 암호화폐

이름	발행 or ICO	특징
아이오타 (IOTA)	2016	사물인터넷을 위한 수수료 없고 채굴자 없이 데이터 무결성을 추구하는 블록체인. 탱글(방향성 비사이클 그래피)이라는 새로운 분산장부 작성 기술을 사용[23]. 이 기술은 서버사용으로부터의 탈중앙화를 유지하며 채굴자 없이 즉 수수료 없는 거래가 가능케 함.
카드다노 (ADA)	2017. 10.1	암호화폐 개발 언어로 적합한 하스켈언어로 만들어진 블록체인임 [24]. 하스켈은 함수형 언어이며 카드다노 백서에 제시된 수학적 표현을 완벽하게 설명하고 입증함. 카드다노는 회계(Accounting)와 컴퓨팅(Computing)을 분리하여 구현되어있음. 우로보로스(Ouroboros)라는 PoS 증명 알고리즘을 사용함.
이오스 (EOS)	2017	이더리움의 PoS에 비해 빠른 트랜잭션 처리가 가능한 비잔틴장애 허용 DPoS(Delegated Proof Of Stake)방식을 사용함[25]. 이오스 Dapp은 사용자는 수료를 지급하지 않고 개발자가 이오스를 지급함.

Ⅲ. 공개형 블록체인과 허가형 블록체인

블록체인 기술은 서로 신뢰할 수 없는 인터넷 환경에서 사람이나 사물들이 중개인 없이 돈이나 자산을 안전하게 교환하는 것이다. 거래들이 기록되어 있는 분산화된 원장을 안전하고 위변조 될 수 없게 관리하기 위한 분산 데이터베이스 및 관련 기술을 말한다. 분산화된 원장은 암호화키를 이용하여 체인 형태로 연결되어 있어 위변조가 불가능(혹은 난이도가 높음)하다. 기존 시스템에는 거래명세를 독점하여 데이터를 보호하고 무결성을 검증하였다. 반대로 블록체인은 거래명세를 모두가 공유하고 내용에 대해 수정이 있는지 서로 감시하고 검증한다. 블록체인 검증에 참여하는 노드들이 많아야 하며 참여, 검증, 저장에 대해 보상이 필요하다.

노드들은 특정 기간(예, 비트코인 매 10분)에 일어난 거래 기록을 담은 블록(block) 단위로 저장한다. 블록을 조작하는 것을 막기 위하여 공동 관리하고 블록들을 해시 함수(예, sha256)를 이용하여 요약하고 연결체인을 만든다. 연결 체인은 이전 블록의 요약을 다음 블록에 추가하여 다음 블록이 완성되면 이전 블록을 수정할 경우 바로 확인할 수 있다. 분산 원장은 모든 블록을 가지고 있어서 내용을 확인할 수 있지만 변경할 수 없다. 위의 특징을 이용하여 P2P 구조로 모든 사람이 변경되지 않은 데이터를 가질 수 있으며 확인할 수 있다. 비트코인 등의 암호화폐의 블록체인은 금전 거래(트랜잭션)를 저장하고 있다. 금전 거래뿐만 아니라 다른 기록 정보를 원본 그대로 보존하는 목적으로 이용할 수 있다.

블록체인 네트워크는 비트코인, 이더리움 등과 같이 분산 원장에 누구나 참여할 수 있는 공개형 블록체인과 권한을 가진 이들만 참여할 수 있는 허가형 블록체인으로 나뉜다. 허가형 블록체인은 운영 규칙에 따라 운영 주체 또는 특정 몇 명만 원장을 만들 수 있다. 허가형 블록체인은 현재 운영하는 시스템에 블록체인을 적용하기 위한 특수한 형태이다.

기업이 허가형 블록체인을 도입하는 이유는 공개형 블록체인(예, 비트코인, 이더리움)을 그대로 적용할 경우 기존 시스템과 맞지 않기 때문이다. 그래서 자신에게 맞는 형태로 변경하여 적용하고 싶기 때문에 허가형 블록체인의 수요가 유지되고 있다. 기업이 허가형 블록체인을 적용하는 이유는 다음과 같다.

1. **거래 처리 시간 단축:** 비트코인은 하나의 블록이 생성되기 위해서는 10분이 소요된다. 블록의 기록이 안정되기 위해서 5개의 블록이 필요하다고 가정한다면 50분이 필요하다. 블록이 생성되는 시간이 비즈니스에서 필요한 요구 시간보다 길기 때문에 보다 적은 시간을 요구한다.
2. **트랜잭션 비용(Transaction fee) 조정:** 블록체인은 데이터를 P2P로 저장하기 때문에 트랜잭션을 저장하기 위해서 일정한 비용을 요구한다. 일반적인 은행 계좌일 경우 송금 시에 보내는 쪽이나 받는 쪽 정책에 따라 일정량의 수수료가 책정된다. 일반적인 암호화폐(예, 비트코인, 이더리움)의 수수료는 송금자의 수수료 납부 정책에 의해 결정된다. 이는 거래를 인증해주는 노드(채굴자)가 네트워크상에 반드시 존재해야 하기 때문이다. 하지만 암호화폐의 가격이 급격히 상승하고 거래량이 증가하여 송금자가 낮은 수수료 책정 시 채굴 노드에 의해 승인되지 않거나 시간이 오래 걸리는 문제가 발생한다. 기업에서 비즈니스로 발생한 트랜잭션이 트랜잭션 비용 때문에 저장이 늦어지거나 거부될 경우 문제가 발생할 수 있다.
3. **해킹 범죄 감소:** 기업은 블록체인을 적용하여 해킹 범죄를 감소시킬 수 있다. 기존의 서버를 공격하여 진행되었던 사이버 공격 대상이 P2P 전체를 대상으로 바뀌었으며 해킹을 위하여 투입되는 비용이 증가한다. 기업은 보안을 위해 소요되는 비용보다 블록체인을 도입으로 인한 유지비용이 적다면 블록체인 도입을 고려할 수 있다.
4. **위-변조 불가능한 프로세스 공유:** 일반 사용자들은 기업에서 저장하고 있는 데이터의 무결성에 관해서 확인할 수 있는 방법은 없었다. 기업은 블록체인을 이용하여 데이터를 저장하고 관리하는지에 대하여 프로세스를 공유할 수 있다. 기업과 사용자는 운용하고 있는 데이터에 대한 신뢰성에 대한 공유와 지속적인 관리가 가능하다.

블록체인 표준화 생태계는 하이퍼레저 프로젝트(Hyperledger Project)와 EEA(Enterprise Ethereum Alliance) 그리고 R3CEV로 구성되어있다. 하이퍼레저 프로젝트는 모든 산업에서 사용할 수 있는 블록체인 기술의 표준화 및 발전을 위한 오픈소스 커뮤니티이다. 리

눅스 재단, IBM, VM웨어, 레드햇, 오라클 등이 모여 오픈소스 프로젝트로 진행 중이다. 현재 200개 이상의 회원사가 참여하고 있으며 국내에서는 한국거래소, 예탁결제원, 코스콤 등이 포함되어 있다. 하이퍼레저 프로젝트 중 하이퍼레저 패브릭이 가장 빠르게 성장하고 있으며 159명의 개발자와 IBM과 인텔을 포함한 28개의 기업에서 패브릭 플랫폼을 지원하고 있다. EEA는 이더리움 기반 허가형 블록체인 컨소시엄이다. 150개 이상의 회원을 확보하고 있다. 삼성SDS, SK텔레콤, 더루프, 코인플러그 등이 참여하고 있다. EEA는 이더리움을 엔터프라이즈급 기술로 발전시켜 개인정보, 기밀성, 확장성, 보안 등 다양한 영역에서 연구 및 개발을 제공한다. 또한, 실시간 트랜잭션 처리 등 다양한 플랫폼 프로젝트들이 있다. R3 CEV는 글로벌 은행 컨소시엄이다. 100개 이상의 회원을 확보했다. 하나, 신한, 우리, 국민, 농협은행 등이 회원이다. 하지만 초기 참여자인 골드만삭스 등 대형 금융사의 이탈하여 상황이 좋지 않다.

표 5는 블록체인 유형별 특징을 보인다. 개방형 블록체인과 허가형 블록체인은 큰 차이점은 네트워크에 접근 권한이다. 개방형 블록체인은 누구나 네트워크에 참

여하고 거래를 형성하고 승인할 수 있다. 허가형 블록체인은 승인된 사용자만 참여가 가능하며 승인된 기관만 거래를 생성이 가능하다. 악의적 참여자에 의한 해킹에 대해서 허가형 블록체인이 공개형 블록체인보다 안전하다.

공개형 블록체인은 채굴을 해서 거래를 증명하고 보상에 대해서 약속한다. 채굴방식으로 PoW(작업증명), PoS(지분증명), DPoS(위임지분증명), PoI(기여증명) 등을 사용한다. 공개형 블록체인은 다수의 사람이 이용하기 때문에 대형 네트워크가 형성되고 다수의 사람이 검증한다. 반면에 허가형 블록체인은 소수의 승인 기관이 거래를 검증 및 수정을 한다. 허가형 블록체인은 작은 네트워크 크기로 인해 거래를 통제 가 가능하지만 투명성과 보안성이 부족하다.

IV. 결 론

이 논문에서는 블록체인 개발현황과 특징에 대해서 알아보았다. 세대가 증가할수록 기능과 거래 속도 등을 강조하였지만 상대적으로 보안성은 낮아지는 특징을 보였다. 블록체인은 공개형태에 따라서 공개형 블록체인과 허가형 블록체인으로 분류된다. 공개형 블록체인은 사람들에게 네트워크에 자유롭게 참여 가능하여 트랜잭션을 모든 사람과 공유하며 안정적인 거래가 가능하다. 허가형 블록체인은 네트워크에 허가된 사용자만 참여할 수 있으며 변형된 형태의 블록체인을 사용한다. 블록체인의 공개형태에 따라 네트워크 크기가 다르며 투명성과 보안성이 달라졌다. 향후 연구로 마이닝 방법 즉, 작업증명(PoW, Proof-of-work)과 지분증명(PoS, Proof-of-stake) 등에 따른 보안성에 대해 연구를 진행할 예정이다.

참 고 문 헌

- [1] 박선중, 김용재, 오석은, "중앙은행 디지털화폐 연구 - 제1부 중앙은행의 디지털화폐 발행 시 법률적 쟁점", 한국은행 금융결제국]
- [2] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008)
- [3] <https://www.bitcoincash.org/>
- [4] <https://bitcoingold.org/>

[표 5] 블록체인 유형별 특징

	개방형 블록체인	허가형 블록체인
특징	<ul style="list-style-type: none"> - 네트워크에 자유로운 참여 가능 - 암호화폐를 이용한 네트워크 유지 및 공유경제 - 트랜잭션을 모든 노드와 공유, 안전한 거래 처리를 위한 프로세스 유지 	<ul style="list-style-type: none"> - 네트워크 노드 참여자를 승인 및 제한 - 암호화폐 부재 - 신속한 트랜잭션 처리, 데이터에 대한 프라이버시 중심
보안성	<ul style="list-style-type: none"> - 높은 수준의 보안성 유지 - 해킹을 위해 큰 비용 소모 요구 	<ul style="list-style-type: none"> - 낮은 수준의 보안성 유지 (불투명성) - 참여자를 식별하여 해킹방지 - 데이터를 일반인에게 공유하지 않음
적용영역	<ul style="list-style-type: none"> - 탈중앙형 트랜잭션 처리가 요구되는 영역 - 암호화폐를 이용한 서비스가 가능한 영역 - 일반인들의 참여와 데이터의 공개가 가능한 영역 	<ul style="list-style-type: none"> - 중재자의 역할이 필요하며 트랜잭션 처리가 필요한 영역 - 기업/기관 등 참여자를 사전에 특정할 수 있는 서비스

- [5] <http://btcd.io/>
- [6] <https://ripple.com/>
- [7] <https://www.stellar.org/>
- [8] <https://litecoin.com/>
- [9] <https://z.cash/>
- [10] <https://getmonero.org/>
- [11] <https://www.dash.org/>
- [12] <https://www.factom.com/>
- [13] <https://zcoin.io/>
- [14] <https://navcoin.org/>
- [15] <https://sia.tech/>
- [16] <https://www.burst-coin.org/>
- [17] <https://storj.io/>
- [18] <https://nem.io/>
- [19] <https://vertcoin.org/>
- [20] <https://www.digibyte.co/>
- [21] <https://www.ethereum.org/>
- [22] <https://qtum.org/>
- [23] <https://www.iota.org/>
- [24] <https://www.cardano.org/zh/home-3/>
- [25] <https://eos.io/>



이 흥 노 (Heung-No Lee)
정회원

1993년 : University of California
전기공학과 졸업

1994년 : University of California
전기공학과 석사

1999년 : University of California
전기공학과 박사

1999년~2002년 : HRL Laboratories Research Staff
Member

2002년~2008년 : University of Pittsburgh Assistant
Professor

2009년~현재 : 광주과학기술원 전기전자컴퓨터공학부 교수
관심분야: 정보이론, 신호처리, 통신/네트워크, 무선 통신인
및 네트워킹, 압축센싱

〈저자소개〉



정 현 준 (Hyunjun Jung)

2008년 2월 : 삼육대학교 컴퓨터과
학과 졸업

2010년 2월 : 숭실대학교 컴퓨터학
과 석사

2017년 8월 : 고려대학교 컴퓨터전
과통신공학과 박사

2017 9월~현재 : 광주과학기술원 센
서지능화연구센터 연구원

관심분야: 사물인터넷, 인공지능, 블록체인

Article

Compressive Sensing Spectroscopy Using a Residual Convolutional Neural Network

Cheolsun Kim, Dongju Park and Heung-No Lee *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea; csk0315@gist.ac.kr (C.K.); toriving@gist.ac.kr (D.P.)

* Correspondence: heungno@gist.ac.kr; Tel.: +82-62-715-2237

Received: 20 December 2019; Accepted: 20 January 2020; Published: 21 January 2020



Abstract: Compressive sensing (CS) spectroscopy is well known for developing a compact spectrometer which consists of two parts: compressively measuring an input spectrum and recovering the spectrum using reconstruction techniques. Our goal here is to propose a novel residual convolutional neural network (ResCNN) for reconstructing the spectrum from the compressed measurements. The proposed ResCNN comprises learnable layers and a residual connection between the input and the output of these learnable layers. The ResCNN is trained using both synthetic and measured spectral datasets. The results demonstrate that ResCNN shows better spectral recovery performance in terms of average root mean squared errors (RMSEs) and peak signal to noise ratios (PSNRs) than existing approaches such as the sparse recovery methods and the spectral recovery using CNN. Unlike sparse recovery methods, ResCNN does not require *a priori* knowledge of a sparsifying basis nor prior information on the spectral features of the dataset. Moreover, ResCNN produces stable reconstructions under noisy conditions. Finally, ResCNN is converged faster than CNN.

Keywords: spectroscopy; compressed sensing; deep learning; inverse problems; sparse recovery; dictionary learning

1. Introduction

There has been considerable interest in producing compact spectrometers having a high spectral resolution, wide working range, and short measuring time. Such a spectrometer can be used in a broad range of fields such as remote sensing [1], forensics [2], and medical applications [3]. Spectrometers that exploit advanced signal-processing methods are promising candidates. The compressive sensing (CS) [4,5] framework makes it possible for a spectrometer to improve its spectral resolution while retaining its compact size. CS spectroscopy comprises two parts: Capturing a spectrum with a small number of compressed measurements and reconstructing the spectrum from the compressed measurements using reconstruction techniques.

To date, for effective signal recovery in CS spectroscopy, three requirements should be satisfied. First, the spectrum should be a sparse signal or capable of sparse representation on a certain basis. Second, the sensing patterns of optical structures should be designed to have a small mutual coherence [6]. Third, appropriate reconstruction algorithms are required. Note that several sparsifying bases have been used in CS spectroscopy such as a family of orthogonal Daubechies wavelets [7], a Gaussian line shape matrix [8,9], and a learned dictionary [10]. Furthermore, numerous optical structures have been proposed to attain the necessary small mutual coherence for sensing patterns such as thin-film filters [11,12], a liquid crystal phase retarder [13], Fabry–Perot filters [7,14], and photonic crystal slabs [15,16]. As algorithms for reconstructing the original signal, two types of basic reconstruction techniques have been developed: greedy iterative algorithms [17,18] and convex relaxation [19,20]. In CS spectroscopy, the reconstruction algorithms have been used with a sparsity

constraint. Additionally, a non-negativity constraint is used in Reference [16,21]. Combining these three considerations, CS spectrometers have shown stable performance for light-emitting diodes (LEDs) and monochromatic lights.

Since not all signals can be represented as sparse on a fixed basis, prior information on structural features of the spectral dataset is therefore required to generate a best-fit sparsifying basis. Furthermore, a high computational cost is required for reconstruction techniques. Recently, deep learning [22] has been emerging as a promising alternative framework for reconstructing the original signal from the compressed measurements.

Mousavi et al. [23] was the first study on image recovery from structured measurements using deep learning. Moreover, a deep-learning framework for inverse problems has been applied in biomedical imaging for imaging through scattering media [24], magnetic resonance imaging [25,26], and X-ray computed tomography [27]. Kim et al. [28] reported the first attempt to use deep learning in CS spectroscopy. They trained a convolutional neural network (CNN) to output the reconstructed signal from the network. From here on the network reported by Kim et al. will be referred to as CNN.

Unlike CNN [28] in which learnable layers were simply stacked and trained to directly reconstruct the original spectrum, we make a residual connection [29] between the input and output of CNN and train the network to reconstruct the original spectrum by referring the input of the network. As a result, the network learns residuals between the input of the network and the original spectrum. It has been reported that it is easier to train a network when using residual connections than to train a plain network that was simply stacked with learnable layers [25,29]. Lee et al. [25] analyzed the topological structure of magnetic resonance (MR) images and the residuals of MR images. They showed that the residuals possessed a simpler topological structure, thus making learning residuals easier than learning the original MR images. In addition, He et al. [29] demonstrated with empirical results that the residual networks are easy to optimize and they achieved improvements in image-recognition tasks. From these works, we gain insights such that adding residual connections to CNN would improve the spectral reconstruction performance in CS spectroscopy.

In this paper, we aim to propose a novel residual convolutional neural network (ResCNN) for recovering an input spectrum from the compressed sensing measurements in CS spectroscopy. The novelty lies in the proposed ResCNN structure, with a moderate depth of learnable layers and a single residual connection, which provides the desired spectral reconstruction performance. The desired performance here means that the proposed ResCNN offers a performance which is better than that of CNNs as well as that of CS reconstruction with its sparsifying base known. In CS reconstruction, the prior knowledge of a fixed sparsifying basis is useful and offers good sparse representation results. However, in general it is a difficult problem to identify a sparsifying basis for various kinds of spectra and apply the identified basis to have the recovery performance improved. In this regard, it is an important advance to find a simple ResCNN which offers good enough performance. It is also worth to note that the proposed ResCNN is tested with the array type CS spectroscopy, discussed in Section 2, which we have designed with an array of multilayer thin-film filters.

The previous works on CS spectroscopy [7,11,13,14,16] have shown decent reconstruction performance but on limited simple sources such as LEDs and monochromatic lights. Using ResCNNs, we are now able to reconstruct more complex spectra, such as spectra with multiplicity of peaks mixed with a gradual rise-and-fall.

The remainder of this paper is organized as follows. In Section 2, we model the optical structure which is used for CS spectroscopy. In Section 3, we describe the system of CS spectroscopy and the proposed ResCNN. In Section 4, simulated experiments are described. Section 5 presents the results of experiments. In Section 6, we discuss the results. Finally, we conclude this paper in Section 7.

2. Optical Structure

Numerous optical structures have been proposed for CS spectroscopy. It has been reported that CS spectrometers, which have various spectral features in the transmission spectrum, show high

spectral-resolving performance [16]. In this work, we used thin-film filters to model CS spectrometers. Thin-film filters demonstrate a variety of spectral features depending on the materials used, the number of layers, and the thicknesses of the layers. Once the structure of thin-film is determined, a transmission value at a given wavelength λ is defined as follows [30]:

$$T(\lambda) = 1 - \frac{1}{2} \left(|\rho_{TE}(\lambda)|^2 + |\rho_{TM}(\lambda)|^2 \right), \quad (1)$$

where $\rho_{TE}(\lambda)$ and $\rho_{TM}(\lambda)$ are amplitude reflection coefficients. The coefficients represent the fraction of the power reflected by a multilayer thin-film in the transverse electric (TE) and transverse magnetic (TM) modes of an incident light, respectively. We summarized recursive processes for calculating amplitude reflection coefficients in Algorithm 1 [11,12,31].

Algorithm 1: Recursive processes for amplitude reflection coefficients.

Input:	λ Structure parameters: $\theta_1, \mathbf{n} = \{n_1, n_2, \dots, n_l\}, \mathbf{d} = \{d_2, d_3, \dots, d_l\}$.
Step 1:	Calculate θ_k, β_k , and N_k using structure parameters. $\theta_k = \sin^{-1} \left(\frac{n_{k-1}}{n_k} \sin \theta_{k-1} \right)$, for $k = 2, 3, \dots, l$. $\beta_k = 2\pi \cos(\theta_k) n_k d_k / \lambda$, for $k = 2, 3, \dots, l$. $N_k = \begin{cases} n_k / \cos \theta_k & \text{for TE} \\ n_k \cos \theta_k & \text{for TM} \end{cases}$, for $k = 2, 3, \dots, l$.
Step 2:	Obtain η_2 by setting $\eta_l = N_l$. For $k = l-1$ to 2 $\eta_k = N_k \frac{\eta_{k+1} \cos \beta_k + j N_k \sin \beta_k}{N_k \cos \beta_k + j \eta_{k+1} \sin \beta_k}$.
Step 3:	Compute $\rho = (N_1 - \eta_2) / (N_1 + \eta_2)$.
Output:	ρ

Here, θ_k is the angle of an incident light passing from k^{th} to $k+1^{\text{th}}$ layer. The refractive index of k^{th} layer is denoted as n_k . d_k denotes the thickness of the k^{th} layer. Given a wavelength vector $\lambda = (\lambda_1 \lambda_2 \dots \lambda_N) \in \mathbb{R}^{1 \times N}$ in the range of interest, i.e., $\lambda_{\max} - \lambda_{\min}$. Let $\Delta\lambda = \frac{\lambda_{\max} - \lambda_{\min}}{N}$. Then, evaluating the function at the integer multiple of $\Delta\lambda$, i.e., $T(\lambda = \lambda_{\min} + n\Delta\lambda)$ for $n = 0, 1, \dots, N-1$, we obtain the vector of transmission spectrum $\mathbf{T}_m \in \mathbb{R}^{1 \times N}$ for the wavelength range. Then, the sensing pattern matrix of optical structures $\mathbf{T} \in \mathbb{R}^{M \times N}$ is obtained by repeating the calculation of \mathbf{T}_m for $m = 1, 2, \dots, M$.

We have used SiNx and SiO₂ for high- and low-refractive index materials, respectively. We numerically generated thin-film filters by alternately stacking high- and low-refractive index materials, changing the number of layers, and varying the thickness of each layer. The number of layers in each filter is in the interval of (19, 24), and the thickness (nm) of each layer is in the interval of (50, 300). Initially, we randomly generated reference filters and compute the mutual coherence among the filters. Then, new filters were generated by changing thicknesses of the layers and the mutual coherence of the filters is compared to the mutual coherence of reference filters. Filters with a smaller mutual coherence then became the new reference filters. This process is repeated until reasonable reference filters with the required small mutual coherence are obtained.

Figure 1 shows the heatmap for the transmission spectra of the reference filters and two selected transmission spectra. In Figure 1a, each of the transmission spectra shows a unique sensing pattern because of the iterative modeling process of the reference filters based on mutual coherence. Figure 1b shows two transmission spectra that correspond to the 15th and 30th rows in the heatmap of reference filters. The transmission spectrum reveals a deep spectral modulation depth and various features such as broadband backgrounds, multiple peaks with a small full width at half maximums (FWHMs), and irregular fluctuations.

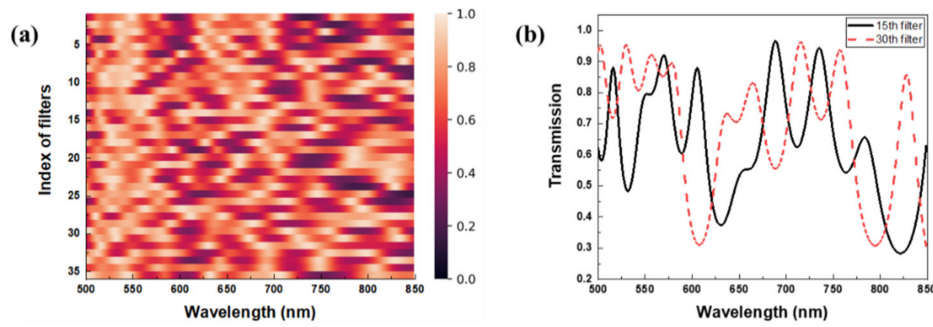


Figure 1. (a) Heatmap of the sensing matrix: each row represents the transmission spectrum of the designed thin-film filter. (b) Two transmission spectra corresponding to the 15th and 30th rows in the sensing matrix.

3. Compressive Sensing (CS) Spectrometers Using the Proposed Residual Convolutional Neural Network (ResCNN)

3.1. CS Spectrometers

In CS spectroscopy, the measurement column vector $\mathbf{y} \in \mathbb{R}^{M \times 1}$ is represented using the following relation:

$$\mathbf{y} = \mathbf{T}\mathbf{x}, \quad (2)$$

where $\mathbf{x} \in \mathbb{R}^{N \times 1}$ is the spectrum column vector of incident light and $\mathbf{T} \in \mathbb{R}^{M \times N}$ is the sensing matrix of the optical structure. Each row of \mathbf{T} represents a transmission spectrum. Because the length of the measurement vector is smaller than the length of the spectrum vector ($M < N$), the system is underdetermined. Conventionally, if \mathbf{x} is a sparse signal or can be sparsely represented in a certain basis, i.e., $\mathbf{x} = \Phi\mathbf{s}$, reconstruction algorithms can determine a unique sparse solution $\hat{\mathbf{S}}$ from the following optimization problem:

$$\min_{\mathbf{s}} \|\mathbf{T}\Phi\mathbf{s} - \mathbf{y}\|_2^2 + \tau\|\mathbf{s}\|_1, \quad (3)$$

where $\Phi \in \mathbb{R}^{N \times N}$ is a sparsifying basis and τ is a regularization parameter. Here, \mathbf{s} is a sparse signal whose components are zero except for a small number of non-zero components. Then, the recovered spectrum $\hat{\mathbf{x}}$ is $\Phi\hat{\mathbf{s}}$. In this paper, we refer to the methods of solving the optimization problem using Equation (3) as sparse recovery.

Typically, except for narrow-band spectra, a spectrum is not a sparse signal, and a fixed sparsifying basis cannot transform all spectra into sparse signals. Clearly, the use of a fixed basis may lead the sparse recovery to struggle, as no fixed basis will transform every signal into a sparse signal. In addition, the sparse recovery is time-consuming and takes a high computational cost.

Our goal is to overcome the limitations of the sparse recovery in CS spectroscopy and recover various kinds of spectra using ResCNN. Figure 2 shows the schematic of the CS spectroscopy system using ResCNN. This system consists of two parts: compressive sampling and dimension extension, and the reconstruction using ResCNN. In the compressive sampling and dimension extension, the measurement vector \mathbf{y} is obtained from Equation (1), which then transforms into $\tilde{\mathbf{x}} \in \mathbb{R}^{N \times 1}$ using a linear transformation. A transform matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$ extends the M dimension of \mathbf{y} to N dimension of $\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}}$ is a representative spectrum corresponding to \mathbf{x} . We used $\tilde{\mathbf{x}}$ as the input for the reconstruction. ResCNN learnt a non-linear mapping between $\tilde{\mathbf{x}}$ and \mathbf{x} , and afforded a reconstructed spectrum $\hat{\mathbf{x}} \in \mathbb{R}^{N \times 1}$. The dimension extension by the transform matrix was used to make it easier for ResCNN to extract features and reconstruct spectra from the non-linear mapping.

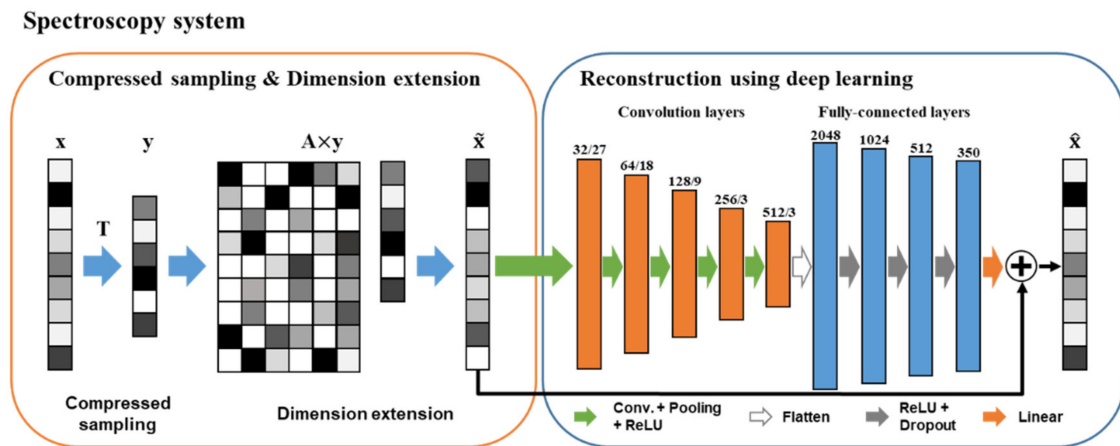


Figure 2. Overview of compressive sensing (CS) spectroscopy system including the proposed residual convolutional neural network (ResCNN): An input spectrum is compressively sampled by the sensing matrix, and the dimension of measurements is extended by the transform matrix. ResCNN is trained to recover the original spectrum from the extended measurements.

3.2. The Proposed ResCNN

As depicted in Figure 2, ResCNN comprises nine learnable layers, five of which are convolution layers, four are fully-connected layers, and one is a residual connection. Convolution layers are used for the feature extraction in the non-linear mapping between \tilde{x} and x . Fully-connected layers are used for the spectra reconstruction. Each of the convolution layers has a set of one-dimensional learnable kernels with specific window sizes. The number of kernels and the window sizes are indicated in Figure 2. After every convolutional layer, the rectified linear unit (ReLU) is used as an activation function, and the subsampling is then applied. We use non-overlapping max-pooling to down-sample the output of the activation function. We stack the convolutional layer, the ReLU, and the subsampling five times. The output of the last subsampling is flattened and then fed into the subsequent four fully-connected layers. The first three layers are followed by the ReLU and dropout in sequence. The dropout is introduced to reduce the overfitting of ResCNN. The output of the last fully-connected layer is fed into a linear activation function. The number of units in each of the fully-connected layers is noted in Figure 2. Unlike CNN [28] in which learnable layers are simply stacked, we make the residual connection that the representative spectrum \tilde{x} and the output of the linear activation function are added up to the reconstructed spectrum \hat{x} . Consequently, \hat{x} is trained to become x . Given training data $\{x_t^i\}_{i=1}^k$, we train ResCNN to minimize a loss function L . We use the mean squared error between the original x_t and recovered \hat{x}_t as the loss function:

$$L = \frac{1}{k} \sum_{i=1}^k \|x_t^i - \hat{x}_t^i\|_2^2. \quad (4)$$

The non-linear mapping that \tilde{x} becomes x can be defined as $H(\tilde{x}) = x$. Because of the residual connection in ResCNN, $H(\tilde{x})$ can be rewritten as $H(\tilde{x}) = F(\tilde{x}) + \tilde{x}$, where $F(\tilde{x})$ is the mapping of the learnable layers. The representative spectrum \tilde{x} is referenced by the residual connection, and then, $F(\tilde{x}) = H(\tilde{x}) - \tilde{x}$. In particular, the mapping of $F(\tilde{x})$ is called a residual mapping; therefore, the learnable layers learn the residual of x and \tilde{x} .

The previous researches [25,29] have used numerous residual connections in very deep neural networks in order to make networks converge faster by avoiding vanishing gradient problems. We use one residual connection between input and output of the moderate depth network. Figure 3 depicts the manner in which a spectrum is recovered in CNN and ResCNN. The learnable layers of CNN directly reconstruct the spectrum from the representative spectrum \tilde{x} . Alternatively, ResCNN reconstructs the

spectrum by passing the representative spectrum \tilde{x} through the residual connection shown in Figure 3b. Consequently, the learnable layers of ResCNN learn to reconstruct residuals.

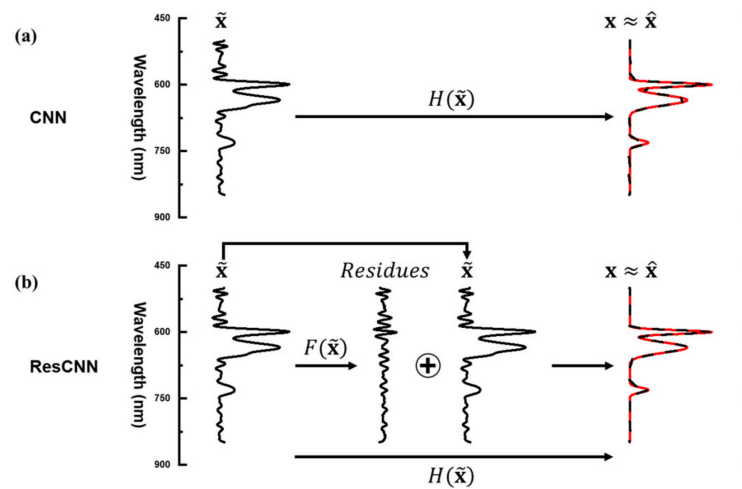


Figure 3. Descriptions of the spectrum recovery process: (a) convolutional neural network (CNN), (b) ResCNN.

4. Simulated Experiments

We reconstructed 350 spectral bands ($N = 350$) using 36 thin-film filters ($M = 36$) whose sensing patterns have a spacing of 1 nm for wavelengths from 500 to 850 nm. We determined the sensing matrix \mathbf{T} , assuming that the incident light falls onto the filters with normal incidence. As the transform matrix \mathbf{A} , we used the Moore–Penrose inverse of the sensing matrix \mathbf{T} , i.e., $\mathbf{A} = \mathbf{T}^T(\mathbf{T}\mathbf{T}^T)^{-1}$.

4.1. Spectral Datasets

To evaluate the performance of ResCNN, we used two synthetic spectral datasets and two measured spectral datasets. The first synthetic dataset is composed of Gaussian distribution functions while the other is composed of Lorentzian distribution functions. These two synthetic datasets were selected as generally these types of functions are used to represent spectral line shapes. As shown in Figure 4, component functions are added to produce the spectra. We generated 12,000 spectra for each dataset. For each spectrum, the number of component functions was generated using a geometric distribution with the probability parameter p set to 0.3. We added one to the number of component functions to prevent the number of component functions from becoming zero. Then, we randomly set a location, a height, and an FWHM of each peak. To set a peak location (nm), an integer number was randomly selected from a uniform distribution with the interval (500, 849). A random number from a uniform distribution in the interval (0, 1) was used for the height. An integer number for an FWHM (nm) was randomly drawn from a uniform distribution with the interval (2, 50). Finally, all of the component functions were summed to generate the spectrum. The height of each generated spectrum was normalized such that it was mapped from zero to one.

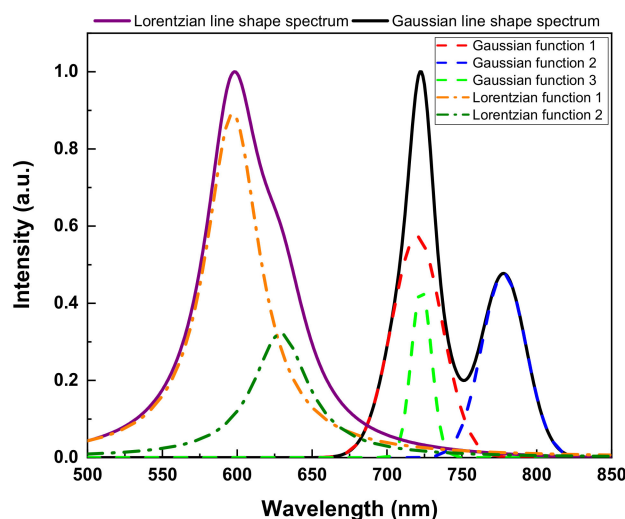


Figure 4. Examples of two synthetic spectra: the solid purple line is composed of two Lorentzian distribution functions (dash-dotted orange and olive lines), and the solid black line is composed of three Gaussian distribution functions (dashed red, blue, and green lines).

As measured datasets, we used the US Geological Survey (USGS) spectral library version 7 [32], and the glossy Munsell colors spectral dataset [33]. The USGS spectral library provides reflectance spectra for artificial materials, coatings, liquids, minerals, organic compounds, soil mixtures, and vegetation. We discarded any spectrum that has missing spectral bands. Then, we extracted the spectrum in the wavelength range of interest (500 to 849 nm) from the wavelength range of the original spectrum (350 to 2500 nm). The measured wavelength range for the glossy Munsell colors spectral dataset, which contains the reflectance spectra of the glossy Munsell color chips, was 380 to 780 nm. The wavelength range of the original spectrum was different from the wavelength range of interest. We decided to use the wavelength range from 400 to 749 nm to ensure each spectrum was set to 350 spectral bands. This selection of wavelengths is reasonable because the wavelengths were located in the center of the wavelength range of the original spectrum, and showed different spectral features with respect to each spectrum. In addition, our aim was to show the reconstruction performance with respect to various kinds of spectra. Finally, each spectrum was normalized such that the height varies from 0 to 1. Overall, 1473 spectra from USGS spectral dataset and 1600 spectra from Munsell color spectral dataset were used for our simulated experiments. Table 1 lists the details of each of the spectral datasets.

Table 1. Description of the spectral datasets.

Dataset	Training/Validation/Test	Avg. Number of Nonzero Values	Description
Gaussian dataset	8000/2000/2000	336.8/350	FWHM (nm) on the interval [2, 50], Height on the interval [0, 1]
Lorentzian dataset	8000/2000/2000	349/350	FWHM (nm) on the interval [2, 50], Height on the interval [0, 1]
US Geological Survey [32]	982/246/245	348.9/350	350–2500 nm, 2151 spectral bands (we use 350 spectral bands in 500–849 nm)
Munsell colors [33]	1066/267/267	349/350	380–780 nm, 401 spectral bands (we use 350 spectral bands in 400–749 nm)

4.2. Data Preprocessing and Training

Given the sensing matrix, the spectral data are compressively sampled as the measurement vector \mathbf{y} shown in Equation (1), and then transformed into the representative spectrum $\tilde{\mathbf{x}}$ by multiplying the transform matrix \mathbf{A} and \mathbf{y} .

In each spectral dataset, the number of training, validation, and test spectra are randomly assigned using a ratio of 4:1:1 for the synthetic and measured data sets, respectively. The validation spectra are used for estimating the number of epochs and tuning the hyper-parameters. To train ResCNN, we used the Adam optimizer [34] implemented in Tensorflow with the batch size of 16 and 250 epochs. The experiments were conducted on an NVIDIA GeForce RTX 2060 graphics processing unit (GPU). Training the architecture can be done in half an hour for each dataset.

4.3. Sparsifying Bases for Sparse Recovery

Using sparse recovery, we evaluated the performance of conventional CS reconstructions to benchmark the performance of ResCNN. As shown in Table 1, the spectra for both the synthetic and measured datasets are dense spectra. Therefore, we must transform the spectra into sparse signals to solve Equation (3). In this section, we considered methods to make a sparsifying basis Φ .

First, we considered a Gaussian line shape matrix as a sparsifying basis. Each column of the matrix comprises a Gaussian distribution function whose length is N . A collection of N Gaussian functions works as a sparsifying basis $\Phi \in \mathbb{R}^{N \times N}$. We generate two Gaussian line shape matrices. Figure 5a shows the heatmap images for two Gaussian line shape matrices. Seven different FWHMs are used to generate the Gaussian distributions. Given an FWHM, Gaussian distributions are generated by shifting the peak location using uniform spacing. To create a small dissimilarity between the two Gaussian line shape matrices, two of the seven FWHMs in Gaussian 1 were replaced with other FWHMs, thus producing Gaussian 2, as shown in Figure 5a.

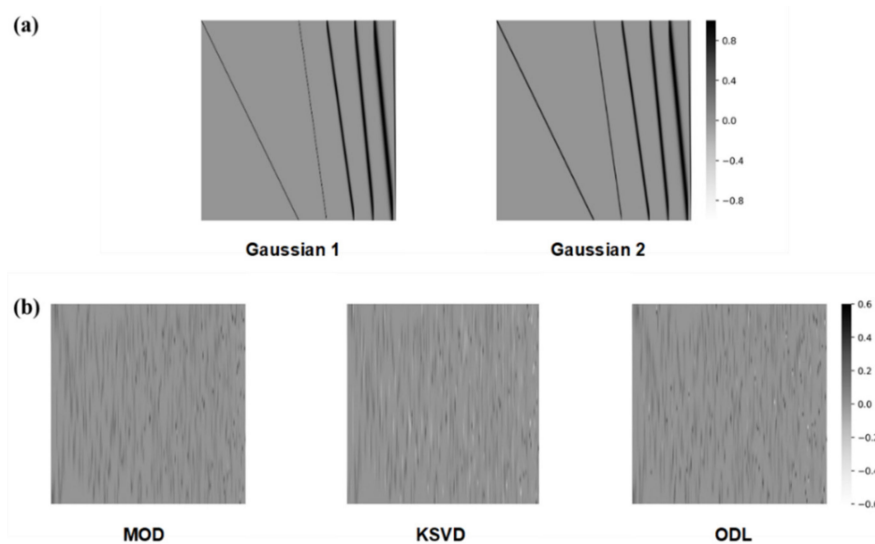


Figure 5. Heatmap images of sparsifying bases that were used in simulated experiments: (a) Gaussian line shape matrices, (b) the learned dictionaries which are from the Gaussian training dataset.

Second, a learned dictionary [35–38] is used as a sparsifying basis. Given a training dataset $\{\mathbf{x}_t^i\}_{i=1}^k$, we can derive a learned dictionary Φ that sparsely represents the training data \mathbf{x}_t by solving the following optimization problem, known as the dictionary learning problem:

$$\min_{\Phi, \mathbf{s}_1^1, \dots, \mathbf{s}_t^k} \sum_{i=1}^k \|\mathbf{x}_t^i - \Phi \mathbf{s}_t^i\|_2^2 + \tau \|\mathbf{s}_t^i\|_1, \quad (5)$$

where τ is a regularization parameter and \mathbf{s}_t^i is i th sparse signal over the training dataset. By fixing an initial guess for the dictionary Φ in Equation (5), we obtain a solution for the sparse signals $\{\mathbf{s}_t^i\}_{i=1}^k$. The dictionary is then updated by solving Equation (5) using the sparse signals obtained. This process is iteratively repeated until convergence is reached and we derive the learned dictionary. We used three dictionary learning methods: method of optimal directions (MOD) [36], K-SVD [37], and online dictionary learning (ODL) [38]. The learned dictionaries are generated for each of the training datasets, and the reconstruction performances are evaluated for each test dataset. Figure 5b shows learned dictionaries identified using the Gaussian training dataset. The learned dictionaries clearly depend on the dictionary-learning methods used. Nevertheless, each column of the dictionaries shows a learned spectral feature from the training dataset.

5. Results

To demonstrate the ability of ResCNN to reconstruct spectra, we evaluated its performance using three different datasets: Synthetic datasets, noisy synthetic datasets, and measured datasets. We used the same hyper-parameters of ResCNN for each of these datasets. Moreover, we adopted $l1_ls$ [39] as the fixed reconstruction algorithm in the sparse recovery. We compared the recovered signal with the original signal by calculating the root mean squared error (RMSE) and the peak signal to noise ratio (PSNR). In addition, the performance of five conventional sparse recovery methods, described in Section 4.3 and CNN was calculated.

5.1. Synthetic Datasets

The two synthetic data sets described in Table 1 were used to perform the signal recovery using sparse recovery and deep learning. Table 2 shows the average RMSE and PSNR for each of the seven methods evaluated. ResCNN shows the smallest average RMSE for both the Gaussian and Lorentzian datasets of 0.0094 and 0.0073, respectively. Moreover, ResCNN shows the largest average PSNR of 49.0 dB for the Lorentzian dataset. For the Gaussian dataset, the sparse recovery method with Gaussian 2 shows the largest average PSNR, 49.7 dB, which is slightly higher than the 47.2 dB for ResCNN. Note that the minor difference between the two Gaussian line shape matrices results in considerable performance difference. However, reconstruction using the learned dictionaries show similar performance across all of the synthetic datasets.

Table 2. Average root mean squared errors (RMSEs) and peak signal to noise ratios (PSNRs) over synthetic datasets.

Dataset	Sparse Recovery					Deep Learning	
	Gaussian 1	Gaussian 2	K-SVD	MOD	ODL	CNN	ResCNN
Gaussian dataset	0.0226 (43.1 dB)	0.0112 (49.7 dB)	0.0172 (40.3 dB)	0.0174 (40.3 dB)	0.0161 (41.1 dB)	0.0132 (40.5 dB)	0.0094 (47.2 dB)
Lorentzian dataset	0.0146 (44.9 dB)	0.0094 (47.5 dB)	0.0136 (42.3 dB)	0.0137 (42.3 dB)	0.0127 (42.9 dB)	0.0101 (42.8 dB)	0.0073 (49.0 dB)

Figure 6 shows the reconstructed test spectra from each of the synthetic datasets. The solid red line (i) is the input spectra from each dataset. ResCNN is shown in dashed black line (ii), while CNN is shown in solid orange lines (iii). The reconstructed spectra using sparse recovery with Gaussian 1 (iv), Gaussian 2 (v), and ODL (vi) are shown in solid green, blue, and purple lines in respectively. Because of the similar performance from each of the learned dictionaries, only the ODL method is shown. The RMSE and PSNR of ResCNN are 0.0138 (37.2 dB) for the spectrum from the Gaussian dataset and 0.0096 (40.4 dB) for the spectrum from the Lorentzian dataset. For the selected spectra, ResCNN achieves superior reconstruction performance compared with the other four reconstructions.

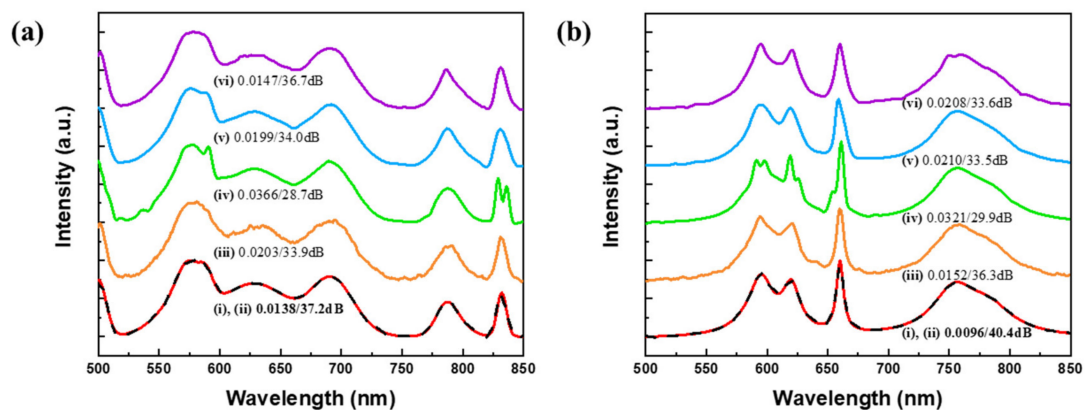


Figure 6. Spectral reconstructions of test spectra in synthetic datasets, (a) Gaussian dataset, (b) Lorentzian dataset. An input spectrum (solid red (i)) is compared with ResCNN (dashed black (ii)), CNN (orange (iii)), sparse recovery: Gaussian 1 (green (iv)), Gaussian 2 (blue (v)), and online dictionary learning (ODL) (purple (vi)). The baselines are shifted for clarity.

Only sparse recovery with Gaussian 1 fails to recover the fine details of the input spectrum. One example of the poor ability of sparse recovery with Gaussian 1 to resolve the signal is the recovery of the peak at ~ 830 and 590 nm being recovered as two neighboring peaks in Figure 6a,b, respectively. CNN was unable to capture the smoothness of the spectral features compared to the other methods.

5.2. Noisy Synthetic Datasets

To verify the stability of ResCNN, we evaluated the accuracy of the reconstruction at various noise levels. Gaussian white noise was added to the measurement vector $\mathbf{n} \in \mathbb{R}^{M \times 1}$ to Equation (2), i.e., $\mathbf{y} = \mathbf{T}\mathbf{x} + \mathbf{n}$. We considered six different noise levels whose signal-to-noise ratios (SNRs) are 15, 20, 25, 30, 35, and 40 dB. The SNR (dB) is defined as $10 \cdot \log_{10}(\|\mathbf{x}\|_2^2 / N\sigma^2)$, where σ is the standard deviation of the noise. Using Gaussian and Lorentzian datasets, we compared the reconstruction performance of ResCNN with the sparse recovery using Gaussian 2, which shows the best reconstruction performances among sparse recovery methods in synthetic datasets. ResCNN was evaluated with the same hyper-parameters that were used for the noise-free datasets. The average RMSE and PSNR for each of the six noise levels are shown in Table 3. While ResCNN was trained using noise-free data, it outperformed the sparse recovery with Gaussian 2 at every noise level, which indicates that ResCNN remains stable even with noisy datasets.

Table 3. Average RMSE and PSNR under various signal-to-noise ratios (SNRs, dB) with synthetic datasets.

		SNR (dB)					
Dataset	Method	15 dB	20 dB	25 dB	30 dB	35 dB	40 dB
Gaussian Dataset	Sparse recovery + Gaussian 2	0.0796 (22.7 dB)	0.0482 (27.1 dB)	0.0308 (31.2 dB)	0.0215 (34.8 dB)	0.0166 (37.9 dB)	0.0138 (40.7 dB)
	ResCNN	0.0671 (24.2 dB)	0.0401 (28.7 dB)	0.0251 (32.9 dB)	0.0171 (36.6 dB)	0.0130 (39.8 dB)	0.0110 (42.4 dB)
Lorentzian Dataset	Sparse recovery + Gaussian 2	0.0817 (22.6 dB)	0.0483 (27.1 dB)	0.0300 (31.2 dB)	0.0201 (35.0 dB)	0.0147 (38.5 dB)	0.0119 (41.4 dB)
	ResCNN	0.0689 (24.1 dB)	0.0404 (28.7 dB)	0.0243 (33.1 dB)	0.0157 (37.1 dB)	0.0113 (40.6 dB)	0.0091 (43.4 dB)

5.3. Measured Datasets

ResCNN was trained using the two measured datasets listed in Table 1, USGS and Munsell colors, and its reconstruction performance was evaluated. In addition, the signal reconstruction was performed using CNN and sparse recovery with five different sparsifying bases. Table 4 reports the average RMSE and PSNR for each of the seven methods. ResCNN achieves the smallest average RMSE and the largest average PSNR for both datasets. In the USGS dataset, the average RMSE and PSNR of ResCNN are 0.0048 and 52.4 dB, respectively. In addition, ResCNN achieves 0.0040 for the average RMSE and 50.0 dB for the average PSNR in the Munsell colors dataset. Similar to synthetic datasets, all of the learned dictionaries provided similar reconstruction performances. In addition, the small differences between Gaussian 1 and 2 show large differences in the RMSE and PSNR. The average RMSE and PSNR of the learned dictionary methods approach the values of ResCNN for Munsell colors dataset because the Munsell colors dataset has simpler spectral features than the other datasets.

Table 4. Average RMSEs and PSNRs for the measured datasets.

Dataset	Sparse Recovery					Deep Learning	
	Gaussian 1	Gaussian 2	K-SVD	MOD	ODL	CNN	ResCNN
USGS [32]	0.0081 (45.3 dB)	0.0061 (48.4 dB)	0.0070 (48.5 dB)	0.0081 (47.4 dB)	0.0074 (47.6 dB)	0.0116 (40.8 dB)	0.0048 (52.4 dB)
Munsell colors [33]	0.0068 (44.6 dB)	0.0050 (47.5 dB)	0.0040 (49.8 dB)	0.0040 (49.9 dB)	0.0042 (49.5 dB)	0.0076 (43.0 dB)	0.0040 (50.0 dB)

Figure 7 shows the reconstruction results of one test spectra from each of the measured datasets. The spectrum for the organic compound dibenzothiophene in the USGS dataset is reconstructed in Figure 7a. The spectrum of Munsell color 5 PB 2/2 is shown in Figure 7b. The solid red lines are the input spectra (i). ResCNN are shown in dashed black lines (ii), and CNN are shown in solid black lines (iii). The spectra of (iv) to (vi) are reconstructed spectra using the sparse recovery with Gaussian 1, Gaussian 2, and K-SVD. Because of the best performance of the K-SVD among the learned dictionaries only the K-SVD method is shown.

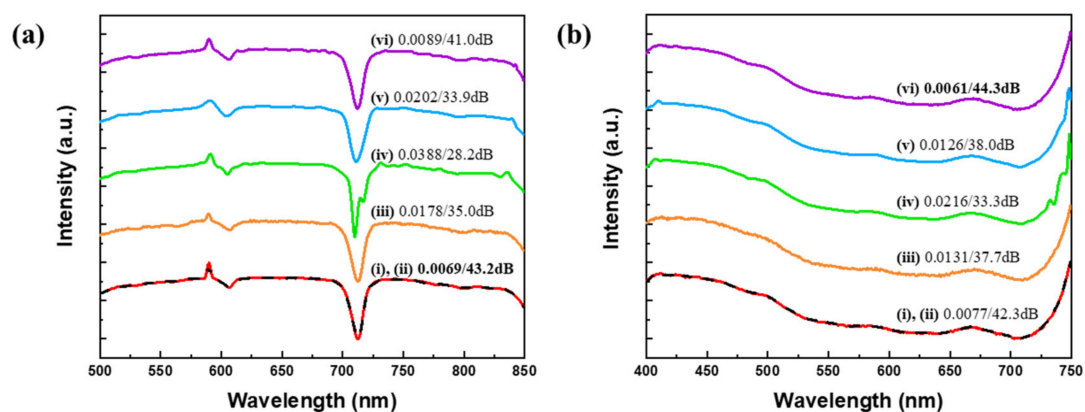


Figure 7. Spectral reconstructions of test spectra in measured datasets: (a) spectrum of organic compound dibenzothiophene in USGS dataset, (b) spectrum of Munsell color 5PB 2/2. The input spectrum (solid red line (i)) is compared with ResCNN (dashed black (ii)), CNN (orange (iii)), sparse recovery: Gaussian 1 (green (iv)), Gaussian 2 (blue (v)), and K-SVD (purple (vi)). The baselines are shifted for clarity.

The RMSE and PSNR for ResCNN are 0.0069 (43.2 dB) for the spectrum from the USGS dataset and 0.0077 (42.3 dB) for the spectrum from the Munsell colors dataset. ResCNN outperforms other approaches for the spectrum from USGS dataset. However, for the spectrum from Munsell colors

dataset, the sparse recovery with K-SVD outperforms ResCNN. ResCNN achieves slightly larger RMSE and smaller PSNR.

The performances of sparse recovery with Gaussian 2 is degraded for measured datasets compared with the performance for synthetic datasets. The measured datasets have rough spectral features unlike the smooth spectral features observed in the synthetic datasets. As a result, the sparse recovery with Gaussian 2 performs worse, because of its inability to represent rough spectral features using Gaussian distribution functions. The performance of sparse recovery with dictionary learning methods are improved for measured datasets compared with the performance of synthetic datasets. Because the number of spectra in measured datasets are smaller than the number of spectra in synthetic datasets. Therefore, finding the best-fit sparsifying basis for measured datasets is easier than finding the best-fit sparsifying basis for synthetic datasets using dictionary-learning methods. Meanwhile, ResCNN shows superior reconstruction performances regardless of spectral features of datasets and the size of datasets.

6. Discussion

As shown in the results, we demonstrate empirically that ResCNN outperforms the sparse recovery methods and the CNN over all datasets. The sparse recovery shows unstable performance because it is highly dependent on the sparsifying basis and spectral features of dataset. This is a direct result of being unable to identify a fixed sparsifying basis that can transform any spectra into a sparse signal, which means the *a priori* structural information such as line shapes and FWHMs is required to select a consistent sparsifying basis. Learned dictionaries are used to cope with the problem of identifying a consistent sparsifying basis. The columns of learned dictionaries are composed of learned spectral features from the training dataset. While this shows an improvement in measured datasets, a learned dictionary is still limited to representing all the spectral features in the large dataset (i.e., synthetic datasets) using linear combinations of columns of the learned dictionary.

Compression approaches for summarizing information with a small number of sensors were proposed in [40]. These approaches can be exploited to generate a sparsifying basis by reducing the loss of spectral information in large datasets.

To improve the reconstruction performance in sparse recovery, pre-defined structure information and side information of unknown target signals were used in [41,42]. The reconstruction of three-dimensional electrical impedance tomography was improved by updating three-dimensional structural correlations using pre-defined structured signals [41]. To recover multi-modal data, a reconstruction framework is proposed in [42] that uses side information in unrolled optimization. Unrolled optimization approaches using deep learning were proposed in [43,44]. Deep-learning architectures were used to train hyper-parameters, such as a gradient regularizer and a step size. Using learned hyper-parameters, it was shown optimized solutions can be obtained within a fixed number of iterations. These proposed approaches for image reconstruction have assumed random sensing matrix and structured or sparse signals. In this work, however, we consider dense spectra and the sensing matrix from thin-film filters for the real implementation. Moreover, the reconstruction performance may change to a sparsifying basis as shown in results because a reconstructed spectrum $\hat{\mathbf{x}}$ should be represented as a linear combination of columns of a fixed sparse basis Φ as $\Phi\hat{\mathbf{s}}$.

For recovering spectra, ResCNN does not require the *a priori* knowledge of a sparsifying basis or prior information of spectral features. During training, ResCNN learns the spectral features using learnable layers, which enable it to recover the fine details for various kinds of spectra without identifying a sparsifying basis.

ResCNN is directly compared with CNN for the synthetic Gaussian dataset in Figure 8a where the mean squared error (Equation (4)) is plotted with respect to the epoch. The mean squared error for CNN and ResCNN are shown in solid black line and solid red line with square symbols, respectively. ResCNN shows a lower mean squared error than that of CNN. Moreover, ResCNN converges faster than CNN, indicating that ResCNN optimizes the learnable layers quicker, as expected based on

previous research using residuals [25,29]. In contrast to the previous research that numerous residual connections were used in very deep neural networks to converge networks faster by avoiding vanishing gradient problem, we achieve spectral reconstruction improvements even with one residual connection in a moderate depth CNN.

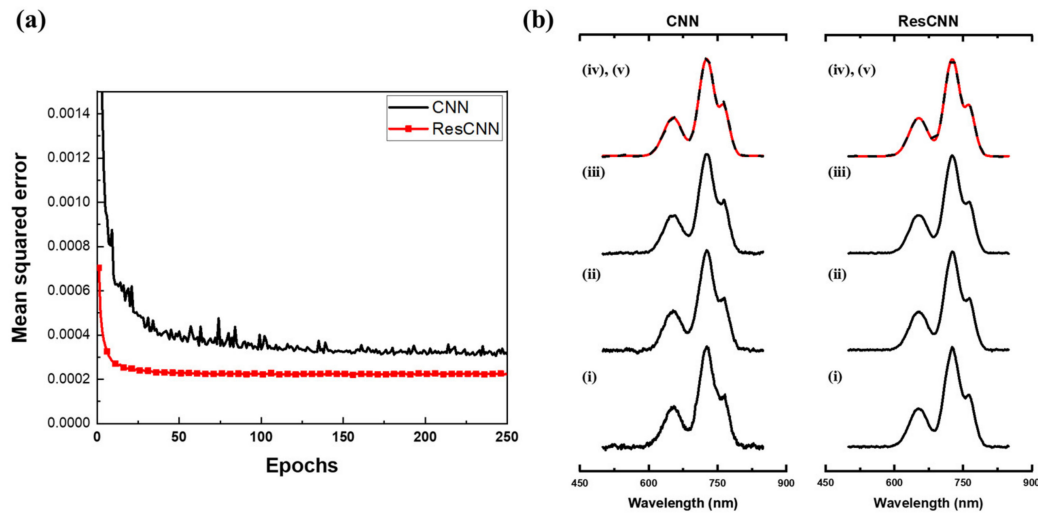


Figure 8. (a) Mean squared error of Gaussian dataset with respect to epochs. Solid black line denotes validation error of CNN, and solid red line with square symbols denotes validation error of ResCNN. (b) Reconstructions of a spectrum with respect to epochs where (i) to (iv) are epochs 1, 50, 150, and 250, respectively. Red line (v) denotes the original spectrum.

The reconstruction of an example spectrum with respect to the number of epochs is shown in Figure 8b. Black lines ((i) to (iv)) are the reconstructed spectra at 1, 50, 150, and 250 epochs, respectively. The solid red line (v) is the original spectrum, and the series of reconstructed spectrum for ResCNN show that the reconstruction converged earlier than CNN. The increased rate of convergence is because of the residual connection in ResCNN. Overall, the reconstruction performance of ResCNN is an improvement over CNN.

Note that both ResCNN and dictionary learning for sparse recovery require a training dataset and an optimization process to learn the spectral features. While this is a time-consuming process, remember that when using a learned dictionary to recover spectra, an iterative reconstruction algorithm is required, which needs additional time and incurs a high computational cost. The benefit of ResCNN is that it gives a reconstructed spectrum immediately once the training is completed.

7. Conclusions

In this paper, we propose a novel ResCNN for recovering the input spectrum from the compressed measurements in CS spectroscopy. As the optical structure for CS spectroscopy, we numerically generated multilayer thin-film filters which have a small mutual coherence. Therefore, we could compressively measure input spectra with unique sensing patterns. To reconstruct the input spectra from the compressively sampled measurements, we modeled ResCNN, which has a moderate-depth of learnable layers and a residual connection. We stacked nine learnable layers: five convolutional layers and four fully-connected layers with a single residual connection between the input and output of the learnable layers. The measurements were extended by a linear transformation and then fed into ResCNN. Finally, ResCNN reconstructed the input spectra. We demonstrated the empirical reconstruction results for ResCNN using synthetic and measured datasets. We compared the reconstruction performance of ResCNN with sparse recovery using five different sparsifying bases and CNN. Compared with sparse recovery methods, ResCNN shows better reconstruction performance without the *a priori* knowledge of either a sparsifying basis or any spectral features of the spectral datasets. On the other hand, the

sparse recovery methods show deviation of reconstruction performances to sparsifying bases and spectral datasets, meaning that a fixed sparsifying basis cannot represent all spectral features of input spectra. Furthermore, ResCNN shows stable reconstruction performances under noisy environments. Compared with CNN, ResCNN shows significant improvement in reconstruction performance and converges faster than CNN. In future work, we will explore compression approaches [40] and unrolled optimization approaches [43,44] for generating a sparsifying basis Φ from the training dataset to fully represent spectra without loss of spectral features.

Author Contributions: Conceptualization, C.K. and H.-N.L.; methodology, C.K.; software, C.K. and D.P.; formal analysis, C.K. and D.P.; investigation, C.K.; data curation, C.K.; writing—original draft preparation, C.K.; writing—review and editing, C.K., D.P. and H.-N.L.; project administration, H.-N.L.; funding acquisition, H.-N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2018R1A2A1A19018665).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Clark, R.N.; Roush, T.L. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *J. Geophys. Res. Solid Earth* **1984**, *89*, 6329–6340. [[CrossRef](#)]
- Izake, E.L. Forensic and homeland security applications of modern portable Raman spectroscopy. *Forensic Sci. Int.* **2010**, *202*, 1–8. [[CrossRef](#)] [[PubMed](#)]
- Kim, S.; Cho, D.; Kim, J.; Kim, M.; Youn, S.; Jang, J.E.; Je, M.; Lee, D.H.; Lee, B.; Farkas, D.L.; et al. Smartphone-based multispectral imaging: System development and potential for mobile skin diagnosis. *Biomed. Opt. Express* **2016**, *7*, 5294–5307. [[CrossRef](#)] [[PubMed](#)]
- Donoho, D.L. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [[CrossRef](#)]
- Eldar, Y.C.; Kutyniok, G. *Compressed Sensing: Theory and Applications*; Cambridge University Press: Cambridge, UK, 2012.
- Candes, E.J.; Eldar, Y.C.; Needell, D.; Randall, P. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harm. Anal.* **2011**, *31*, 59–73. [[CrossRef](#)]
- Oiknine, Y.; August, I.; Blumberg, D.G.; Stern, A. Compressive sensing resonator spectroscopy. *Opt. Lett.* **2017**, *42*, 25–28. [[CrossRef](#)]
- Kurokawa, U.; Choi, B.I.; Chang, C.-C. Filter-based miniature spectrometers: Spectrum reconstruction using adaptive regularization. *IEEE Sens. J.* **2011**, *11*, 1556–1563. [[CrossRef](#)]
- Cerjan, B.; Halas, N.J. Toward a Nanophotonic Nose: A Compressive Sensing-Enhanced, Optoelectronic Mid-Infrared Spectrometer. *ACS Photonics* **2018**, *6*, 79–86. [[CrossRef](#)]
- Oiknine, Y.; August, I.; Stern, A. Multi-aperture snapshot compressive hyperspectral camera. *Opt. Lett.* **2018**, *43*, 5042–5045. [[CrossRef](#)]
- Kim, C.; Lee, W.-B.; Lee, S.K.; Lee, Y.T.; Lee, H.-N. Fabrication of 2D thin-film filter-array for compressive sensing spectroscopy. *Opt. Lasers Eng.* **2019**, *115*, 53–58. [[CrossRef](#)]
- Oliver, J.; Lee, W.-B.; Lee, H.-N. Filters with random transmittance for improving resolution in filter-array-based spectrometers. *Opt. Express* **2013**, *21*, 3969–3989. [[CrossRef](#)] [[PubMed](#)]
- August, Y.; Stern, A. Compressive sensing spectrometry based on liquid crystal devices. *Opt. Lett.* **2013**, *38*, 4996–4999. [[CrossRef](#)] [[PubMed](#)]
- Huang, E.; Ma, Q.; Liu, Z. Etalon Array Reconstructive Spectrometry. *Sci. Rep.* **2017**, *7*, 40693. [[CrossRef](#)] [[PubMed](#)]
- Wang, Z.; Yu, Z. Spectral analysis based on compressive sensing in nanophotonic structures. *Opt. Express* **2014**, *22*, 25608–25614. [[CrossRef](#)] [[PubMed](#)]
- Wang, Z.; Yi, S.; Chen, A.; Zhou, M.; Luk, T.S.; James, A.; Nogan, J.; Ross, W.; Joe, G.; Shahsafi, A. Single-shot on-chip spectral sensors based on photonic crystal slabs. *Nat. Commun.* **2019**, *10*, 1020. [[CrossRef](#)]
- Pati, Y.C.; Rezaifar, R.; Krishnaprasad, P.S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–3 November 1993; pp. 40–44.

18. Dai, W.; Milenkovic, O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inf. Theory* **2009**, *55*, 2230–2249. [CrossRef]
19. Chen, S.S.; Donoho, D.L.; Saunders, M.A. Atomic decomposition by basis pursuit. *SIAM Rev.* **2001**, *43*, 129–159. [CrossRef]
20. Candes, E.; Tao, T. Decoding by linear programming. *arXiv* **2005**, arXiv:math/0502327. [CrossRef]
21. Oliver, J.; Lee, W.; Park, S.; Lee, H.-N. Improving resolution of miniature spectrometers by exploiting sparse nature of signals. *Opt. Express* **2012**, *20*, 2613–2625. [CrossRef]
22. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef]
23. Mousavi, A.; Baraniuk, R.G. Learning to invert: Signal recovery via deep convolutional networks. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2272–2276.
24. Li, Y.; Xue, Y.; Tian, L. Deep speckle correlation: A deep learning approach toward scalable imaging through scattering media. *Optica* **2018**, *5*, 1181–1190. [CrossRef]
25. Lee, D.; Yoo, J.; Ye, J.C. Deep residual learning for compressed sensing MRI. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, Australia, 18–21 April 2017; pp. 15–18.
26. Mardani, M.; Gong, E.; Cheng, J.Y.; Vasanawala, S.S.; Zaharchuk, G.; Xing, L.; Pauly, J.M. Deep Generative Adversarial Neural Networks for Compressive Sensing MRI. *IEEE Trans. Med. Imaging* **2019**, *38*, 167–179. [CrossRef] [PubMed]
27. Jin, K.H.; McCann, M.T.; Froustey, E.; Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **2017**, *26*, 4509–4522. [CrossRef] [PubMed]
28. Kim, C.; Park, D.; Lee, H.-N. Convolutional neural networks for the reconstruction of spectra in compressive sensing spectrometers. In *Optical Data Science II*; International Society for Optics and Photonics: Bellingham, WA, USA, 2019; Volume 10937, p. 109370L.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Macleod, H.A. *Thin-Film Optical Filters*; CRC Press: Boca Raton, FL, USA, 2010.
31. Barry, J.R.; Kahn, J.M. Link design for nondirected wireless infrared communications. *Appl. Opt.* **1995**, *34*, 3764–3776. [CrossRef] [PubMed]
32. Kokaly, R.F.; Clark, R.N.; Swayze, G.A.; Livo, K.E.; Hoefen, T.M.; Pearson, N.C.; Wise, R.A.; Benzel, W.M.; Lowers, H.A.; Driscoll, R.L. *USGS Spectral Library Version 7 Data: US Geological Survey Data Release*; United States Geological Survey (USGS): Reston, VA, USA, 2017.
33. University of Eastern Finland. Spectral Color Research Group. Available online: <http://www.uef.fi/web/spectral/-spectral-database> (accessed on 2 August 2019).
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
35. Chen, G.; Needell, D. Compressed sensing and dictionary learning. *Finite Fram. Theory* **2016**, *73*, 201.
36. Engan, K.; Aase, S.O.; Husoy, J.H. Method of optimal directions for frame design. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258), Phoenix, AZ, USA, 15–19 March 1999; Volume 5, pp. 2443–2446.
37. Aharon, M.; Elad, M.; Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal. Process.* **2006**, *54*, 4311–4322. [CrossRef]
38. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G. Online dictionary learning for sparse coding. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–17 June 2009; pp. 689–696.
39. Koh, K.; Kim, S.-J.; Boyd, S. An interior-point method for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res.* **2007**, *8*, 1519–1555.
40. Martino, L.; Elvira, V. Compressed Monte Carlo for distributed Bayesian inference. *arXiv* **2018**, arXiv:1811.0505.
41. Liu, S.; Wu, H.; Huang, Y.; Yang, Y.; Jia, J. Accelerated Structure-Aware Sparse Bayesian Learning for 3D Electrical Impedance Tomography. *IEEE Trans. Ind. Inform.* **2019**. [CrossRef]
42. Tsiliogianni, E.; Deligiannis, N. Deep coupled-representation learning for sparse linear inverse problems with side information. *IEEE Signal. Process. Lett.* **2019**, *26*, 1768–1772. [CrossRef]

43. Diamond, S.; Sitzmann, V.; Heide, F.; Wetzstein, G. Unrolled optimization with deep priors. *arXiv* **2017**, arXiv:1705.08041.
44. Gilton, D.; Ongie, G.; Willett, R. Neumann Networks for Linear Inverse Problems in Imaging. *IEEE Trans. Comput. Imaging* **2019**. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Two-Wired Active Spring-Loaded Dry Electrodes for EEG Measurements

Seungchan Lee ¹, Younghak Shin ², Anil Kumar ¹, Kiseon Kim ¹ and Heung-No Lee ^{1,*}

¹ School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea; futuremax7@gmail.com (S.L.); anilkdee@gmail.com (A.K.); kskim@gist.ac.kr (K.K.)

² LG CNS AI&BigData Research Center, Seoul 07795, Korea; shinyh0919@gmail.com

* Correspondence: heungno@gist.ac.kr; Tel.: +82-62-715-2237

Received: 4 October 2019; Accepted: 17 October 2019; Published: 21 October 2019



Abstract: Dry contact electrode-based EEG acquisition is one of the easiest ways to obtain neural information from the human brain, providing many advantages such as rapid installation, and enhanced wearability. However, high contact impedance due to insufficient electrical coupling at the electrode-scalp interface still remains a critical issue. In this paper, a two-wired active dry electrode system is proposed by combining finger-shaped spring-loaded probes and active buffer circuits. The shrinkable probes and bootstrap topology-based buffer circuitry provide reliable electrical coupling with an uneven and hairy scalp and effective input impedance conversion along with low input capacitance. Through analysis of the equivalent circuit model, the proposed electrode was carefully designed by employing off-the-shelf discrete components and a low-noise zero-drift amplifier. Several electrical evaluations such as noise spectral density measurements and input capacitance estimation were performed together with simple experiments for alpha rhythm detection. The experimental results showed that the proposed electrode is capable of clear detection for the alpha rhythm activation, with excellent electrical characteristics such as low-noise of 1.131 μV_{RMS} and 32.3% reduction of input capacitance.

Keywords: EEG measurements; active electrodes; spring-loaded dry electrodes; two-wired electrodes; bootstrapping topology

1. Introduction

During the last few decades, dry contact electrode-based electroencephalogram (EEG) acquisition [1] is one of the easiest ways to obtain neural information from the human brain in real time. This type of electrode is rapidly replacing conventional wet electrodes, which have been used in a variety of applications such as patient monitoring of neurological disorders [2], brain-computer interfaces [3], and biofeedback [4]. Nowadays, dry electrodes are integrated into portable commercial devices with wearable technologies to provide personal services such as healthcare and home diagnostics to improve the quality of life. These electrodes are designed to eliminate the need for electrolytic gels, which makes the installation process simple with a short setup time and also prevents an increase in impedance due to drying of gels. However, the absence of conductive gels means that controlling the contact impedance at the electrode-scalp interface is more difficult than using the conventional wet electrodes. Therefore, the impedance characteristics and the physical contact capability of the electrode device have become crucial design considerations for practical electrolyte-free EEG measurements.

Considering these design challenges, many researchers have endeavored to design better dry electrodes with various innovative ideas. Generally, these electrodes can be classified into three different categories: microelectromechanical systems (MEMS)-based electrodes, capacitive electrodes,

and finger-shaped electrodes based on the probe shape and electrical coupling topology at the electrode–scalp interface.

In the MEMS-based dry electrodes [5,6], an array of microneedles are employed to penetrate the 10–40 μm thickness outer skin layer of the scalp. Spiky microneedles, which have lengths of 100–210 μm [7], around 150 μm [8], and 300 μm [9], are typically fabricated on a silicon wafer using special etching processes. In addition to silicon-based materials, a brush-type carbon nanotube-based [10], chitosan/Au-TiO₂ nanotube-based [11], and polydimethylsiloxane (PDMS) substrate-based MEMS electrodes [12] have been developed for various kinds of electrophysiological sensing. Although the tip of microneedles can pass directly through into the inner skin layer to create a direct DC-coupled interface with the scalp surface, their complicated and costly fabrication process and infection risks still remain as practical constraints. In addition, EEG measurements on a hairy scalp are still limited because of the fragile and microscopic needles, which do not penetrate the hair layer effectively.

Capacitive electrodes are generally designed by building AC-coupled non-contact interfaces between the scalp surface and electrodes, utilizing insulation materials such as a hair layer, cotton fabric or printed circuit board (PCB) [13]. This AC-coupled interface can function as a capacitor on the electrode frontend, due to which the acquired biopotentials pass across the electrically insulated layer. With regard to this, Sullivan et al. [14] and Chi et al. [15,16] have proposed PCB plate-based capacitive electrodes equipped with discrete off-the-shelf components or a customized application-specific integrated circuit (ASIC). Capacitive electrodes based on soft insulating materials such as polymer foam [17], PDMS [18] and carbon nanotube [19] have also been introduced. However, there are still many design issues related to measurement distortion such as gain attenuation and phase drift due to the AC-coupled interface [20].

Finger-shaped dry electrodes have also been developed for direct-contact biopotential measurements. In these electrodes, the shape of the probe part has been designed to penetrate the hair layer; therefore, DC-coupled interfaces can be easily made by touching the probes to the scalp surface. From this idea, a shrinkable spring-loaded probe-based passive dry electrode [21], a brush-type flexible dry electrode [22,23], a pin-shaped conductive polymer-based dry electrode [24,25] and a 3D-printed dry-fingered electrode [26] have also been proposed. However, high and unstable contact impedance due to the electrolyte-free interface remains a major challenge in this type of dry electrode.

One possible approach to solve this issue is that the electrode device itself supplies conductive liquid to lower the contact impedance. This method has been presented in the literature [27], but the semi-dry approach still has some of the same problems as the wet types. Another approach is to embed supplementary active circuitry in the electrode device to electronically maximize the input impedance characteristics of the dry electrodes. Following this approach, active electrodes [28] with various circuit topologies designed using off-the-shelf discrete components [29,30] and ASICs [15,31,32] have been proposed. This overall research trend suggests that dry electrodes require that the electrode device is able to make reliable contact with the scalp surface and have high input impedance characteristics.

To meet these requirements, this study proposes a 2-wired active spring-loaded dry electrode to simultaneously achieve high-precision and electrolyte-free EEG monitoring. The proposed electrode is designed with a combination of spring-loaded probes and an active buffer circuit. The finger-shaped probes are able to penetrate the hairs on the scalp without prior preparation, and their shrinkable spring-loaded structures provide mechanical flexibility to each probe for adjustable contact intensity along the curvature of the uneven scalp surface. These structural advantages effectively improve the contact efficiencies of the electrodes with the scalp surface. The zero-drift amplifier-based active buffer circuit provides low-noise impedance conversion to stabilize the intractable impedance characteristics of the dry electrodes caused by the absence of the conductive paste. In the design of the active circuit, the 2-wired bootstrap topology reduces the number of wire connections and provides further enhancement of the input impedance by reducing the input capacitance.

To achieve low-noise and attenuation-free EEG measurements, an equivalent circuit model and amplifier requirements for active circuits were theoretically analyzed in the design process. Evaluations of the electrical characteristics such as spectral noise power density and input capacitance were also performed along with a simple alpha rhythm detection test to verify the EEG feature detection capability.

The contribution of this study is to present an optimized design for an active dry electrode for EEG measurements by combining the electronically maximized impedance characteristic and the physically maximized contact capability of the electrode device.

The remainder of the paper is organized as follows: Section 2 provides detailed descriptions of the design and implementation methods along with an electrical analysis of the equivalent circuit model. The evaluation of the electrical characteristics as well as the experimental methodology for alpha rhythm detection is presented in Section 3. Section 4 summarizes several results, including the evaluation of the electrical characteristics and alpha rhythm detection capability. Finally, a brief discussion of this study and a summary of the proposed electrode development are given in Sections 5 and 6.

2. Design and Implementation

2.1. Two-Wired Active Electrode Design

Active electrodes require an active power supply. At least three wired connections are needed, instead of a single wire, for both the power supply and signal transmission. Compared to conventional passive electrodes that do not require a power supply, the additional wires make it difficult to handle rigid wires and increase the design complexity of the biopotential acquisition system. To reduce the number of wires for the active electrodes, a bootstrap technique [33] was employed for the proposed active dry electrodes. This technique reduces the number of electrode wires by replacing the conventional voltage-based power supply with a current source-based power supply. The power supply rails and signal transmission lines can be shared over a single wire, resulting in an active electrode design that requires only two wire connections.

Figure 1 shows the simplified schematic of the bootstrap technique-based active electrode system using an operational amplifier buffer. The half-power supply bootstrap scheme is implemented by connecting the amplifier's positive power supply rail with its signal output node to a current source. At this point, the current source I_s feeds current to the positive power rail of the amplifier, while the signal output node of the amplifier consumes the surplus current. The signal output voltage is therefore determined as follows:

$$V_o = \frac{(I_s - I_{q+})R_o + A_{ol}V_i}{A_{ol} + 1} \approx V_i \quad (1)$$

where A_{ol} , R_o , and I_{q+} are the amplifier's open-loop gain, output impedance, and quiescent current on the positive power supply rail, respectively.

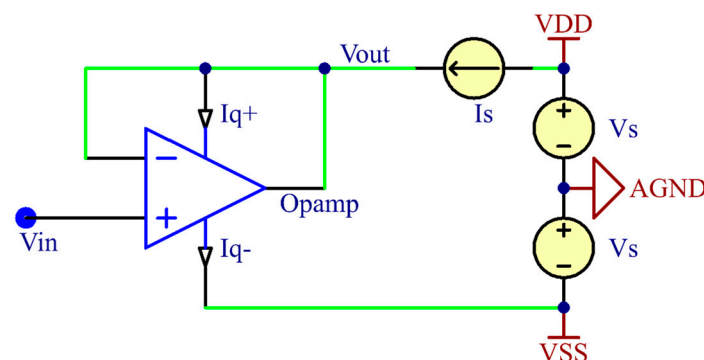


Figure 1. Simplified schematic of bipolar two-wired active electrode with bootstrapping topology.

Generally, the open-loop gain of an amplifier is very large, so the current biasing effects on the output node are neglected. Therefore, the output node voltage will be followed to the input node voltage, and the bootstrapped wire connected with output node can then be used as a signal output link for the active electrode system. However, this circuit design lowers the voltage delivered to the amplifier's positive power supply rail unintentionally, making it difficult to meet the minimum operating voltage for normal amplifier operation in some special cases. To avoid this cases, the operating voltage range of the amplifier needs to be checked. This requirement is discussed further in Section 2.3.2.

The unity-gain buffer configuration allows transformation from the low impedance of the biopotential source to the possible highest impedance [34]. Because the input impedance of the buffer circuit is determined as the differential input impedance multiplied by the open-loop gain, this configuration enables maximizing the electrode impedance. The extremely high input impedance of the dry electrode enables virtually perfected isolation between the source and load, and thus eliminates the loading effects. This property helps to provide a robust signal, which is hardly affected by motion artifacts and power line interferences.

2.2. Electrical Model Analysis and Design Considerations

To investigate the electrical characteristics such as source-to-output gain and input-referred noise of the active circuits, we analyzed the electrical coupling model of the skin–electrode interface for the proposed active circuit. A general electrical model of the active electrode circuit was analytically studied by Chi [13]. Figure 2 shows an equivalent electrical model of the proposed active dry electrode reinterpreted from the general active electrode model. In this circuit model, V_s and V_o denote the biopotential source generated from the human brain and output node of the active circuit, respectively. R_s and C_s represent the resistive and capacitive properties of the scalp–electrode interface established by dry contact of the spring-loaded probes, respectively. R_a and C_a indicate the input resistance and capacitance of the amplifier, respectively. C_p denotes the parasitic capacitance [35] originating from the voltage difference between the signal input and output through active shielding. A_v is the gain of the circuit and is set to unity because the proposed active circuit is designed to operate under a buffer configuration. In order to easily calculate the gain and input-referred noise of the circuit model, the resistances and capacitances have been substituted in parallel at the interface layer ($R_s // C_s$) and input node of the amplifier ($R_a // C_a$) for impedance Z_s and Z_a , respectively. Using nodal analysis, the formulation for source-to-output gain of the equivalent circuit model can be derived as:

$$G(j\omega) = \frac{V_o}{V_s} = \frac{Z_a}{Z_s + Z_a + (1 - A_v)j\omega C_p Z_s Z_a} = \frac{R_a(j\omega R_s C_s + 1)}{j\omega R_s R_a [C_a + C_s + (1 - A_v)C_p] + R_s + R_a} \quad (2)$$

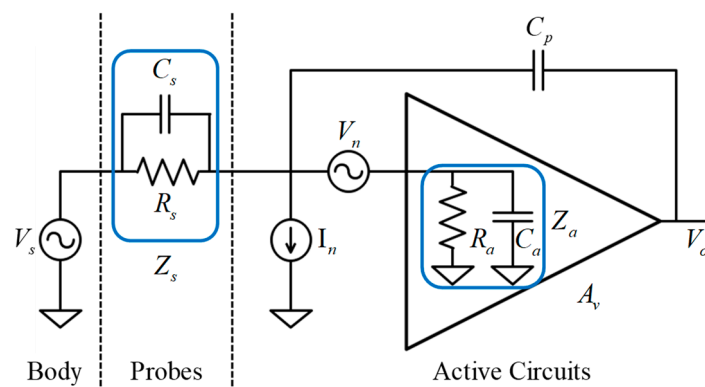


Figure 2. Equivalent circuit model of the proposed active spring-loaded electrodes.

With a low-frequency biopotential source, the contributions of the resistive components are relatively high because of the reduction of the w factor. In the extreme DC case, where w is equal to zero, this gain formula simply changes to $R_a/(R_s + R_a)$. As the value of R_a increases, R_s becomes negligible, which means that the input impedance specification of the amplifier directly affects the gain attenuation of the low-frequency biopotential source.

Conversely, with a high-frequency biopotential source, the contribution of the capacitive components increases. Hence, C_s needs to be maximized, while C_a and C_p need to be minimized in order to avoid gain attenuation of the high-frequency biopotential source. C_p can be minimized by suppressing the leakage current between the input and output nodes. This can be achieved by shielding the input node with the output node of the same potential as the input node. C_a is the amplifier's internal parasitic capacitance that originates from between the input node and both of the power supply rails [36]. Thus, this parasitic capacitance can be considered as a combination of the capacitance built up between the input node and the positive rail (C_{a+}) and between the input node and negative rail (C_{a-}). Applying the bootstrap topology to the proposed active circuit, the voltage difference between the signal output node, which has the same potential as the signal input node, and the positive voltage supply rail of the amplifier can be minimized. Therefore, C_{a+} can be effectively eliminated, and the total capacitance of C_a can also be minimized. C_s is involved in the electrode contact efficiency with the scalp surface. When using non-flexible rigid probes, it is difficult to achieve tight contact with the scalp, resulting in an air gap between the probes and scalp surface. This air gap is equivalent to another extra capacitor, which is connected with C_s in series. Consequently, the total capacitance of C_s will be reduced because of the series connection of two individual capacitors. The flexible spring-loaded probes, on the other hand, can easily adjust their contact intensities in accordance with the curvature of the scalp surface, thus preventing to the building of air gaps. Therefore, the maximization of C_s can be achieved by employing spring-loaded probes.

To quantitatively analyze the noise performance of the active circuit, the noise voltage with respect to the biopotential source input can be expressed as:

$$N_{in} = \left(\frac{Z_s + Z_a + jwC_pZ_sZ_a}{Z_a} \right) V_n + Z_s I_n \quad (3)$$

and the power density, which is equal to the root-mean-squared (RMS) power of the input-referred noise voltage, can also be derived as:

$$N_{in,rms}^2 = \left(\frac{|Z_s + Z_a + jwC_pZ_sZ_a|^2}{|Z_a|^2} \right) V_{n,rms}^2 + |Z_s|^2 I_{n,rms}^2 \quad (4)$$

where $V_{n,rms}^2$ and $I_{n,rms}^2$ denote the RMS-squared power of the voltage and current noise sources V_n and I_n , respectively. These noise sources are derived from the noise model of the amplifier [37], and these parameters depend on the electrical characteristics specified in the amplifier datasheet. Therefore, amplifier selection is a key optimization factor for low-noise biopotential acquisition, and it will be discussed in Section 2.3.2.

For low input-referred noise performance, it is obvious that the operand terms multiplied with the voltage and current noise sources need to be minimized. To lower the voltage noise V_n , Z_a firstly needs to be maximized. The bootstrapping topology provides low input capacitance characteristics by reducing the parasitic capacitance of the amplifier, resulting in high input impedance of the amplifiers. C_p should also be minimized for further reduction of the voltage noise term, which can be achieved by preventing leakage current with robust shielding of the input node. The current noise is typically dominated by the scalp–electrode coupling impedance Z_s , which is inversely proportional to the electrode contact efficiency. To lower the current noise I_n , high contact efficiency is required, meaning that low coupling impedance with low resistance and high capacitance must be achieved. These requirements can be achieved by equipping multiple spring-loaded probes in the design of

the proposed electrode. Installation of the twelve parallelly connected probes lowers the resistive impedance, which in turn prevents poor electrical coupling caused by loose installation of the electrode unit. In addition, the probe's shrinkable structure fills the air gaps caused by microcontact failures at the scalp–electrode interface, thereby continuously keeping high capacitance characteristics.

2.3. Design of Active Dry Electrodes

2.3.1. Spring-Loaded Probes

The EEG signals are acquired using the spring-loaded probes (SK100 Series [19], Leeno Industrial Inc., Pusan, Korea) by contact with the scalp surface. In each electrode, a total of 12 probes are soldered to the active circuit PCB, inside the electrode. Each probe consists of four components: plunger, barrel, spring, and probe receptacles. The plunger has a cylindrical shape, which is coated with beryllium copper and gold plated over nickel. The plating materials are biocompatible, and they do not induce any allergic reactions. The tip of the plunger, which is in contact with the scalp, has a round shape to minimize stabbing pain. The plunger is combined with a barrel and spring to make a spring-loaded structure. With the aid of the embedded springs, the probe is shrinkable up to a maximum of 1.5 mm along with the barrel. The initial pressure of the spring is only 10 g in the preloaded state of the probe. When the probe shrinks to its minimum length, the spring delivers up to 54 g of pressure to the scalp. Therefore, this linearly increasing force ensures that appropriate contact pressure continues to be applied along the uneven scalp surface. The probes also relieve pain by absorbing the excess pressure in the vertical direction. In the electrical specifications, the resistance of each probe is given as less than 50 m Ω , which is sufficiently low for conducting biopotential signals.

2.3.2. Amplifier Specifications

With reference to the electrical model analyses for the proposed active circuits in Section 2.2, it can be observed that the input-referred noise is primarily affected by the electrical specifications of the amplifier. In addition to this, the circuit design with bootstrap topology lowers the voltage on the positive power supply rail in proportion to the input biopotential voltage, which may not meet the minimum voltage requirement for normal amplifier operation. The other amplifier specifications, including offset voltage, input bias current, and quiescent current, should also be checked for the DC-coupled circuit design and longer operation times.

To fulfill these particular requirements, an OPA378 operational amplifier (Texas Instruments, Dallas, TX, USA) [38] was employed for the proposed electrodes. This amplifier provides outstanding characteristics such as low-noise, minimal input offset, a wide acceptable range of power supply voltages, and low power consumption optimized for battery-powered medical instruments. These key parameters are summarized in Table 1.

Because of the microscopic amplitudes of the EEGs, the noise characteristic is the most important parameter in the design of a biopotential sensor, which is indicated as the noise voltage and its spectral densities in the datasheet. According to the IEC standard [28], input-referred noise below 6 μV_{PP} is acceptable for EEG acquisition systems, and the OPA378 fulfills this condition.

Table 1. Electrical characteristics of the OPA378 operational amplifier.

Electrical Parameters	Characteristics
Voltage noise	0.4 μV_{PP} at 0.1–10 Hz
Noise power spectral density	20 nV/ $\sqrt{\text{Hz}}$ at 1 kHz
Offset voltage and offset drift	20 μV and 0.1 $\mu\text{V}/^\circ\text{C}$
Input capacitance	5 pF with common mode
Input bias current	± 150 pA, max. 550 pA
Power supply voltage range	2.2–5.5 V (rail-to-rail)
Quiescent current	125 μA , max. 150 μV

For low-noise EEG measurements in the frequency bands near DC, the offset voltage and its drift need to be checked because they represent measurement precision at the DC region. In low-frequency bands close to DC, $1/f$ noise, called flicker noise [39], is more dominant than other type noises. This type of noise is amplified when approaching the DC region, for which the noise power spectral density is inversely proportional to the square root of the frequency, making it a major noise contributor to the low-frequency band near DC. When a large DC offset is coupled directly with the input of the EEG acquisition system, it can saturate the high-gain preamplifiers and diminish their dynamic range. To mitigate the DC offset, operational amplifiers equipped with advanced circuit design techniques such as auto-calibration and chopping have been introduced and are known as zero-drift amplifiers [28,40]. Utilizing the auto-calibration technique, a signal pathway of the OPA378 continuously corrects the incoming offset voltage every 3 μs with a 350 kHz sample-and-hold circuit. Therefore, this auto-calibration technique maintains a noise voltage density of 20 nV down to 1 Hz and achieves a noise voltage of 0.4 μV_{PP} in the bandwidth of 0.1–10 Hz, thereby extending the acceptable low-frequency range of measurements without an AC-coupled high-pass filter.

As mentioned in Section 2.1, the bootstrap topology lowers the input capacitance of the amplifier by connecting its positive power supply rail to the signal output node, while also lowering the range of voltages supplied to the amplifier. Normal operation cannot be guaranteed, when the supply voltage range does not meet the minimum voltage requirement of 2.2 V. Based on the proposed circuit design, as the voltage of -2.5 V is already supplied to the negative power rail of the amplifier, the common-mode voltage of the input biopotential should be kept at least -0.3 V to meet the minimum voltage requirement. This condition is practically unlikely because of the stable offset characteristics of the amplifier. Nonetheless, if the common-mode voltage of the input node drops below -0.3 , the amplifier will be turned off, and thereby the output node of the circuit can be left in a floating state.

Moreover, an on-chip electromagnetic interference (EMI) filter with 25 MHz cutoff frequency provides outstanding EMI suppression. This feature prevents offset shifts in the amplifier output caused by EMI and allows more precise measurements. The low current consumption of up to 150 μA easily enables multichannel and battery-powered instrumentation.

2.3.3. Circuit Design and Implementation

The schematic of the proposed active dry electrode and its prototype images are shown in Figure 3a,b, respectively. The proposed system comprises two individual parts—the electrode unit and auxiliary board.

The electrode unit is cylindrical in shape with a diameter of 11 mm and a height of 17 mm. The electrode is composed of the 12 spring-loaded probes, OPA378 amplifier, and CMOD6001 low-leakage diode (Central Semiconductor, Hauppauge, NY, USA), and these are installed in the electrode PCB embedded in the 3D-printed electrode housing. All probes are electrically connected to each other, and the measured biopotentials are delivered to the input node of the amplifier. The buffered biopotentials are then finally transferred to the auxiliary board through the bootstrapped wire, which is connected to the current-sourcing device. Concurrently, this current-sourcing device in the auxiliary board supplies a bias current for the amplifier operation through the same bootstrapped wire. The diode is inserted between the amplifier output and positive rail to keep the output voltage swing lower than that of the positive rail by the inherent forward voltage drop of the diode [41]. Even though the amplifier supports rail-to-rail output that allows maximizing the output swing over the entire range of the supply voltage, this diode is necessary to keep an extra margin for low-distortion voltage output and low power consumption.

The auxiliary board is designed to provide a constant current source and bipolar voltage power using linear regulators, current source devices, and numerous decoupling capacitors. To supply low-noise voltage for the $+2.5$ V and -2.5 V rails, ADM7154 and ADP7183 linear regulators (Analog Devices, Norwood, MA, USA) were used. These regulators provide extremely low-noise performance of 1.6 μV_{RMS} and 4 μV_{RMS} along with high power supply rejection ratios, which are optimized for

noise-sensitive applications. A REF200 (Texas Instruments, Dallas, TX, USA) [42], which is embedded with two 100 μA current sources, was used as the current-sourcing device. By connecting the regulated 2.5 V rail to the current sources, the device is capable of simultaneously powering two channels of the proposed electrodes. Although the current-sourcing capability is limited to 100 μA per channel, the current requirement for the positive rail of the amplifier is only 75 μA , which is half of the maximum current consumption of 150 μA , thus ensuring sufficient current supply. All electrical components are small and surface mounted type, thereby making it easy to design portable size instruments.

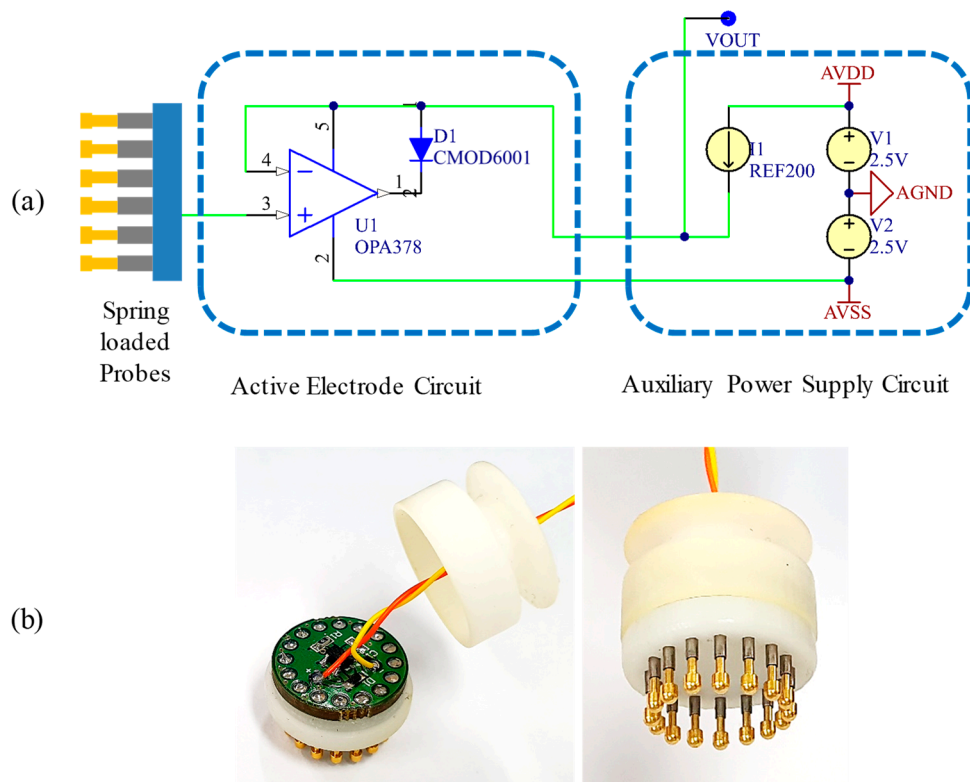


Figure 3. Actual implemented schematic (a) and images (b) of the proposed active dry electrode. The proposed electrode system comprises the electrode unit itself and an auxiliary circuit board for the voltage and current power supplies. In the electrical schematic, decoupling capacitors for stabilized voltage supplies are omitted for simplicity.

3. Evaluation and Experiment

3.1. Noise Characteristics

In the design of electronic circuit-based sensors, the noise floor of the sensing signals is a key parameter for determining the integrity of the acquired data. To evaluate the noise characteristics of the proposed active electrode circuit, noise power spectral densities were analyzed using an FFT-based spectrum analyzer (Keysight 35670A, Santa Rosa, CA, USA), which can quickly capture the spectral information of analog signals utilizing Fourier analysis and digital signal processing techniques. With this instrument, the total noise output of a circuit can be estimated by shorting the circuit's input node to ground potential and measuring the power spectral densities at the output node of the circuit. To compare the noise measurements of the proposed 2-wired bootstrap buffered circuit, a bipolar-powered 3-wired conventional buffered circuit was implemented as a target of comparison. For the two types of active circuits, 1600-point power spectral densities were measured over the 0.1–200 Hz bandwidth. These measurements were repeated 50 times and averaged for a smoother representation. The measured noise spectra were transmitted to a laptop using a USB-type GPIB

interface and instrument control software (Keysight VEE Pro 9.2, Santa Rosa, CA, USA). To reject noise interference, this evaluation was performed within an aluminum enclosure.

In the analysis stage, Pearson correlation coefficients and Wilcoxon signed-rank test was used to measure orientational and statistical similarities between the two pairs of noise spectral densities. To compare actual noise voltages in the EEG bandwidth precisely, RMS voltages were also calculated from the measured noise spectral densities by taking the squared values of the given voltage spectral density, integrating within the specified frequency range, and computing the square root.

3.2. Input Capacitance

In the electrode design for EEG measurements, high input impedance is an essential characteristic for further signal conditioning processes. High input impedance also implies low input capacitance at higher frequencies. To investigate the impedance characteristics of the proposed electrode circuit, the input capacitances for the proposed circuit (2-wired bootstrap buffered circuit) and its counterpart (3-wired conventional buffered circuit) were analyzed. Since the input capacitance of the operational amplifier is typically lower than a few picofarads, the direct measurements for observing input capacitance using a multimeter are not practical because of its poor error tolerance. In order to measure the input capacitance of the operational amplifier-based circuit, a large resistor was inserted in series with the input node of the amplifier. This configuration set up a first-order RC lowpass filter in combination with the internal capacitance of the amplifier. Through the frequency response analysis for the circuits, the input capacitances can be inversely estimated by evaluating the -3 dB cutoff frequencies. Detailed information on this methodology is described in [43].

The same spectrum analyzer was used to investigate the input-to-output frequency responses for the test circuits. After inserting a $2\text{ M}\Omega$ resistor as a large source resistor R_s , a 100 mV_{pp} sinusoidal sweep was applied to the input node of the target circuit in accordance with 800 log-scaled bins arranged over the 1–51.2 kHz bandwidth. The swept source was routed to the input probe of the spectrum analyzer using a signal splitter, while the output probe was connected to the output node of the target circuit, unlike the test setup in [43]. This is because a unity-gain buffer configuration allows the input signal of the amplifier to be identically measured at the output node of the circuit without the need for a high-impedance FET probe. From this setup, dB-scaled Bode plots can be obtained from the frequency analyses, and we can estimate the input capacitance of the test circuit using the following equation: $C_{in} = 1/(2\pi R_s f_{-3dB})$. All tests were carried out using customized test PCBs that were carefully designed with active shielding to avoid other parasitic capacitances.

3.3. Alpha Rhythm Detection Experiment

Alpha rhythm, the most prominent feature of an EEG, can be easily utilized as a benchmark tool for testing the detection capabilities of real EEG features. When users close their eyes, the spectral power of the alpha rhythm band (8–15 Hz) is amplified compared to other spectral ranges, and vice versa when the users open their eyes. By comparing the spectral activation for the alpha rhythm when the eyes are closed or open, we can evaluate the practical applicability of the proposed electrode for real EEG monitoring.

For this purpose, ten trials were performed for a subject. A single trial consisted of maintaining the eye-open state for 12.5 ± 2.5 s and the eye-closed state for 10 s. For every transition of instruction, a beep sound was used to inform the subject of the command changes. Alternative electrodes such as a 3-wired active buffered electrode and a passive dry electrode, as well as the proposed electrode, were employed for the comparison of EEG measurements. All electrode implementations were equipped with the same spring-loaded probe for dry contact with the scalp surface. These three electrodes were installed as close as possible to the Fz position according to the international 10–20 system. Disposable wet electrodes were also attached to the skin behind the left and right earlobes as a reference and a bias electrode, respectively. Experiments were conducted using MATLAB 2014a (Mathworks, Natick, MA,

USA) and the Cogent 2000 toolbox, and EEG measurements were recorded using the ADS1299-based EEG acquisition system which was built in a previous study [44].

Offline analyses for the EEG measurements were also performed using MATLAB 2014a. The raw EEG dataset was filtered with a 4th order zero-phase 0.5–40 Hz bandpass Butterworth filter. From the filtered EEG dataset, the epochs for 5 s corresponding to each condition were extracted based on the recorded event triggers. The spectral power values were also calculated to precisely compare the spectral activation for the alpha rhythm. The 10-s EEG measurements were also visualized before and after the fifth eye-close instruction for time-series waveform comparisons. In addition, the similarity of the bandpass-filtered waveforms was evaluated in terms of Pearson correlation coefficients.

4. Results

4.1. Noise Power Spectral Density

The comparison of the noise power spectral densities for the proposed active electrode circuit and its conventional counterpart are depicted in Figure 4a. Compared to the noise power spectral densities of the conventional 3-wired buffer circuit, the proposed 2-wired bootstrap topology shows a similar trend along with a correlation coefficient of 0.953 within the EEG bandwidth of 0.5–50 Hz. However, in the Wilcoxon signed-rank test, a nonparametric statistical method for testing a hypothesis of paired data, the two paired noise spectral densities do not show statistical similarities with a low significant level ($p < 0.0001$) at the same EEG bandwidth. In a complementary analysis for a sum of the difference between the paired noise spectral densities, we found that the proposed design produces more noise by $2.0433 \text{ nV } \sqrt{\text{Hz}}$ on average than the 3-wired counterpart. Consequently, this extra noise leads to a small difference between the estimated RMS noise voltages (i.e., $1.131 \text{ } \mu\text{V}_{\text{RMS}}$ with the proposed 2-wired topology vs. $1.017 \text{ } \mu\text{V}_{\text{RMS}}$ with the 3-wired counterpart). The slightly increased RMS noise in the proposed topology is due to an increase in noise power at lower frequency bands below 1 Hz. The reason for this is the positive rail voltage of the proposed topology, which has continuously changed in accordance with the voltage corresponding to the acquired input signal, instead of being supplied from a low-noise constant voltage source. This unfixed supply voltage, combined with thermal noise and other interference, seems to result in minor extra noise in the low-frequency region.

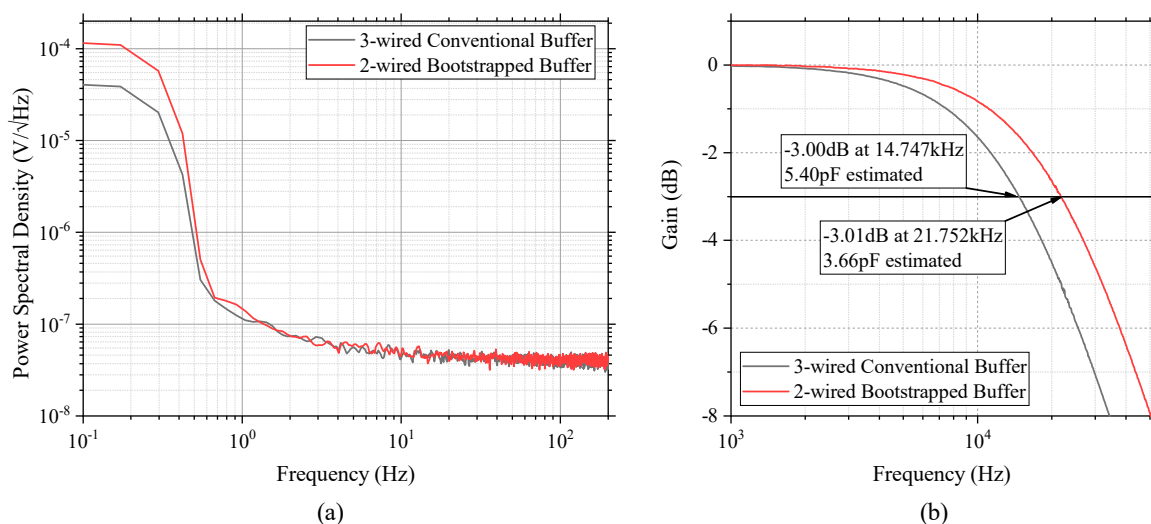


Figure 4. (a) Measurements of the noise power spectral densities and (b) input capacitance estimation results for the proposed active electrode circuit and its alternative implementation (2-wired bootstrapped buffered topology vs. 3-wired conventional buffered topology).

4.2. Input Capacitance Estimation

The spectral analysis results for investigation of the -3 dB cutoff frequencies and the estimated input capacitance from the results are depicted in Figure 4b. For the two types of circuit configuration, the differences in the cutoff frequencies is about 7 kHz, resulting in a 1.74 pF reduction in the input capacitance for the proposed bootstrap configuration compared to the conventional buffer design approach. The impedance of the amplifier is represented as $R/(j\omega RC + 1)$, because it is simplified as a parallel combination of resistance and capacitance. Therefore, an approximately 32.2% reduction in the input capacitance leads to roughly 147.5% impedance boosting within the EEG bandwidth. This impedance boosting effect makes the measurement more robust against artifacts and EMI interference.

4.3. Experimental Results of Alpha Rhythm Detection

Figure 5 shows the experimental results for the alpha rhythm detection test measured by three types (2-wired active, 3-wired active, and passive) of dry electrodes. The captured time-series waveforms on the left side of figures were extracted from the EEG measurements near the onset time of the fifth task from among the 10 trials. In these waveforms, a red vertical line indicates the start time for the eye-close instruction. Within one second after task onset, slightly large voltage swings were observed in all measurements while the subject's eyelids are closed. After the swings, it was confirmed that clear alpha waves were identified with their distinguished oscillations of measured waveforms. These evoked alpha waves are easily noticeable in the spectral analysis. The figures on the right side show the results of the event-related spectral analysis for each electrode measurements. These spectral comparisons clearly visualize the maximized spectral differences evoked near the 10 Hz, which belong to the alpha rhythm. Specifically, the maximum spectral differences for two different tasks were observed at 12.4 dB at 10.1 Hz for the proposed 2-wired active electrode, 11.28 dB at 10.06 Hz for the 3-wired active electrode, and 13.83 dB at 9.98 Hz for the passive electrode. These spectral analysis results confirmed that EEG feature detection can be fully achieved using the proposed electrode.

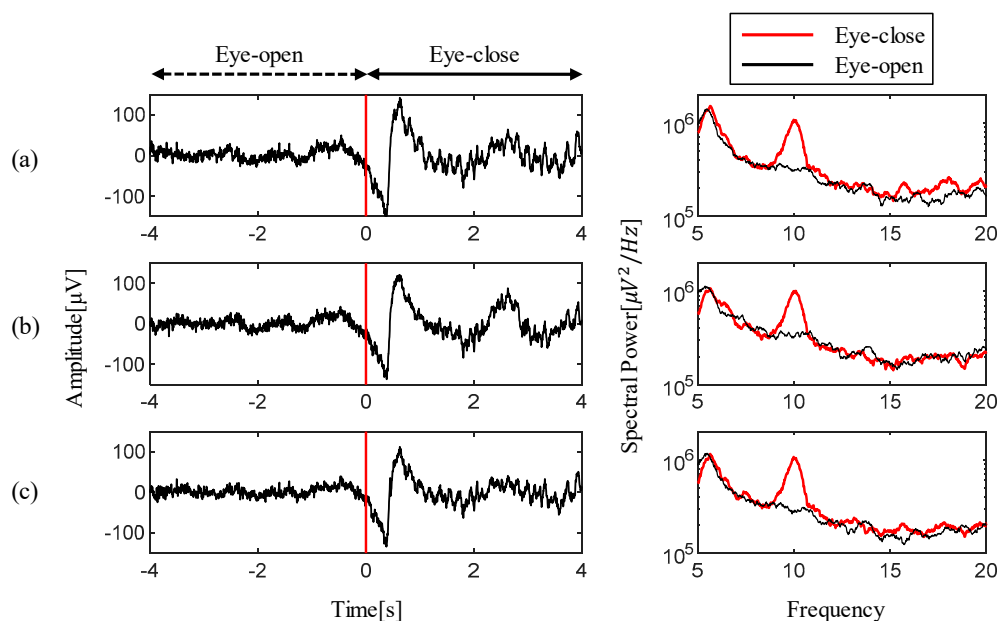


Figure 5. EEG measurements and their spectral comparisons for (a) proposed 2-wired active dry electrode, (b) alternative 3-wired active dry electrode, and (c) passive dry electrode. On the left, the red vertical line on the EEGs indicate the task onset timing for the eye-close instructions. During the eye-close session, activated alpha waves are commonly observed in the time-series and spectral visualization results for all types of electrodes.

The comparison of the correlation coefficients for each paired EEG waveforms is summarized in Table 2. The correlation coefficient between the EEG waveforms using the passive dry electrode and the proposed 2-wired topology is ρ_2 ; the correlation coefficient between the EEG waveforms using the passive dry electrode and the 3-wired counterpart is ρ_3 ; and the correlation coefficient between the proposed 2-wired topology and the 3-wired counterpart is ρ_{23} . An insignificant difference between ρ_2 and ρ_3 indicates that the proposed 2-wired electrode is sufficient to achieve measurements nearly equivalent to the conventional 3-wired design approach. A slight decrease in the value ρ_{23} , compared to ρ_2 and ρ_3 , is supposed to be caused by the difference in common-mode voltages in accordance with a difference in the design topology.

Table 2. Comparison of correlation coefficients for each paired EEG datasets.

2-wired Active vs. Passive (ρ_2)	3-wired Active vs. Passive (ρ_3)	2-wired Active vs. 3-wired Active (ρ_{23})
0.8536	0.8657	0.7854

5. Discussion

Theoretical analysis of the equivalent circuit model for the proposed electrode indicated that the electrical specifications of the amplifier have a significant effect on measurement characteristics such as input-referred noise and gain attenuation. As standard specifications in the datasheet, the offset voltage and the 0.1–10 Hz peak-to-peak noise voltage are involved with not only the precision of the common-mode voltages, but also noise characteristics within the low-frequency bands near DC, associated with $1/f$ noise. Since even EEG waves with very low-frequency bands (0.1–4 Hz), including delta waves and slow oscillations, are often used for sleep studies [45], the examination of these specifications is required to verify the $1/f$ noise characteristics. The OPA378, a zero-drift amplifier with 0.1–10 Hz RMS noise of 0.4 μV and offset voltage of 20 μV , provides excellent low-noise characteristics, but noise boosting is still observed at lower frequencies below 1 Hz in the actual noise measurements. This is because the $1/f$ noise is generated internally from the quantum mechanical random process inherent in all semiconductor devices, including the amplifier to be measured and the measurement instrument itself. This means it is difficult to eliminate $1/f$ noise completely. Nevertheless, the proposed electrode still presented excellent low noise characteristics of 1.131 μV_{RMS} within an EEG bandwidth of 0.5–50 Hz along with noise power spectral densities of 139 nV/ $\sqrt{\text{Hz}}$ at 1 Hz and 49 nV/ $\sqrt{\text{Hz}}$ at 10 Hz. These measurements are comparable with previous studies (7.4 μV_{RMS} within a bandwidth of 1–1000 Hz in the [41], and 200 nV/ $\sqrt{\text{Hz}}$ at 1 Hz in the [15]).

On the other side, the ratio of the noise characteristics versus power consumption also needs to be checked to consider the entire power consumption of the EEG acquisition system. There is a trade-off relationship between power consumption and noise performance [46], which means that increased power consumption of the amplifier results in better low-noise characteristics in general. Amplifiers that require higher power can be used in the active electrodes for better noise performance, but this results in an increase in the overall power requirement of the instrument with numerous channels. For example, the state-of-the-art operational amplifier OPA188 (Texas Instruments, Dallas, TX, USA) exhibits a better noise voltage of 250 nV_{PP} over the 0.1–10 Hz bandwidth, which is a 37.5% lower noise voltage compared to that of the OPA378. However, its current consumption is typically increased 3.6 times more to 450 μA . The proposed electrode is designed to consume up to 150 μA of current per channel, resulting in a total of only 2.4 mA for 16 channels, thus it can be continuously operated for about 40 hours even with a 100 mAh small lithium polymer battery. This low-power operation adequately meets a design requirement for battery-powered mobile instruments.

Compared with previous studies [18,20,29], another difference in the frontend circuit design is the exclusion of a bias current path. In those previous studies, a large value resistor in the T Ω range or parallel connection of two reverse diodes is generally used as the bias current path. This is necessary to prevent voltage saturation at the input node of the amplifier caused by incoming bias currents,

but it also generates a lot of thermal noise due to the high physical resistance. The problem here is that irrespective of the resistance value, degradation of the amplifier's input impedance cannot be avoided. The proposed electrode omits the design of the bias current path, but the built-in protection circuitries embedded in the amplifier can fulfill this role to effectively prevent electrical overstress at the input node and degradation of the high input impedance.

6. Conclusions

In this study, we have proposed a two-wired active spring-loaded dry electrode to conduct electrolyte-free EEG monitoring. By combining spring-loaded probes with the active buffer circuit, the proposed electrode design simultaneously enables electronically maximized input impedance, and physically maximized contact capability. In the design process, the equivalent circuit model for the electrode circuit and its associated electrical parameters such as noise and gain attenuation were analyzed to obtain low-noise and attenuation-free EEG measurements. Based on the analysis, the active circuit was designed based on low-cost discrete components and the low-noise and low-offset zero-drift amplifier. The complete electrode device was implemented by combining the active buffer circuit with spring-loaded probes and a 3D-printed housing. Through several evaluations included the alpha rhythm detection test, the proposed electrodes were found to have a low-noise characteristic of $1.131 \mu\text{V}_{\text{RMS}}$ within the EEG bandwidth of 0.5–50 Hz and the capability to clearly detect an alpha rhythm near 10 Hz. In addition, by applying the bootstrap topology to the proposed electrode design, the proposed electrode only requires a two-wired connection with an approximate 32.2% reduction in the input capacitance. This leads to an impedance boosting of roughly 147.5% within the EEG bandwidth. In our future work, we plan to design a portable instrument for mobile EEG monitoring based on the proposed electrode system.

Author Contributions: Conceptualization, S.L. and K.K.; Data curation, S.L.; Formal analysis, S.L. and Y.S.; Funding acquisition, H.-N.L.; Investigation, S.L. and Y.S.; Methodology, S.L. and K.K.; Project administration, H.-N.L.; Resources, H.-N.L.; Software, S.L.; Supervision, A.K. and H.-N.L.; Visualization, S.L.; Writing – original draft, S.L.; Writing – review & editing, Y.S., A.K. and H.-N.L.

Funding: This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MISP) under Grants [NRF-2018R1A2A1A19018665] and [NRF-20151A2A1A05001826], and by the Brain Research Program through the NRF funded by the Ministry of Science, ICT & Future Planning under Grant [NRF-2016M3C7A1905475].

Acknowledgments: The authors would like to thank all involved personnel and institutions for their valuable technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lopez-Gordo, M.A.; Sanchez-Morillo, D.; Valle, F.P. Dry EEG Electrodes. *Sensors* **2014**, *14*, 12847–12870. [[CrossRef](#)] [[PubMed](#)]
2. Smith, S.J.M. EEG in the diagnosis, classification, and management of patients with epilepsy. *J. Neurol. Neurosurg. Psychiatry* **2005**, *76*, 2–7. [[CrossRef](#)] [[PubMed](#)]
3. Nicolas-Alonso, L.F.; Gomez-Gil, J. Brain Computer Interfaces—A Review. *Sensors* **2012**, *12*, 1211–1279. [[CrossRef](#)] [[PubMed](#)]
4. Millán, J.D.R.; Rupp, R.; Müller-Putz, G.R.; Murray-Smith, R.; Giugliemma, C.; Tangermann, M.; Vidaurre, C.; Cincotti, F.; Kübler, A.; Leeb, R.; et al. Combining Brain–Computer Interfaces and Assistive Technologies: State-of-the-Art and Challenges. *Front Neurosci.* **2010**, *4*. [[CrossRef](#)]
5. Ren, L.; Liu, B.; Zhou, W.; Jiang, L. A Mini Review of Microneedle Array Electrode for Bio-Signal Recording: A Review. *IEEE Sens. J.* **2019**, *1*. [[CrossRef](#)]
6. Yao, S.; Zhu, Y. Nanomaterial-Enabled Dry Electrodes for Electrophysiological Sensing: A Review. *JOM* **2016**, *68*, 1145–1155. [[CrossRef](#)]
7. Griss, P.; Enoksson, P.; Tolvanen-Laakso, H.K.; Merilainen, P.; Ollmar, S.; Stemme, G. Micromachined electrodes for biopotential measurements. *J. Microelectromechanical Syst.* **2001**, *10*, 10–16. [[CrossRef](#)]

8. Dias, N.S.; Carmo, J.P.; da Silva, A.F.; Mendes, P.M.; Correia, J.H. New dry electrodes based on iridium oxide (IrO) for non-invasive biopotential recordings and stimulation. *Sens. Actuators A Phys.* **2010**, *164*, 28–34. [[CrossRef](#)]
9. O'Mahony, C.; Pini, F.; Blake, A.; Webster, C.; O'Brien, J.; McCarthy, K.G. Microneedle-based electrodes with integrated through-silicon via for biopotential recording. *Sens. Actuators A Phys.* **2012**, *186*, 130–136. [[CrossRef](#)]
10. Ruffini, G.; Dunne, S.; Fuentemilla, L.; Grau, C.; Farrés, E.; Marco-Pallarés, J.; Watts, P.C.P.; Silva, S.R.P. First human trials of a dry electrophysiology sensor using a carbon nanotube array interface. *Sens. Actuators A Phys.* **2008**, *144*, 275–279. [[CrossRef](#)]
11. Song, Y.; Li, P.; Li, M.; Li, H.; Li, C.; Sun, D.; Yang, B. Fabrication of chitosan/Au-TiO₂ nanotube-based dry electrodes for electroencephalography recording. *Mater. Sci. Eng. C* **2017**, *79*, 740–747. [[CrossRef](#)] [[PubMed](#)]
12. Zhang, H.; Pei, W.; Chen, Y.; Guo, X.; Wu, X.; Yang, X.; Chen, H. A Motion Interference-Insensitive Flexible Dry Electrode. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1136–1144. [[CrossRef](#)] [[PubMed](#)]
13. Chi, Y.M.; Jung, T.P.; Cauwenberghs, G. Dry-Contact and Noncontact Biopotential Electrodes: Methodological Review. *IEEE Rev. Biomed. Eng.* **2010**, *3*, 106–119. [[CrossRef](#)] [[PubMed](#)]
14. Sullivan, T.J.; Deiss, S.R.; Cauwenberghs, G. A Low-Noise, Non-Contact EEG/ECG Sensor. In Proceedings of the 2007 IEEE Biomedical Circuits and Systems Conference, Montreal, QC, Canada, 27–30 November 2007; pp. 154–157.
15. Chi, Y.M.; Maier, C.; Cauwenberghs, G. Ultra-High Input Impedance, Low Noise Integrated Amplifier for Noncontact Biopotential Sensing. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2011**, *1*, 526–535. [[CrossRef](#)]
16. Chi, Y.M.; Wang, Y.; Wang, Y.; Maier, C.; Jung, T.; Cauwenberghs, G. Dry and Noncontact EEG Sensors for Mobile Brain–Computer Interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2012**, *20*, 228–235. [[CrossRef](#)] [[PubMed](#)]
17. Baek, H.J.; Kim, H.S.; Heo, J.; Lim, Y.G.; Park, K.S. Brain–computer interfaces using capacitive measurement of visual or auditory steady-state responses. *J. Neural Eng.* **2013**, *10*, 024001. [[CrossRef](#)] [[PubMed](#)]
18. Lee, S.M.; Kim, J.H.; Byeon, H.J.; Choi, Y.Y.; Park, K.S.; Lee, S.-H. A capacitive, biocompatible and adhesive electrode for long-term and cap-free monitoring of EEG signals. *J. Neural Eng.* **2013**, *10*, 036006. [[CrossRef](#)]
19. Lee, S.M.; Kim, J.H.; Park, C.; Hwang, J.; Hong, J.S.; Lee, K.H.; Lee, S.H. Self-Adhesive and Capacitive Carbon Nanotube-Based Electrode to Record Electroencephalograph Signals from the Hairy Scalp. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 138–147. [[CrossRef](#)]
20. Baek, H.J.; Lee, H.J.; Lim, Y.G.; Park, K.S. Comparison of pre-amplifier topologies for use in brain-computer interface with capacitively-coupled EEG electrodes. *Biomed. Eng. Lett.* **2013**, *3*, 158–169. [[CrossRef](#)]
21. Liao, L.-D.; Wang, I.-J.; Chen, S.-F.; Chang, J.-Y.; Lin, C.-T. Design, Fabrication and Experimental Validation of a Novel Dry-Contact Sensor for Measuring Electroencephalography Signals without Skin Preparation. *Sensors* **2011**, *11*, 5819–5834. [[CrossRef](#)]
22. Grozea, C.; Voinescu, C.D.; Fazli, S. Bristle-sensors—Low-cost flexible passive dry EEG electrodes for neurofeedback and BCI applications. *J. Neural Eng.* **2011**, *8*, 025008. [[CrossRef](#)] [[PubMed](#)]
23. Gao, K.; Yang, H.; Liao, L.; Jiang, C.; Zhao, N.; Wang, X.; Li, X.; Yang, B.; Liu, J. A Novel Bristle-shaped Semi-dry Electrode with Low Contact Impedance and Ease of Use Features for EEG Signal Measurement. *IEEE Trans. Biomed. Eng.* **2019**, *1*. [[CrossRef](#)] [[PubMed](#)]
24. Chen, Y.-H.; de Beeck, M.O.; Vanderheyden, L.; Carrette, E.; Mihajlović, V.; Vanstreels, K.; Grundlehner, B.; Gadeyne, S.; Boon, P.; Van Hoof, C. Soft, Comfortable Polymer Dry Electrodes for High Quality ECG and EEG Recording. *Sensors* **2014**, *14*, 23758–23780. [[CrossRef](#)] [[PubMed](#)]
25. Fiedler, P.; Mühle, R.; Griebel, S.; Pedrosa, P.; Fonseca, C.; Vaz, F.; Zanow, F.; Haueisen, J. Contact Pressure and Flexibility of Multipin Dry EEG Electrodes. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 750–757. [[CrossRef](#)] [[PubMed](#)]
26. Krachunov, S.; Casson, A.J. 3D Printed Dry EEG Electrodes. *Sensors* **2016**, *16*, 1635. [[CrossRef](#)]
27. Li, G.; Zhang, D.; Wang, S.; Duan, Y.Y. Novel passive ceramic based semi-dry electrodes for recording electroencephalography signals from the hairy scalp. *Sens. Actuators B Chem.* **2016**, *237*, 167–178. [[CrossRef](#)]
28. Xu, J.; Mitra, S.; Hoof, C.V.; Yazicioglu, R.F.; Makinwa, K.A.A. Active Electrodes for Wearable EEG Acquisition: Review and Electronics Design Methodology. *IEEE Rev. Biomed. Eng.* **2017**, *10*, 187–198. [[CrossRef](#)]
29. Huang, Y.J.; Wu, C.Y.; Wong, A.M.K.; Lin, B.S. Novel Active Comb-Shaped Dry Electrode for EEG Measurement in Hairy Site. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 256–263. [[CrossRef](#)]


30. Guerrero, F.N.; Spinelli, E.M. A Two-Wired Ultra-High Input Impedance Active Electrode. *IEEE Trans. Biomed. Circuits Syst.* **2018**, *12*, 437–445. [CrossRef]
31. Guermandi, M.; Cardu, R.; Scarselli, E.F.; Guerrieri, R. Active Electrode IC for EEG and Electrical Impedance Tomography with Continuous Monitoring of Contact Impedance. *IEEE Trans. Biomed. Circuits Syst.* **2015**, *9*, 21–33. [CrossRef]
32. Pourahmad, A.; Dehghani, R. Two-Wired Current Modulator Active Electrode for Ambulatory Biosignal Recording. *IEEE Trans. Biomed. Circuits Syst.* **2019**, *13*, 15–25. [CrossRef] [PubMed]
33. Degen, T.; Torrent, S.; Jackel, H. Low-Noise Two-Wired Buffer Electrodes for Bioelectric Amplifiers. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 1328–1332. [CrossRef] [PubMed]
34. An Applications Guide for Op Amps. Available online: <http://www.ti.com/lit/an/snoa621c/snoa621c.pdf> (accessed on 17 October 2019).
35. Lányi, S. The noise of input stages with low parasitic capacitance. *Meas. Sci. Technol.* **2001**, *12*, 1456. [CrossRef]
36. Lanyi, S.; Pisani, M. A high-input-impedance buffer. *IEEE Trans. Circuits Syst. I Fundam. Theory Appl.* **2002**, *49*, 1209–1211. [CrossRef]
37. Noise Analysis in Operational Amplifier Circuits. Available online: <http://www.ti.com/lit/an/slva043b/slva043b.pdf> (accessed on 17 October 2019).
38. OPA378 Datasheet. Available online: <http://www.ti.com/lit/ds/symlink/opa378.pdf> (accessed on 5 September 2019).
39. Low Noise Signal Conditioning for Sensor-Based Circuits. Available online: <https://www.analog.com/media/en/technical-documentation/technical-articles/MS-2066.pdf> (accessed on 17 October 2019).
40. To Chop or Auto-Zero: That Is the Question. Available online: <https://www.analog.com/media/en/technical-documentation/technical-articles/MS-2062.pdf> (accessed on 17 October 2019).
41. Degen, T.; Jackel, H. A pseudodifferential amplifier for bioelectric events with DC-offset compensation using two-wired amplifying electrodes. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 300–310. [CrossRef]
42. REF200 Datasheet. Available online: <http://www.ti.com/lit/ds/symlink/ref200.pdf> (accessed on 5 September 2019).
43. Measure the Input Capacitance of an Op Amp. Available online: <https://pdfserv.maximintegrated.com/en/an/AN5086.pdf> (accessed on 17 October 2019).
44. Lee, S.; Shin, Y.; Kumar, A.; Kim, M.; Lee, H. Dry Electrode-Based Fully Isolated EEG/fNIRS Hybrid Brain-Monitoring System. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 1055–1068. [CrossRef]
45. Achermann, P.; Dijk, D.-J.; Brunner, D.P.; Borbély, A.A. A model of human sleep homeostasis based on EEG slow-wave activity: Quantitative comparison of data and simulations. *Brain Res. Bull.* **1993**, *31*, 97–113. [CrossRef]
46. Design Trade-Offs for Single-Supply Op Amps. Available online: <https://pdfserv.maximintegrated.com/en/an/AN656.pdf> (accessed on 17 October 2019).



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Optimally sectioned and successively reconstructed histogram sub-equalization based gamma correction for satellite image enhancement

Himanshu Singh¹  · Anil Kumar¹ · L. K. Balyan¹ · H. N. Lee²

Received: 3 May 2018 / Revised: 16 January 2019 / Accepted: 18 February 2019 /
Published online: 1 March 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

This paper presents an overall quality enhancement approach especially for dark or poorly illuminated images with a core objective to re-allocate the processed pixels using recursive histogram sub-division. An information preserved and image content based behavioral reconstruction inspired adaptive stopping criterion based on pixel-wise relative L_2 -norm basis (which itself is intuitively related to optimal PSNR value) is proposed in this paper, so that highly adaptive gamma value-set can be derived out of it for sufficient enhancement. Due to this adaptive behavior of the intensity distribution the gamma value-set when derived from it, is obviously highly adaptive and here individual gamma values are evaluated explicitly raised over reconstructed intensity values, unlike conventional gamma correction methods. This adaptiveness makes the entire methodology highly capable for covering a wide variety of images, due to which robustness of the algorithm also increases. The proposed methodology has been verified on various dark images. The simulation results authenticate the overall enhancement (contrast as well as entropy enhancement along with sharpness enhancement) achieved by the proposed has been found superior to other dark image enhancement techniques.

Keywords Sub-histograms, Gamma Correction · Image quality enhancement · Adaptive thresholding, Peak signal to noise ratio (PSNR)

✉ Himanshu Singh
himanshu.iitj@gmail.com

Anil Kumar
anilkdee@gmail.com

L. K. Balyan
lokendra.balyan@gmail.com

H. N. Lee
heungno@gist.ac.kr

Extended author information available on the last page of the article

1 Introduction

Remotely acquired digital imagery in diverse circumstances and its indispensable contribution for social welfare, demands an efficient quality enhancement as a core part of image pre-processing. In this manner, the required information can be restored and required parametric features can be sufficiently extracted according to the demand [21]. Researchers get highly fascinated by histogram equalization (HE) [5] and its efficiently modified variations due to their simplicity and less computational complexity. Obviously, the global HE cannot preserve local spatial features of the image which limits the amount of quality enhancement in all parts of the image and hence, researchers started looking for distributing histogram into its constituting sub-histograms for local histogram modifications [8, 15, 17, 18]. Fuzzy inspired histogram smoothing followed by local maxima based sub-division has been also proposed as Brightness preserving dynamic fuzzy HE (BPDFHE) [15]. Exposure-based sub-image HE (ESIHE) [17] has been proposed for low exposure images, where image exposure is utilized for sub-division. Afterward, median-mean dependent sub-image-clipped HE (MMSICHE) [18] has been introduced where histogram clipping is based on the median with bisecting each section to obtain four sub-images, so that they can be equalized locally. Later, recursive-ESIHE (R-ESIHE) [19] by iterative usage of ESIHE till exposure reduced to a predefined threshold. Also, its multi-level histogram separation version termed as recursively separated-ESIHE (RS-ESIHE) [19] has been also introduced. Later on, the averaging histogram equalization (AVGHEQ) [11], HE based optimal profile compression (HEOPC) [30] method for color image enhancement followed by HE with maximum intensity coverage (MAXCOVER) [31] have been also proposed. Also, the adaptive gamma correction with weighting distribution (AGCWD) [7] and its efficient variations [16, 20, 24–31] have been also proposed for dark images. Afterward, the intensity and edge-based adaptive unsharp masking filter (IEAUMF) [10] based enhancement have been also proposed by employing the unsharp masking filter for edge augmentation. Sigmoid mapping through cosine transformed Regularized-HE [4] has been also proposed. Recently, getting fascinated by artificial intelligence and deep learning based methods, various methodologies have been also proposed namely, LIME: Low-light image enhancement via illumination map estimation (LIME) [6], Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans (DPE) [3], Learning to See in the Dark (LSD) [2], and Learning a deep single image contrast enhancer from multi-exposure images (LDSICEM) [1]. In the same sequence, although several kinds of enhancement methodologies have been proposed till date for widely diverse characteristics of images from various domains, (contextual literature survey is explicitly presented in [21, 24, 25]), still most of them are lagging when it comes to the matter of enhancement of different domain images through a single approach. In this paper, a robust and highly adaptive end-to-end framework is proposed for quality enhancement of almost all kind of images. On the first sight, the term “gamma correction” seems somehow conventional; but any approach which is capable for computing the quality enhanced intensity distribution out of the input intensity distribution through raising radical powers comes under the head of the gamma correction. Decision making of adaptive gamma value-set precisely for each individual intensity level of the image, is still an open problem, as most of the proposed gamma based (radically powered) algorithms lead to over-enhancement and extreme ends’ saturation, and hereby proposed algorithm seems free from these drawbacks due to deciding a novel kind of gamma value set through “optimal PSNR based perfectly re-allocated and reconstructed” intensity distribution. Here, as such no greedy behavior based optimization algorithm is involved for a blind random search, and hence, the approach is not iterative as a whole. It needs only 2–4 iterations at most for thresholds identification and

subsequent histogram division based on optimal PSNR value, but gamma value-set evaluation non-iterative at all. Here, a precisely re-allocated intensity-span is derived through reconstruction of the image by considering first and second moment for histogram sub-division, and later the cumulative distribution of the reconstructed images itself is utilized for deriving a gamma value set. The corresponding individual values from this set when raised up as radicals over the reconstructed and re-allocated intensity levels of the image under consideration leads to the overall quality enhancement. Remaining manuscript is drafted as follows: after brief literature survey and basic introduction in section 1; section 2 explains the proposed algorithm followed by its stepwise framework. Later, section 3 deals with the experimentation followed by corresponding results and discussion; and in section 4, conclusions are drawn.

2 Proposed methodology

Hue-Saturation-Intensity (HSI) colour image model is generally utilized for separation of chromatic as well as non-chromatic image information. For the proposed quality enhancement for the colour images, hue and saturation channels can be kept unaltered along with relevant processing over intensity channel. The entire methodology using process-flow diagram is presented in Fig. 1, and the corresponding step-wise procedure is as follows:

Step 1: Initially, all three channels (R , G , B) are linearly stretched for dynamic range expansion. For R -channel:

$$R(u, v) \leftarrow \frac{R(u, v) - R_{\min}}{R_{\max} - R_{\min}} \quad (1)$$

Here, $R_{\max} = \max \{R(u, v)\}$ and $R_{\min} = \min \{R(u, v)\}$ for all the pixel elements (u, v) for R -channel. Similarly, other two channels can be stretched.

Step 2: Extraction of intensity (luminance or V -channel) information after RGB to HSI colour space conversion as:

$$[H(u, v), S(u, v), I(u, v)]^T = T_{RGB}^{HSI}[R(u, v), G(u, v), B(u, v)]^T, \quad (2)$$

Here, T_{RGB}^{HSI} is RGB to HSI transformation process.

Step 3: Histogram $\{H(h)\}$ of the luminance channel is employed for further processing. Here, $H(h)$ is count of pixels having h^{th} intensity value. Set $a \leftarrow \min(h)$ and $b \leftarrow \max(h)$ which also represents the entire range of histogram starting from its lowest pixel intensity value to largest pixel intensity value. Calculate the mean (μ) and standard deviation (σ) for this operational range $[a, b]$ of the histogram /sub-histogram (for next level division), using:

$$\mu = \frac{\sum_{h=a}^b hH(h)}{\sum_{h=a}^b H(h)}, \quad (3)$$

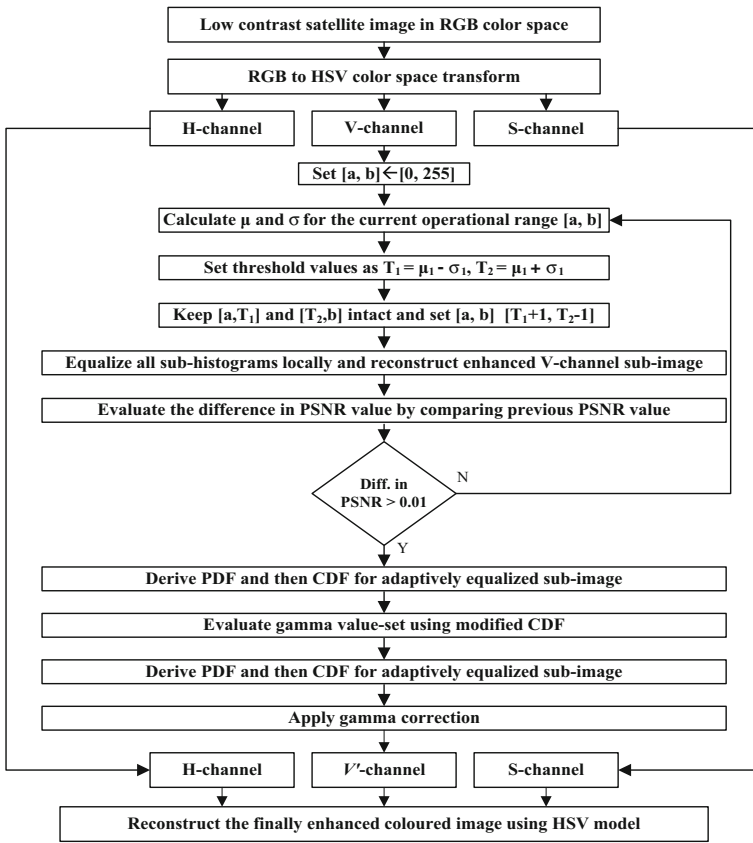


Fig. 1 Process Flow for the proposed methodology

$$\sigma = \left(\frac{\sum_{h=a}^b (h-\mu)^2 H(h)}{\sum_{h=a}^b H(h)} \right)^{1/2}, \tag{4}$$

- Step 4: Set two threshold values i.e. $T_1 = \mu - \sigma$ and $T_2 = \mu + \sigma$, so that the “the operational region” (mentioned in Step 3) can be distributed into its further sub-regions.
- Step 5: Store $[a, T_1]$ and $[T_2, b]$ as two parts of the histogram without further distributing them so that they can be retained as such till their equalization in subsequent steps. Consider $[T_1 + 1, T_2 - 1]$ as sub-histogram region $H_k(h)$ so that operations can perform the next step so that it can be adaptively distributed in further recursive steps.
- Step 6: Cumulative distribution function (CDF) for each k^{th} sub-histogram can be evaluated as:

$$cdf_k(h) = \frac{1}{N_k} \sum_{h_{k+1}}^{h_{k+1}} H_k(h), \tag{5}$$

Here, intensity span of every k^{th} histogram can be considered in the range $[h_k + 1 \rightarrow h_{k+1}]$. Here, N_k is the net pixel count in k^{th} sub-histogram.

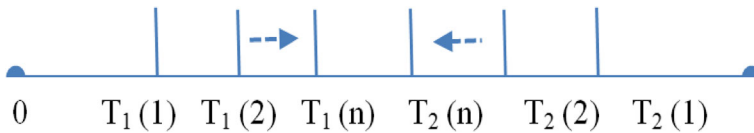


Fig. 2 Multilevel thresholding of intensity value axis

Step 7: Equalize all sub-histograms independently as:

$$\hat{I}_k = I_{k_min} + (I_{k_max} - I_{k_min}) * cdf_k(h), \tag{6}$$

Step 8: Overall reconstructed image can be derived as:

$$\hat{I} = \hat{I}_1 \cup \hat{I}_2 \cup \hat{I}_3 \cup \dots \cup \hat{I}_k; \tag{7}$$

Step 9: Calculate the value of PSNR in dB for enhanced intensity channel obtained in this iteration with reference to that in previous iteration as [31]:

$$PSNR = 10 \log_{10} \frac{255}{MSE}, \tag{8}$$

Here, RSME is root-mean-square error, defined as [31]:

$$MSE = \frac{1}{M \times N} \left\| \hat{I} - I \right\|_2^2, \tag{9}$$

Here, I and \hat{I} are input and output images for every iteration. Find the difference of PSNR value in this step with that obtained in the previous step.

Table 1 Number of iterations and corresponding threshold values evaluated for images under consideration

Image S. No.	No. of iterations (i _{max})	Threshold values in lower intensity region T ₁ (i)	Threshold values in higher intensity region T ₂ (i)
1.	2	[21, 37]	[95, 54]
2.	2	[29, 48]	[102, 68]
3.	2	[34, 49]	[99, 73]
4.	2	[19, 42]	[95, 81]
5.	3	[46, 68, 87]	[139, 112, 94]
6.	2	[25, 48]	[98, 75]
7.	2	[33, 51]	[118, 89]
8.	2	[37, 45]	[94, 53]
9.	2	[21, 47]	[85, 61]
10.	2	[31, 48]	[122, 78]
11.	2	[40, 57]	[99, 73]
12.	2	[34, 59]	[126, 93]
13.	2	[32, 49]	[119, 84]
14.	2	[35, 53]	[121, 78]
15.	2	[28, 52]	[117, 81]
16.	3	[68, 84, 102]	[140, 122, 110]
17.	3	[33, 42, 98]	[150, 127, 111]
18.	2	[23, 45]	[95, 81]
19.	2	[31, 48]	[122, 78]
20.	2	[34, 53]	[96, 83]

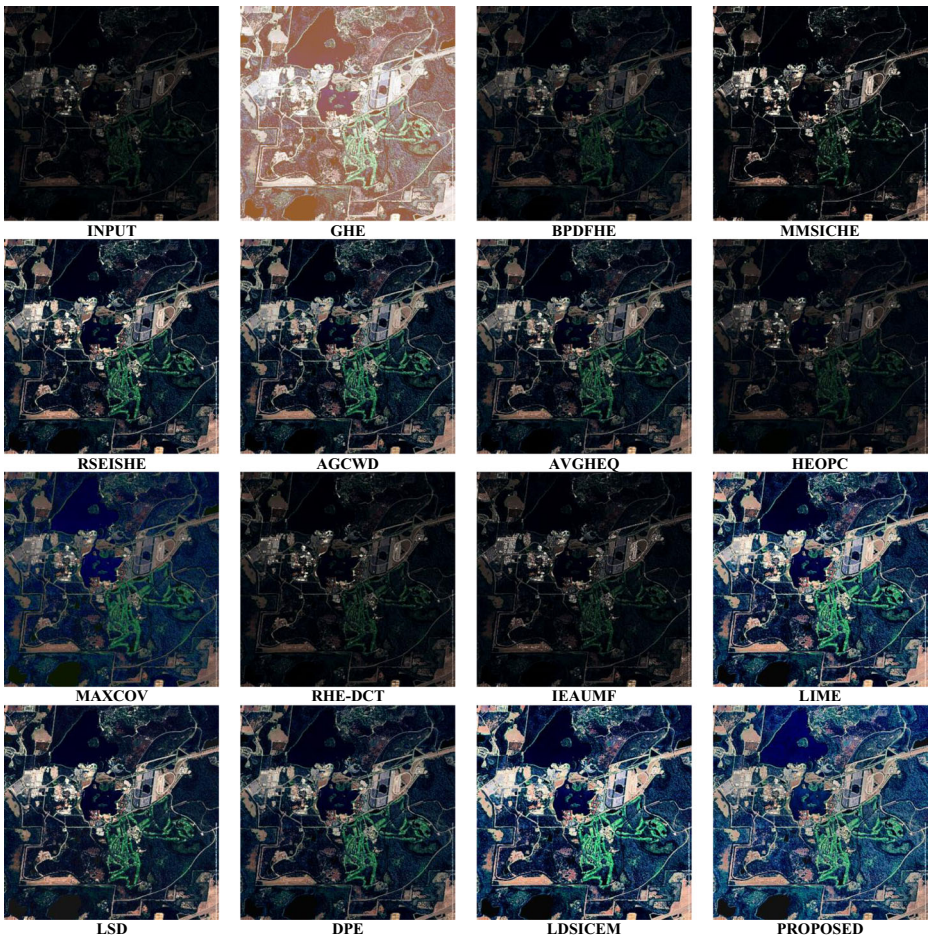


Fig. 3 Quality enhanced results of different algorithms for “Image 1”

- Step 10: Now, follow the optimal PSNR criterion to decide the requirement of next level thresholding. Here, recursion is aborted if difference in PSNR values (obtained in successive steps) gets reduced to less than 0.01 dB. In other words, next level thresholding has to be aborted when PSNR value gets saturated, as this saturation symbolizes insignificant further image division/reconstruction; and hence, not appreciated.
- Step 11: If the optimal PSNR criterion as mentioned in step-10 is not achieved, then assign $[a, b] \leftarrow [T_1 + 1, T_2 - 1]$ and repeat steps 3–9 for further adaptive separation; and hence, adaptively equalized output can be achieved.
- Step 12: Afterwards, cumulative distribution has to be derived reconstructed image so that the adaptive gamma value-set can be derived as:

$$\gamma(i) = 1 - cdf_m(i), \quad (10)$$

Finally, the enhanced output is achieved as:

$$I_{en}(i) = [I(i)]^{\gamma(i)}, \quad (11)$$

Table 2 Brightness (B) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.071	0.692	0.101	0.125	0.239	0.225	0.178	0.097	0.267	0.094	0.097	0.359	0.270	0.229	0.396	0.436
2	0.181	0.578	0.208	0.230	0.286	0.321	0.229	0.227	0.342	0.236	0.242	0.388	0.421	0.339	0.470	0.377
3	0.257	0.605	0.266	0.295	0.333	0.408	0.319	0.322	0.385	0.344	0.324	0.425	0.454	0.410	0.571	0.474
4	0.145	0.678	0.198	0.175	0.259	0.273	0.194	0.181	0.425	0.185	0.187	0.343	0.383	0.321	0.467	0.406
5	0.352	0.638	0.398	0.379	0.345	0.521	0.637	0.449	0.528	0.455	0.467	0.485	0.510	0.478	0.636	0.599
6	0.316	0.544	0.316	0.350	0.328	0.427	0.394	0.394	0.462	0.403	0.407	0.465	0.450	0.420	0.546	0.454
7	0.127	0.599	0.159	0.178	0.290	0.289	0.244	0.158	0.314	0.165	0.171	0.363	0.424	0.347	0.499	0.431
8	0.297	0.524	0.306	0.331	0.340	0.427	0.386	0.370	0.409	0.388	0.376	0.433	0.470	0.456	0.606	0.469
9	0.175	0.595	0.185	0.210	0.274	0.317	0.218	0.218	0.401	0.226	0.235	0.372	0.396	0.402	0.541	0.424
10	0.313	0.549	0.320	0.347	0.350	0.466	0.396	0.389	0.472	0.402	0.405	0.512	0.474	0.484	0.621	0.497
11	0.235	0.566	0.271	0.279	0.338	0.406	0.292	0.258	0.406	0.311	0.314	0.454	0.457	0.470	0.608	0.466
12	0.235	0.590	0.269	0.277	0.372	0.409	0.319	0.293	0.413	0.328	0.312	0.459	0.490	0.477	0.616	0.473
13	0.569	0.569	0.292	0.320	0.355	0.443	0.365	0.350	0.463	0.356	0.371	0.480	0.437	0.430	0.583	0.502
14	0.552	0.552	0.332	0.356	0.387	0.438	0.462	0.408	0.478	0.410	0.413	0.452	0.470	0.461	0.594	0.468
15	0.586	0.586	0.285	0.287	0.342	0.396	0.343	0.323	0.395	0.328	0.326	0.407	0.448	0.444	0.585	0.451
16	0.631	0.631	0.200	0.223	0.286	0.327	0.237	0.235	0.380	0.238	0.246	0.380	0.407	0.355	0.496	0.406
17	0.608	0.608	0.362	0.404	0.348	0.493	0.542	0.492	0.540	0.491	0.498	0.504	0.460	0.458	0.590	0.528
18	0.570	0.570	0.218	0.252	0.295	0.351	0.292	0.270	0.353	0.280	0.287	0.424	0.465	0.400	0.527	0.435
19	0.623	0.623	0.156	0.219	0.301	0.254	0.375	0.261	0.412	0.269	0.276	0.440	0.381	0.329	0.457	0.386
20	0.695	0.695	0.173	0.231	0.265	0.284	0.293	0.251	0.388	0.260	0.263	0.379	0.397	0.294	0.417	0.359

Table 3 Contrast (V) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.007	0.018	0.011	0.043	0.057	0.052	0.035	0.009	0.015	0.013	0.017	0.080	0.054	0.041	0.065	0.098
2	0.029	0.056	0.047	0.066	0.071	0.076	0.047	0.044	0.041	0.056	0.065	0.105	0.083	0.082	0.074	0.131
3	0.030	0.052	0.036	0.060	0.049	0.060	0.049	0.046	0.052	0.059	0.050	0.088	0.094	0.062	0.058	0.094
4	0.032	0.024	0.043	0.057	0.069	0.068	0.057	0.049	0.050	0.055	0.057	0.091	0.074	0.074	0.073	0.127
5	0.009	0.061	0.052	0.034	0.024	0.031	0.054	0.018	0.029	0.033	0.057	0.081	0.082	0.053	0.051	0.080
6	0.056	0.072	0.058	0.082	0.067	0.083	0.088	0.085	0.076	0.107	0.114	0.116	0.090	0.091	0.084	0.126
7	0.014	0.047	0.022	0.052	0.062	0.060	0.047	0.021	0.027	0.028	0.032	0.082	0.073	0.066	0.069	0.112
8	0.044	0.080	0.049	0.070	0.058	0.067	0.072	0.067	0.072	0.080	0.085	0.097	0.093	0.068	0.066	0.096
9	0.030	0.048	0.032	0.058	0.065	0.068	0.050	0.047	0.048	0.057	0.064	0.090	0.092	0.083	0.066	0.123
10	0.027	0.071	0.035	0.052	0.041	0.055	0.044	0.042	0.054	0.057	0.074	0.081	0.091	0.077	0.059	0.101
11	0.020	0.070	0.041	0.050	0.052	0.057	0.029	0.023	0.045	0.046	0.055	0.081	0.090	0.078	0.064	0.105
12	0.017	0.052	0.036	0.047	0.047	0.052	0.034	0.026	0.036	0.047	0.054	0.081	0.086	0.076	0.063	0.100
13	0.068	0.068	0.031	0.051	0.052	0.063	0.045	0.040	0.053	0.048	0.082	0.094	0.100	0.068	0.066	0.106
14	0.080	0.080	0.062	0.084	0.087	0.075	0.100	0.095	0.089	0.107	0.099	0.111	0.093	0.085	0.076	0.112
15	0.051	0.051	0.046	0.065	0.067	0.066	0.070	0.063	0.065	0.073	0.068	0.090	0.098	0.078	0.065	0.108
16	0.041	0.041	0.035	0.060	0.066	0.071	0.054	0.049	0.051	0.056	0.064	0.094	0.091	0.078	0.069	0.124
17	0.064	0.064	0.040	0.059	0.046	0.064	0.085	0.066	0.066	0.078	0.105	0.103	0.095	0.087	0.076	0.118
18	0.061	0.061	0.039	0.066	0.066	0.074	0.066	0.053	0.055	0.069	0.078	0.106	0.079	0.088	0.078	0.125
19	0.037	0.037	0.023	0.068	0.080	0.048	0.118	0.056	0.049	0.067	0.081	0.131	0.090	0.078	0.070	0.127
20	0.017	0.017	0.035	0.077	0.079	0.075	0.099	0.068	0.057	0.079	0.093	0.134	0.081	0.080	0.085	0.134

Table 4 Entropy (H) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	5.475	3.860	5.216	5.357	5.421	5.405	6.045	5.541	5.780	5.746	5.769	5.422	5.431	6.335	6.142	7.745
2	6.190	5.989	5.788	6.126	6.102	6.021	6.381	6.234	6.318	6.268	6.533	6.097	6.070	6.708	6.340	7.279
3	7.262	6.768	6.902	7.210	7.125	6.968	7.533	7.370	7.648	7.662	7.528	7.159	7.034	7.804	7.383	7.856
4	6.060	4.663	5.720	5.990	5.929	5.830	6.168	6.120	6.359	6.205	6.247	5.944	5.925	6.786	6.460	7.492
5	6.673	7.220	6.495	6.603	6.586	6.549	7.477	7.110	7.430	7.540	7.698	6.566	6.551	7.845	7.438	7.715
6	7.065	6.864	6.671	7.031	7.000	6.637	7.282	7.101	7.182	7.126	7.278	6.939	6.958	7.330	6.845	7.693
7	6.147	5.657	5.905	6.051	6.087	6.034	6.748	6.196	6.553	6.323	6.471	6.078	6.058	7.159	6.820	7.760
8	7.404	7.407	7.026	7.348	7.275	7.024	7.717	7.545	7.696	7.666	7.732	7.270	7.230	7.916	7.438	7.886
9	6.681	5.740	6.315	6.620	6.536	6.433	6.728	6.753	6.912	6.764	6.957	6.567	6.528	7.425	6.995	7.798
10	7.268	6.946	6.894	7.200	7.165	7.003	7.563	7.316	7.691	7.621	7.823	7.165	7.091	7.937	7.411	7.829
11	7.024	6.859	6.695	6.966	6.946	6.855	7.316	7.118	7.580	7.394	7.577	6.949	6.864	7.884	7.366	7.871
12	6.950	6.128	6.701	6.888	6.865	6.795	7.223	7.038	7.398	7.530	7.451	6.854	6.800	7.948	7.412	7.882
13	6.627	6.627	6.539	6.861	6.831	6.734	7.187	6.917	7.347	7.247	7.583	6.833	6.661	7.657	7.261	7.767
14	7.379	7.379	6.891	7.299	7.243	6.808	7.573	7.430	7.494	7.382	7.599	7.202	7.161	7.832	7.287	7.849
15	6.154	6.154	6.767	7.113	6.994	6.814	7.438	7.191	7.473	7.346	7.424	7.049	6.914	7.778	7.281	7.838
16	5.792	5.792	6.057	6.414	6.333	6.227	6.518	6.571	6.667	6.711	6.732	6.358	6.255	7.071	6.694	7.600
17	7.161	7.161	6.621	6.918	6.850	6.728	7.467	7.288	7.348	7.519	7.532	6.863	6.830	7.745	7.237	7.721
18	6.411	6.411	6.381	6.677	6.645	6.496	6.972	6.909	7.001	6.881	7.049	6.599	6.626	7.270	6.806	7.668
19	5.181	5.181	5.608	5.876	5.895	5.727	6.364	6.240	6.296	6.554	6.568	5.816	5.878	6.596	6.258	7.127
20	3.937	3.937	5.341	5.679	5.655	5.465	5.781	5.832	5.769	6.273	5.997	5.544	5.660	6.026	5.731	6.826

Table 5 Sharpness (S) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.037	0.056	0.053	0.075	0.115	0.115	0.090	0.044	0.074	0.054	0.060	0.156	0.102	0.104	0.176	0.175
2	0.067	0.091	0.082	0.095	0.104	0.108	0.085	0.082	0.079	0.094	0.138	0.128	0.111	0.116	0.145	0.125
3	0.045	0.059	0.049	0.060	0.058	0.064	0.057	0.055	0.061	0.072	0.074	0.076	0.080	0.097	0.124	0.113
4	0.071	0.085	0.099	0.088	0.120	0.130	0.097	0.087	0.095	0.102	0.108	0.152	0.134	0.153	0.208	0.169
5	0.035	0.082	0.082	0.066	0.056	0.063	0.083	0.050	0.063	0.072	0.181	0.102	0.102	0.099	0.118	0.110
6	0.074	0.084	0.076	0.090	0.083	0.090	0.093	0.092	0.087	0.106	0.175	0.106	0.093	0.102	0.117	0.110
7	0.067	0.125	0.086	0.117	0.145	0.146	0.126	0.082	0.098	0.097	0.109	0.172	0.162	0.157	0.206	0.175
8	0.054	0.080	0.059	0.070	0.064	0.069	0.070	0.067	0.071	0.084	0.132	0.081	0.084	0.096	0.114	0.107
9	0.057	0.074	0.059	0.072	0.086	0.093	0.072	0.070	0.080	0.082	0.115	0.107	0.107	0.112	0.136	0.111
10	0.051	0.088	0.057	0.067	0.063	0.072	0.066	0.064	0.073	0.079	0.178	0.088	0.091	0.102	0.115	0.105
11	0.066	0.120	0.092	0.100	0.105	0.110	0.067	0.066	0.083	0.103	0.146	0.131	0.137	0.141	0.160	0.150
12	0.064	0.103	0.092	0.098	0.105	0.112	0.089	0.079	0.098	0.110	0.158	0.139	0.145	0.146	0.167	0.159
13	0.081	0.081	0.044	0.059	0.058	0.060	0.052	0.049	0.058	0.060	0.186	0.077	0.081	0.086	0.104	0.097
14	0.088	0.088	0.072	0.082	0.085	0.083	0.093	0.087	0.086	0.103	0.108	0.098	0.093	0.110	0.126	0.112
15	0.083	0.083	0.078	0.085	0.092	0.095	0.093	0.087	0.091	0.105	0.109	0.109	0.115	0.139	0.164	0.141
16	0.067	0.067	0.058	0.069	0.080	0.086	0.070	0.067	0.069	0.078	0.110	0.098	0.096	0.101	0.129	0.110
17	0.071	0.071	0.056	0.073	0.062	0.060	0.076	0.066	0.067	0.081	0.176	0.090	0.091	0.095	0.100	0.091
18	0.137	0.137	0.101	0.124	0.134	0.149	0.135	0.121	0.122	0.144	0.176	0.178	0.152	0.165	0.199	0.172
19	0.060	0.060	0.046	0.068	0.086	0.071	0.106	0.074	0.068	0.091	0.135	0.118	0.089	0.095	0.123	0.104
20	0.040	0.040	0.057	0.080	0.086	0.088	0.097	0.080	0.073	0.098	0.135	0.117	0.088	0.097	0.123	0.113

Table 6 Colorfulness (C) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.048	0.921	0.060	0.060	0.132	0.131	0.100	0.055	0.185	0.052	0.051	0.222	0.141	0.285	0.177	0.242
2	0.136	0.289	0.078	0.086	0.108	0.125	0.086	0.085	0.137	0.090	0.096	0.151	0.153	0.369	0.397	0.392
3	0.228	0.474	0.131	0.135	0.166	0.215	0.156	0.160	0.202	0.170	0.159	0.212	0.237	0.337	0.342	0.356
4	0.240	0.180	0.164	0.160	0.210	0.222	0.167	0.152	0.278	0.158	0.162	0.267	0.274	0.521	0.624	0.595
5	0.359	0.592	0.252	0.229	0.211	0.316	0.368	0.269	0.313	0.276	0.295	0.308	0.323	0.811	0.881	0.898
6	0.325	0.249	0.083	0.089	0.084	0.118	0.104	0.103	0.122	0.106	0.114	0.124	0.116	0.380	0.385	0.389
7	0.075	0.408	0.074	0.083	0.139	0.138	0.117	0.075	0.169	0.078	0.080	0.176	0.203	0.304	0.380	0.376
8	0.023	0.240	0.054	0.058	0.059	0.077	0.068	0.064	0.072	0.069	0.067	0.079	0.085	0.204	0.249	0.246
9	0.224	0.292	0.085	0.101	0.125	0.145	0.101	0.101	0.211	0.105	0.107	0.170	0.177	0.327	0.262	0.342
10	0.194	0.337	0.092	0.097	0.103	0.144	0.119	0.118	0.142	0.118	0.127	0.152	0.135	0.546	0.572	0.590
11	0.250	0.341	0.115	0.121	0.143	0.169	0.117	0.109	0.163	0.132	0.132	0.189	0.190	0.550	0.566	0.576
12	0.215	0.342	0.120	0.124	0.171	0.186	0.144	0.134	0.203	0.149	0.144	0.206	0.224	0.538	0.578	0.563
13	0.319	0.319	0.111	0.118	0.133	0.172	0.141	0.135	0.172	0.135	0.141	0.182	0.165	0.544	0.584	0.598
14	0.327	0.327	0.110	0.118	0.128	0.145	0.152	0.134	0.159	0.134	0.136	0.150	0.154	0.465	0.459	0.471
15	0.367	0.367	0.117	0.110	0.138	0.174	0.139	0.129	0.162	0.133	0.132	0.172	0.196	0.539	0.612	0.564
16	0.456	0.456	0.136	0.159	0.194	0.222	0.161	0.160	0.261	0.162	0.168	0.256	0.265	0.537	0.656	0.606
17	0.562	0.562	0.208	0.228	0.199	0.291	0.304	0.278	0.309	0.285	0.288	0.292	0.262	0.564	0.626	0.627
18	0.321	0.321	0.088	0.104	0.119	0.141	0.117	0.109	0.150	0.113	0.110	0.170	0.183	0.489	0.491	0.499
19	0.320	0.320	0.087	0.121	0.166	0.147	0.211	0.134	0.234	0.153	0.150	0.254	0.200	0.121	0.135	0.139
20	0.153	0.153	0.120	0.160	0.183	0.203	0.205	0.166	0.244	0.183	0.186	0.268	0.241	0.599	0.680	0.650

Table 7 GLCM Homogeneity (GH) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.903	0.759	0.820	0.843	0.641	0.652	0.704	0.863	0.686	0.839	0.833	0.556	0.634	0.644	0.539	0.501
2	0.760	0.687	0.733	0.715	0.677	0.669	0.718	0.721	0.677	0.705	0.629	0.636	0.631	0.650	0.627	0.615
3	0.801	0.758	0.789	0.771	0.757	0.734	0.761	0.762	0.740	0.719	0.713	0.713	0.708	0.635	0.612	0.597
4	0.760	0.712	0.684	0.740	0.629	0.618	0.695	0.715	0.644	0.693	0.687	0.575	0.566	0.558	0.523	0.511
5	0.815	0.674	0.659	0.722	0.730	0.704	0.668	0.753	0.709	0.679	0.522	0.621	0.621	0.605	0.603	0.590
6	0.723	0.693	0.717	0.691	0.703	0.695	0.680	0.681	0.675	0.669	0.552	0.671	0.672	0.651	0.654	0.629
7	0.762	0.602	0.716	0.705	0.590	0.592	0.624	0.713	0.628	0.691	0.673	0.552	0.540	0.562	0.528	0.516
8	0.767	0.696	0.755	0.742	0.742	0.729	0.726	0.732	0.721	0.701	0.607	0.707	0.699	0.641	0.640	0.621
9	0.781	0.701	0.770	0.755	0.701	0.681	0.742	0.743	0.673	0.723	0.673	0.651	0.647	0.627	0.621	0.611
10	0.788	0.698	0.772	0.749	0.757	0.723	0.743	0.747	0.715	0.712	0.524	0.691	0.692	0.651	0.659	0.641
11	0.742	0.633	0.693	0.694	0.659	0.641	0.726	0.734	0.678	0.660	0.596	0.609	0.606	0.583	0.594	0.561
12	0.726	0.648	0.673	0.672	0.634	0.611	0.663	0.690	0.636	0.622	0.565	0.572	0.566	0.554	0.567	0.528
13	0.696	0.696	0.806	0.766	0.769	0.753	0.778	0.788	0.747	0.757	0.544	0.713	0.712	0.672	0.661	0.647
14	0.685	0.685	0.721	0.711	0.705	0.706	0.690	0.701	0.696	0.681	0.652	0.688	0.682	0.628	0.633	0.616
15	0.699	0.699	0.708	0.708	0.677	0.660	0.677	0.690	0.665	0.658	0.642	0.646	0.634	0.569	0.568	0.556
16	0.755	0.755	0.776	0.759	0.724	0.708	0.749	0.753	0.705	0.731	0.686	0.685	0.680	0.663	0.641	0.629
17	0.689	0.689	0.734	0.691	0.723	0.720	0.679	0.703	0.694	0.665	0.538	0.656	0.651	0.632	0.653	0.636
18	0.580	0.580	0.644	0.619	0.578	0.554	0.577	0.597	0.561	0.573	0.538	0.520	0.513	0.522	0.515	0.494
19	0.755	0.755	0.812	0.758	0.695	0.746	0.668	0.723	0.722	0.694	0.627	0.657	0.669	0.670	0.641	0.638
20	0.804	0.804	0.776	0.729	0.708	0.713	0.692	0.714	0.715	0.690	0.646	0.668	0.711	0.676	0.653	0.641

Table 8 GLCM Energy (GE) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.553	0.210	0.313	0.526	0.091	0.111	0.152	0.434	0.118	0.392	0.425	0.035	0.079	0.083	0.033	0.022
2	0.141	0.081	0.145	0.128	0.081	0.076	0.106	0.107	0.056	0.111	0.101	0.061	0.045	0.070	0.067	0.052
3	0.114	0.073	0.102	0.106	0.074	0.053	0.078	0.078	0.057	0.060	0.070	0.050	0.051	0.037	0.045	0.030
4	0.241	0.161	0.148	0.228	0.084	0.079	0.153	0.173	0.055	0.163	0.166	0.053	0.038	0.051	0.045	0.030
5	0.177	0.051	0.042	0.105	0.080	0.060	0.047	0.098	0.066	0.055	0.032	0.031	0.031	0.032	0.048	0.036
6	0.069	0.055	0.069	0.060	0.065	0.060	0.051	0.050	0.038	0.057	0.045	0.055	0.045	0.046	0.066	0.039
7	0.240	0.055	0.177	0.213	0.053	0.056	0.072	0.149	0.061	0.137	0.139	0.039	0.027	0.039	0.032	0.025
8	0.089	0.040	0.077	0.081	0.067	0.055	0.055	0.060	0.051	0.051	0.045	0.050	0.044	0.032	0.056	0.030
9	0.182	0.084	0.168	0.165	0.090	0.073	0.134	0.131	0.054	0.130	0.126	0.053	0.054	0.041	0.056	0.033
10	0.097	0.053	0.086	0.081	0.076	0.053	0.064	0.066	0.048	0.050	0.027	0.040	0.043	0.033	0.068	0.037
11	0.099	0.039	0.078	0.086	0.051	0.037	0.075	0.088	0.045	0.053	0.053	0.029	0.032	0.026	0.051	0.025
12	0.104	0.064	0.072	0.087	0.047	0.035	0.060	0.076	0.047	0.043	0.061	0.026	0.025	0.023	0.050	0.023
13	0.057	0.057	0.102	0.091	0.080	0.066	0.082	0.089	0.055	0.075	0.040	0.051	0.064	0.042	0.058	0.040
14	0.045	0.045	0.064	0.062	0.054	0.058	0.046	0.054	0.046	0.056	0.048	0.056	0.044	0.033	0.063	0.030
15	0.078	0.078	0.073	0.083	0.055	0.045	0.055	0.061	0.043	0.057	0.056	0.044	0.044	0.027	0.044	0.025
16	0.115	0.115	0.146	0.143	0.088	0.075	0.114	0.114	0.055	0.111	0.107	0.064	0.064	0.060	0.061	0.042
17	0.051	0.051	0.083	0.064	0.077	0.073	0.048	0.058	0.052	0.048	0.031	0.048	0.048	0.041	0.067	0.043
18	0.054	0.054	0.091	0.084	0.053	0.042	0.054	0.060	0.034	0.061	0.059	0.033	0.023	0.031	0.040	0.024
19	0.140	0.140	0.198	0.166	0.092	0.121	0.080	0.104	0.079	0.105	0.099	0.079	0.072	0.080	0.073	0.058
20	0.248	0.248	0.193	0.164	0.137	0.133	0.133	0.138	0.104	0.142	0.139	0.116	0.150	0.119	0.106	0.093

Table 9 GLCM Correlation (GC) values for comparative quantitative evaluation among various algorithms

S.No	INPUT	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF	LIME	LSD	DPE	LDSICEM	PROPOSED
1	0.516	0.544	0.500	0.534	0.544	0.533	0.550	0.506	0.357	0.480	0.432	0.480	0.540	0.530	0.462	0.284
2	0.575	0.613	0.592	0.587	0.597	0.603	0.586	0.584	0.591	0.573	0.417	0.607	0.610	0.595	0.592	0.529
3	0.802	0.816	0.803	0.805	0.804	0.817	0.811	0.810	0.813	0.770	0.755	0.822	0.819	0.678	0.647	0.601
4	0.663	0.436	0.583	0.652	0.602	0.554	0.647	0.660	0.636	0.604	0.551	0.547	0.546	0.469	0.416	0.388
5	0.656	0.757	0.719	0.682	0.701	0.712	0.724	0.690	0.708	0.663	0.174	0.723	0.724	0.639	0.639	0.570
6	0.662	0.674	0.662	0.666	0.657	0.665	0.667	0.666	0.663	0.653	0.474	0.671	0.670	0.652	0.649	0.600
7	0.357	0.365	0.351	0.342	0.348	0.344	0.351	0.345	0.343	0.320	0.277	0.346	0.348	0.334	0.325	0.277
8	0.801	0.779	0.792	0.792	0.799	0.794	0.803	0.803	0.802	0.761	0.603	0.806	0.794	0.700	0.677	0.644
9	0.679	0.670	0.681	0.691	0.690	0.674	0.692	0.691	0.665	0.659	0.528	0.677	0.672	0.647	0.627	0.602
10	0.714	0.720	0.731	0.736	0.716	0.747	0.727	0.725	0.737	0.712	0.386	0.755	0.758	0.691	0.683	0.613
11	0.561	0.604	0.568	0.547	0.572	0.605	0.653	0.599	0.677	0.550	0.359	0.606	0.607	0.553	0.549	0.447
12	0.535	0.601	0.558	0.548	0.564	0.572	0.562	0.550	0.543	0.546	0.295	0.579	0.575	0.530	0.505	0.405
13	0.761	0.761	0.798	0.782	0.800	0.837	0.814	0.812	0.829	0.787	0.393	0.831	0.820	0.758	0.762	0.723
14	0.722	0.722	0.732	0.740	0.740	0.722	0.738	0.741	0.743	0.699	0.693	0.731	0.727	0.647	0.622	0.606
15	0.710	0.710	0.724	0.715	0.729	0.724	0.726	0.728	0.731	0.675	0.663	0.726	0.724	0.559	0.540	0.501
16	0.679	0.679	0.698	0.707	0.711	0.702	0.709	0.709	0.730	0.664	0.543	0.706	0.702	0.655	0.643	0.600
17	0.760	0.760	0.736	0.710	0.717	0.787	0.779	0.776	0.783	0.732	0.460	0.750	0.730	0.702	0.736	0.712
18	0.417	0.417	0.462	0.471	0.465	0.438	0.463	0.466	0.472	0.417	0.321	0.439	0.442	0.434	0.398	0.366
19	0.721	0.721	0.743	0.794	0.768	0.719	0.753	0.754	0.759	0.698	0.545	0.721	0.758	0.721	0.697	0.659
20	0.658	0.658	0.718	0.742	0.736	0.684	0.722	0.729	0.743	0.670	0.546	0.692	0.706	0.670	0.671	0.630

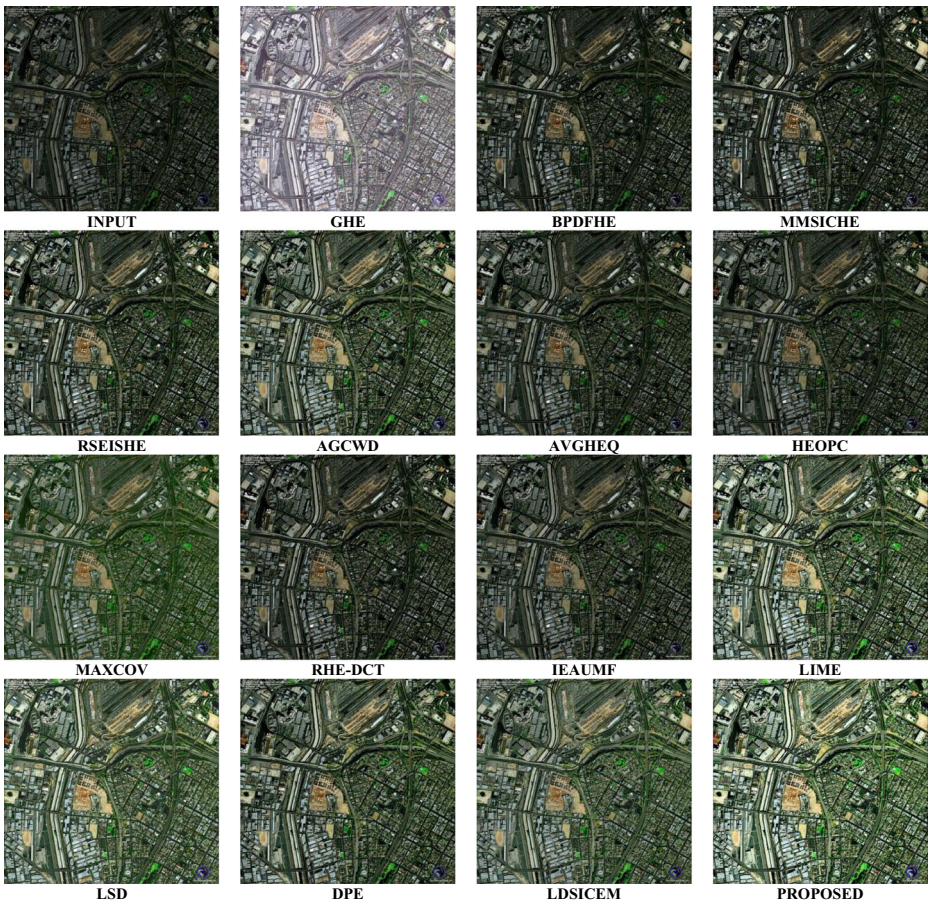


Fig. 4 Quality enhanced results of different algorithms for “Image 2”

Step 13: Finally, the enhanced image can be obtained as:

$$[R(u, v), G(u, v), B(u, v)]^T = T_{HSI}^{RGB} [H(u, v), S(u, v), \hat{I}(u, v)]^T, \tag{12}$$

Here, T_{HSI}^{RGB} is HSI to RGB transformation process.

At the first attempt, two (2) threshold values are identified and hence, results into three (3) sub-histograms, followed by their individual equalization. If the stopping criterion will not get satisfied (i.e., PSNR >0.01 dB), then both of the above thresholds will be treated as extreme end of the middle sub-histogram which is further subdivided in the similar fashion as mentioned above. Hence, the new threshold values will be identified in-between the previous threshold values. In this manner, by the end of second attempt of division, there will be four (4) threshold values and accordingly five (5) sub-histograms. In most of these cases, it is insignificant to looking forward for further sub-division.



Fig. 5 Quality enhanced results of different algorithms for “Image 3”

3 Experimental results: performance evaluation and comparisons

Multilevel iterative thresholds are shown in Fig. 2. Table 1 lists the number of iterations and corresponding threshold values evaluated iteratively (as shown in Fig. 3) for all test images. The iteration-count varies adaptively according to the intensity spread of the image. Performance evaluation and comparison is done by proper reimplementing of some very popular state-of-the-art enhancement methodologies namely, GHE [5], BPDFHE [15], MMSICHE [18], RSEISHE [19], AGCWD, AVGHEQ [11], HEOPC [22], MAXCOV [23], RHE-DCT [4], IEAUMF [10], LIME [6], LSD [3], DPE [2] and LDSICEM [1]. Quantitative analysis (Tables 2, 3, 4, 5, 6, 7, 8 and 9) is done by using 8 reliable statistical performance measures namely, average brightness (B), average contrast (V), average discrete information content (or entropy, E), sharpness (S), and colorfulness (C) of the image. Considering intensity value $I(u, v)$ for pixel element located at u^{th} row and v^{th} column of its equivalent image $M \times N$ matrix whose size is similar to that of corresponding intensity channel of the image, and its performance measures can be formulated as follows.

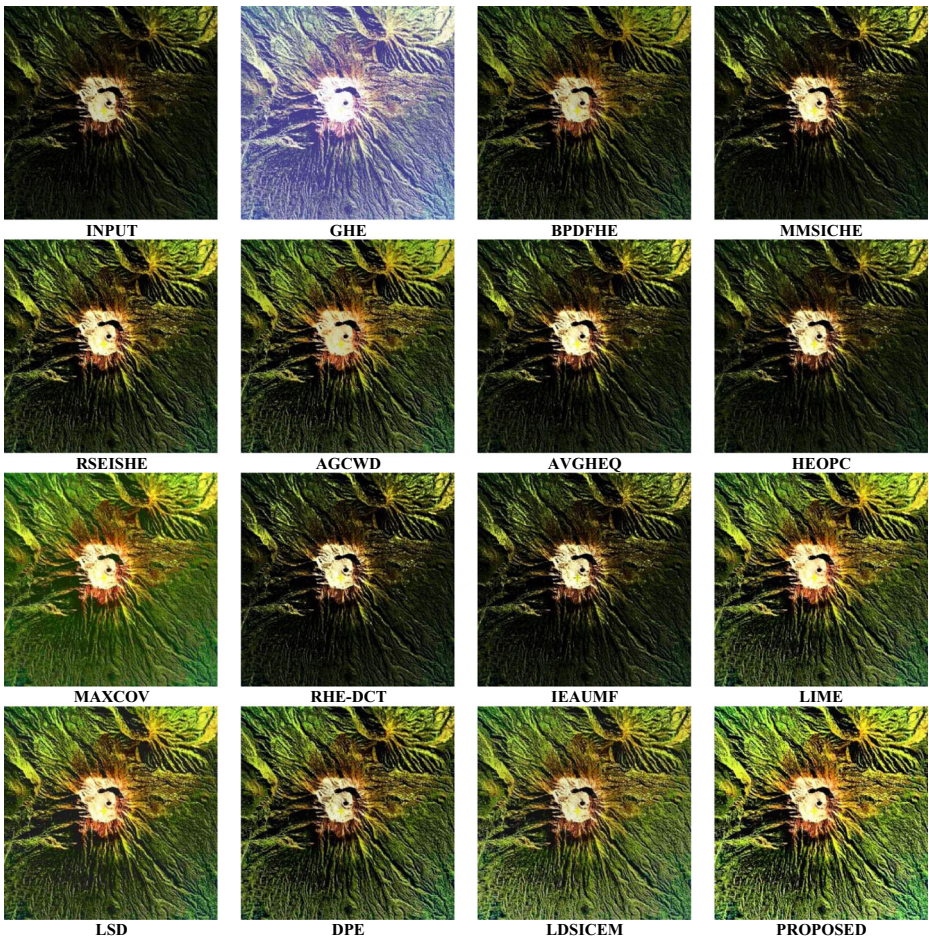


Fig. 6 Quality enhanced results of different algorithms for “Image 4”

Mean represents the average intensity value [11], which indirectly informs about the average image brightness level for the image under consideration. Brightness (B) or mean can be expressed as:

$$B = \frac{1}{M^*N} \sum_{u=1}^M \sum_{v=1}^N I(u, v), \tag{13}$$

Likewise, intensity spread or variance (V) or contrast indicates the amount of intensity deviation per pixel with respect to the mean intensity level (B) of the image, as:

$$V = \frac{1}{M^*N} \sum_{u,v} I(u, v)^2 - \left(\frac{1}{M^*N} \sum_{u,v} I(u, v) \right)^2, \tag{14}$$

In this manner, the total sum of the intensity dispersions (w.r.t. mean level) can be identified as contrast and obviously it should be high for proper quality enhancement. In addition, for proper information content evaluation, Shannon entropy based characterization can be applied as:

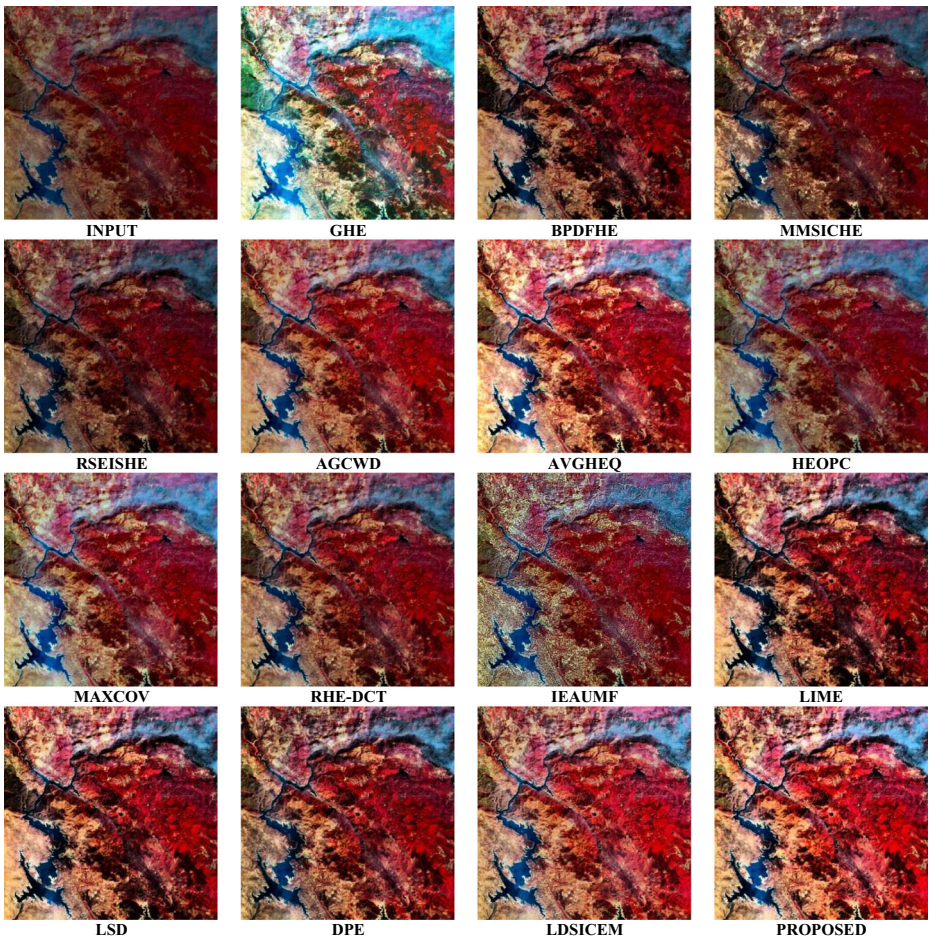


Fig. 7 Quality enhanced results of different algorithms for “Image 5”

$$H = - \sum_{i=0}^{I_{\max}} p_i \log_2(p_i), \tag{15}$$

where, $p_i = n_i / (M \times N)$ is the possibility of existence of i^{th} level of intensity, and I_{\max} is the maximum available intensity. Here, $M \times N$ represents the total number of pixels present in an image. The gradient is obtained from:

$$S = \frac{1}{M \times N} \sum_{u,v} \left(\sqrt{\Delta u^2 + \Delta v^2} \right), \tag{16}$$

$\Delta u = I_{enh}(u, v) - I_{enh}(u + 1, v)$ and $\Delta v = I_{enh}(u, v) - I_{enh}(u, v + 1)$ are the local gradients of enhanced image. Higher the gradient value more will be the sharpness of image. Along with above intensity based measures, colorfulness is also used for proper evaluation of the quality of color images. The colorfulness can be expressed numerically, as:

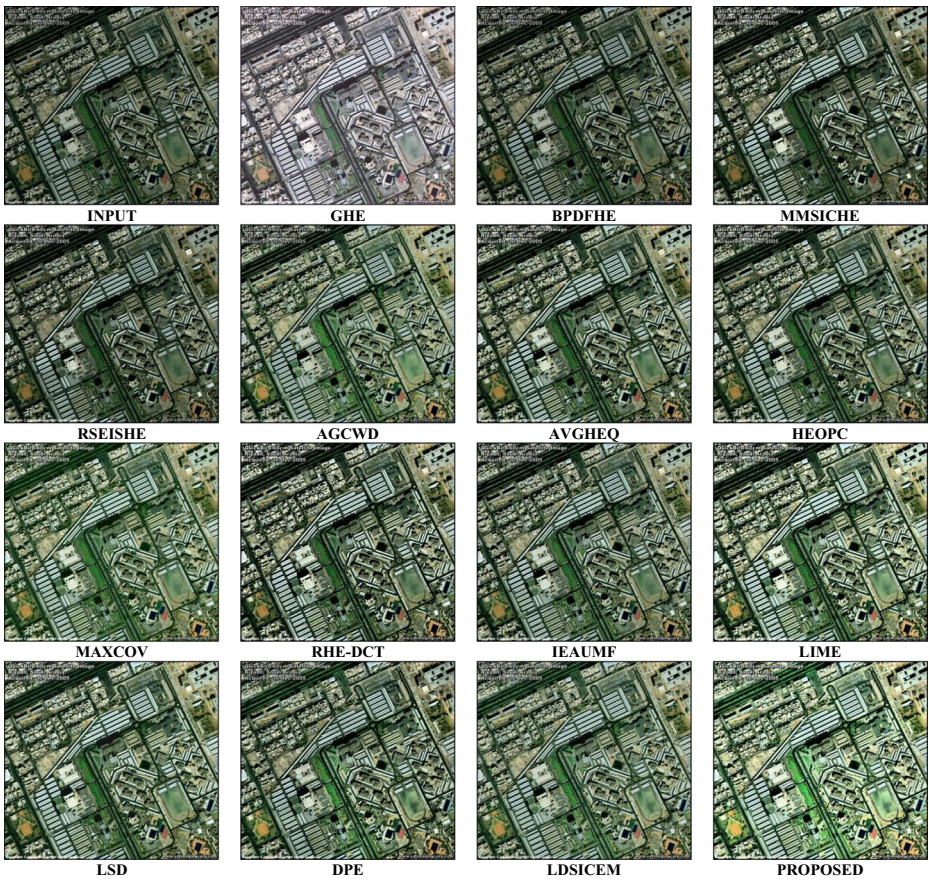


Fig. 8 Quality enhanced results of different algorithms for “Image 6”

$$C = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \tag{17}$$

$$\Delta rg = R - G, \tag{18}$$

$$\Delta yb = 0.5(R + G) - B, \tag{19}$$

Here, μ_{rg} , μ_{yb} are the mean values and σ_{rg} , σ_{yb} are the standard deviation values of Δrg , Δyb respectively. Spatial co-occurrence of the image pixels are usually avoided while evaluating the intensity based indices, and hence, to resolve it, Grey-Level Co-occurrence Matrix based performance indices also plays a significant role for texture and other spatially influenced properties. Overall statistical and spatial behavior w.r.t. reference pixel can be derived by calculating the pixel-wise average for all four directional matrices:

$$GLCM = 0.25(GLCM_0 + GLCM_{\pi/4} + GLCM_{\pi/2} + GLCM_{3\pi/4}); \tag{20}$$

In this paper, three well known GLCM based indices, i.e. GLCM-Correlation, GLCM-Energy and GLCM-Homogeneity are evaluated. Any element of the GLCM matrix $\Psi(m, n)$, is usually

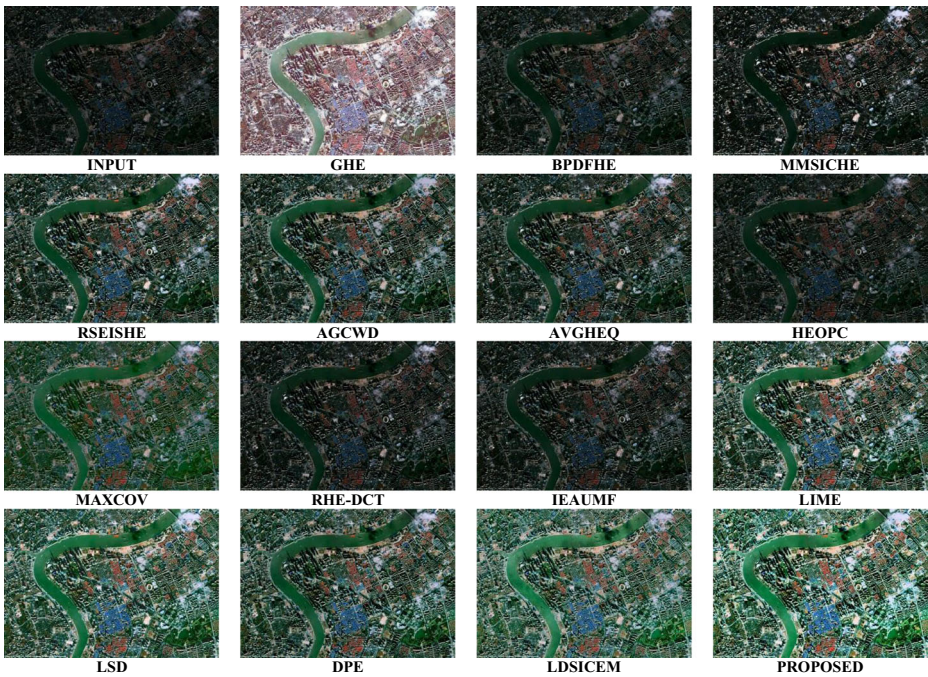


Fig. 9 Quality enhanced results of different algorithms for “Image 7”

evaluated by considering the n^{th} neighboring pixel w.r.t. m^{th} pixel, and later on, by calculating the μ_m , μ_n , σ_m , and σ_n as the corresponding mean values and standard deviation values respectively. GLCM-correlation (GC) stands for the interdependency for the corresponding neighborhood of the pixels w.r.t. reference pixels, expressed as:

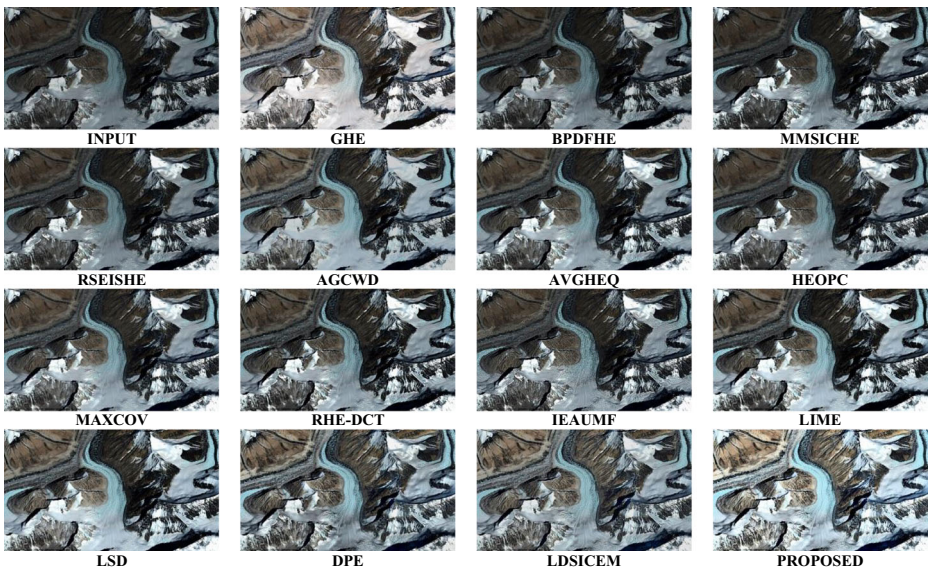


Fig. 10 Quality enhanced results of different algorithms for “Image 8”

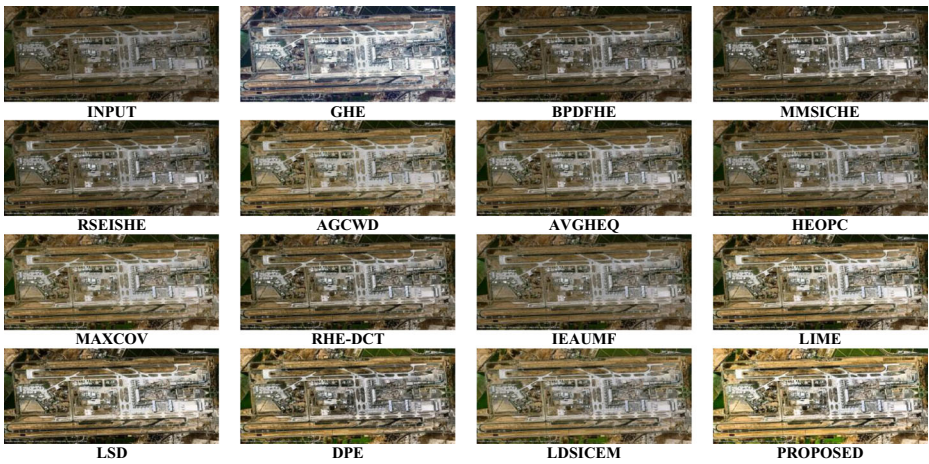


Fig. 11 Quality enhanced results of different algorithms for “Image 9”

$$GC = \frac{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (m-\mu_m)(n-\mu_n)\Psi(m, n)}{\sigma_m \cdot \sigma_n}, \tag{21}$$

GLCM-Energy (*GE*) can be characterized by normalized count of repeated pairs. Intuitively, these are responsible for uniformity of texture, and hence, expressed as:

$$GLCM-Energy(GE) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \Psi(m, n)^2, \tag{22}$$

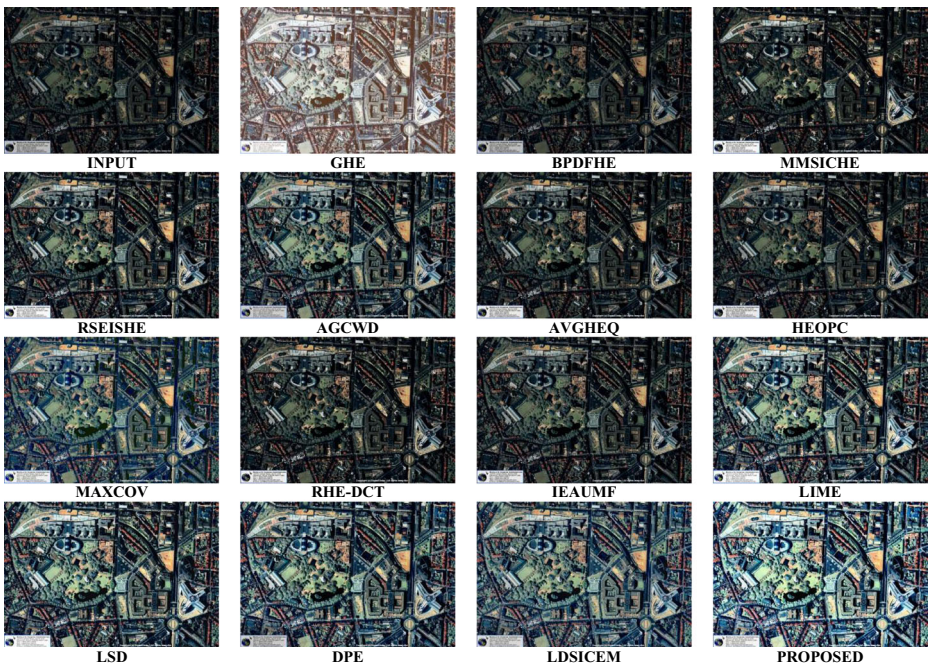


Fig. 12 Quality enhanced results of different algorithms for “Image 10”

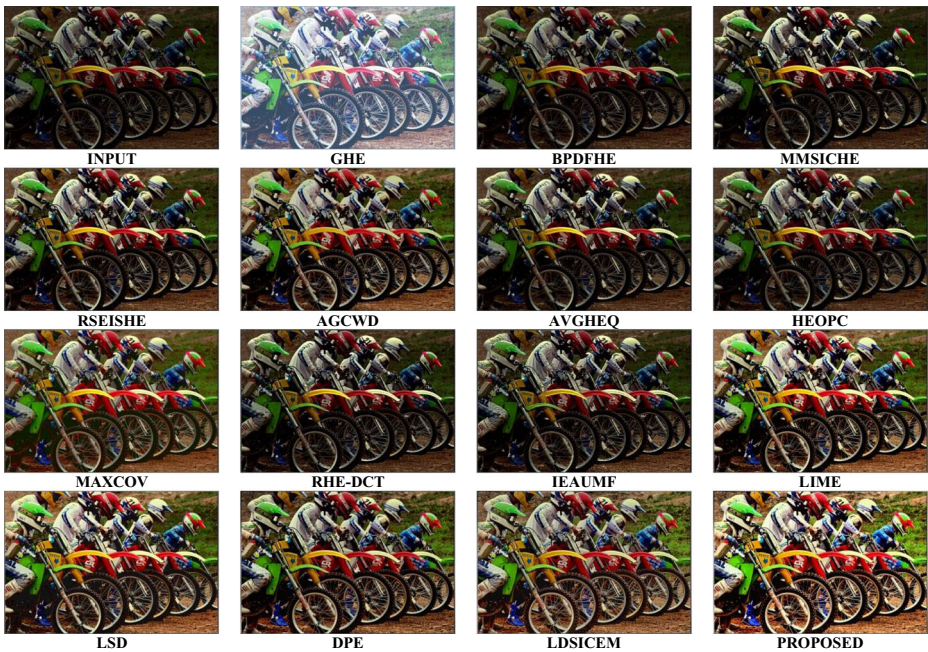


Fig. 13 Quality enhanced results of different algorithms for “Image 11”

GLCM-homogeneity (GH) can be characterized by the closeness of neighboring pixels with reference pixels. Intuitively, these are also responsible for uniformity of texture, and hence, expressed as:

$$GH = - \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \Psi(m, n) \log_2 \Psi(m, n), \quad (23)$$

Qualitative (visual) analysis for enhancement of images is shown in Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 23. Comparative evaluation for Brightness (B), Contrast (V), Entropy (E), Sharpness (S), colourfulness (C), GLCM-homogeneity (GH), GLCM-energy (GE), GLCM-correlation (GC) are listed in Tables 2 to 9, respectively. It can be easily noticed from the tabular results that both entropy and contrast are highly desirable along with image sharpness content of the information. Also, certain amount of brightness should be also increased, which is also desired for clear contrast evaluation in case of dark images.

Also, for identifying the textural improvement, GLCM based performance measures like GLCM- are also employed and the excellence of the proposed model, and the lower value are desired for GLCM-homogeneity, GLCM-energy, GLCM-correlation for better visualization in context of both human as well as machine-vision perspective.

Finally, it can be easily concluded that this approach outperforms the other state-of-the-art approaches. The novelty of the work can be justified as the re-allocation of intensity levels for corresponding pixel elements is so precise due to least successive differential change in PSNR value which ensures that further division or further reconstruction is obviously redundant. As this statistical moment-based redistribution needs only 2–4 iterations at most for subsequent histogram division, otherwise this approach is free from iterative greedy algorithms and hence system complexity is not

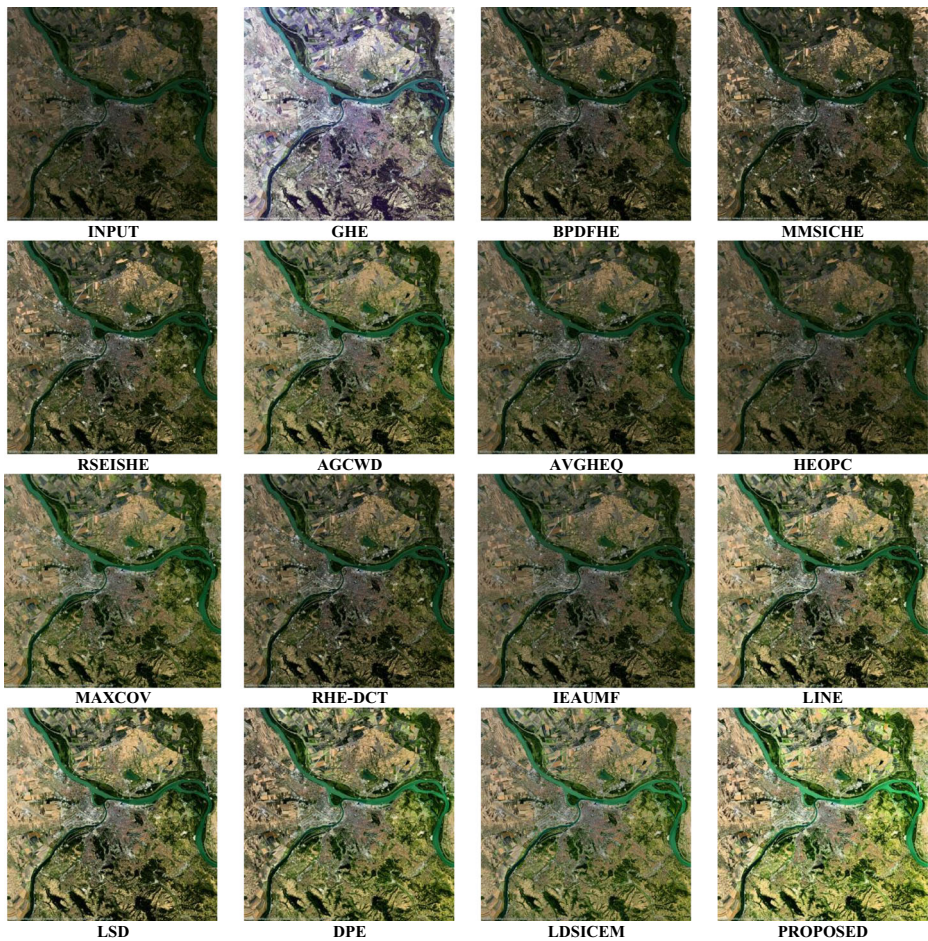


Fig. 14 Quality enhanced results of different algorithms for “Image 12”

so high. Due to this adaptive behavior of the intensity distribution the gamma value-set when derived from it, is obviously highly adaptive and here individual gamma values those evaluated explicitly raised over reconstructed intensity values, unlike conventional gamma correction methods. Unlike greedy algorithms, it is a parameter-free approach, hence no pre-specified count for subdivisions. It imparts the better gamma-corrected intensity distribution throughout the dynamic range. In addition multiple repetitive equalizations like other methods have been avoided for extreme intensity levels according to the image behavior. Here, only the in-between middle range ($\mu_1 - \sigma_1, \mu_1 + \sigma_1$) is only operated for further sub-division (which is also limited to 2–3 iterations) the range and rest of the intensity values themselves decide their adaptive gamma value-set locally. This is the sole region that over-enhancement (which leads to saturated patches) and under-enhancement (which leads to dark patches) can be easily avoided and hence, naturally looking, quality enhanced images can be achieved. Desired time-complexity analysis is also presented in Table 10 and Fig. 22, by executing the proposed method as well as all the state-of-the-art methodologies in a similar environment. The running time is calculated as an averaged execution time for a set of 120 test images.

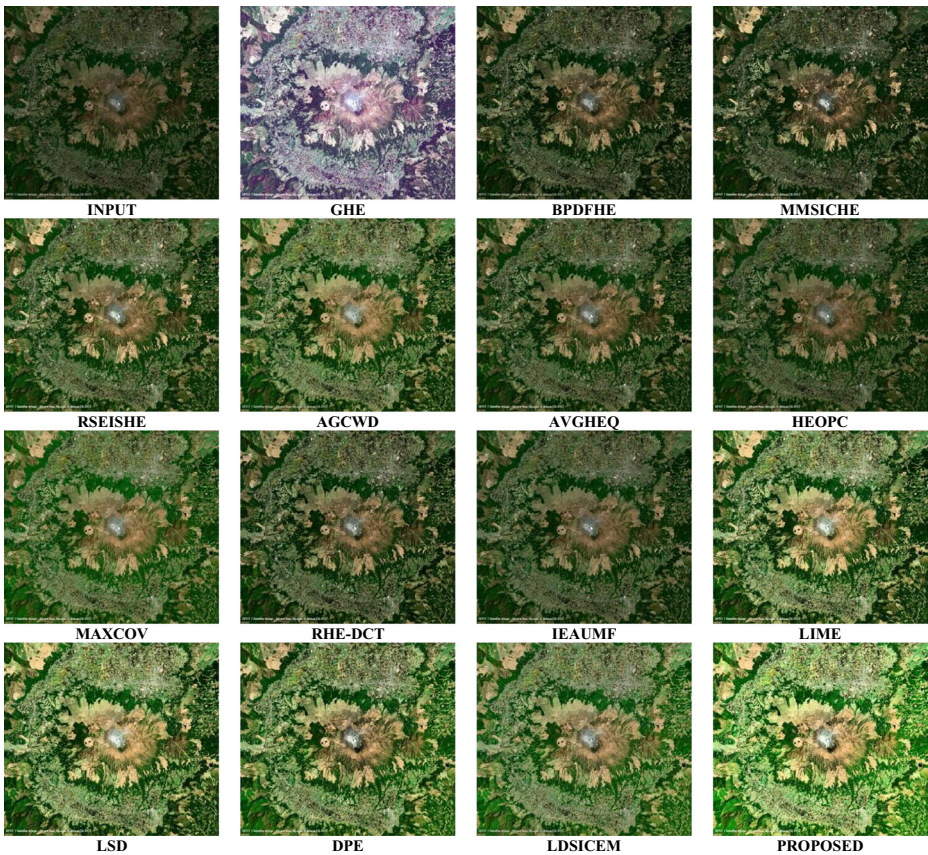


Fig. 15 Quality enhanced results of different algorithms for “Image 13”

4 Conclusion

In this paper, a new quality enhancement approach especially for dark or poorly illuminated images with a core objective to re-allocate the processed pixels using recursive histogram sub-division along with an adaptive stopping criterion based on pixel wise relative L_2 -norm basis (which itself is intuitively related to optimal PSNR value). Employing such kind information preserved signal reconstruction based stopping criterion makes the desired intensity distribution easy achievable in less iterations and hence complexity hike due iterative behaviour can be easily compensated to a great extent. Hence, iteration count only ranges from 2 to 3. Perfectly reconstructed, moment-centered piecewise sub-equalized statistical distribution which intuitively leads to the adaptive or image dependent evaluation of the desired gamma value-set, so that precise re-allocation of the transformed intensity bin-values. Due to this adaptive behavior of the intensity distribution the gamma value-set when derived from it, is obviously highly adaptive and here individual gamma values are evaluated explicitly raised over reconstructed intensity values, unlike conventional gamma correction methods. This adaptiveness makes the entire methodology highly capable for covering a wide variety of images, due to which robustness of the algorithm also increases. The proposed methodology has been verified on various dark images. The desired performance has been achieved visually and also measured by using relevant image quality matrices.

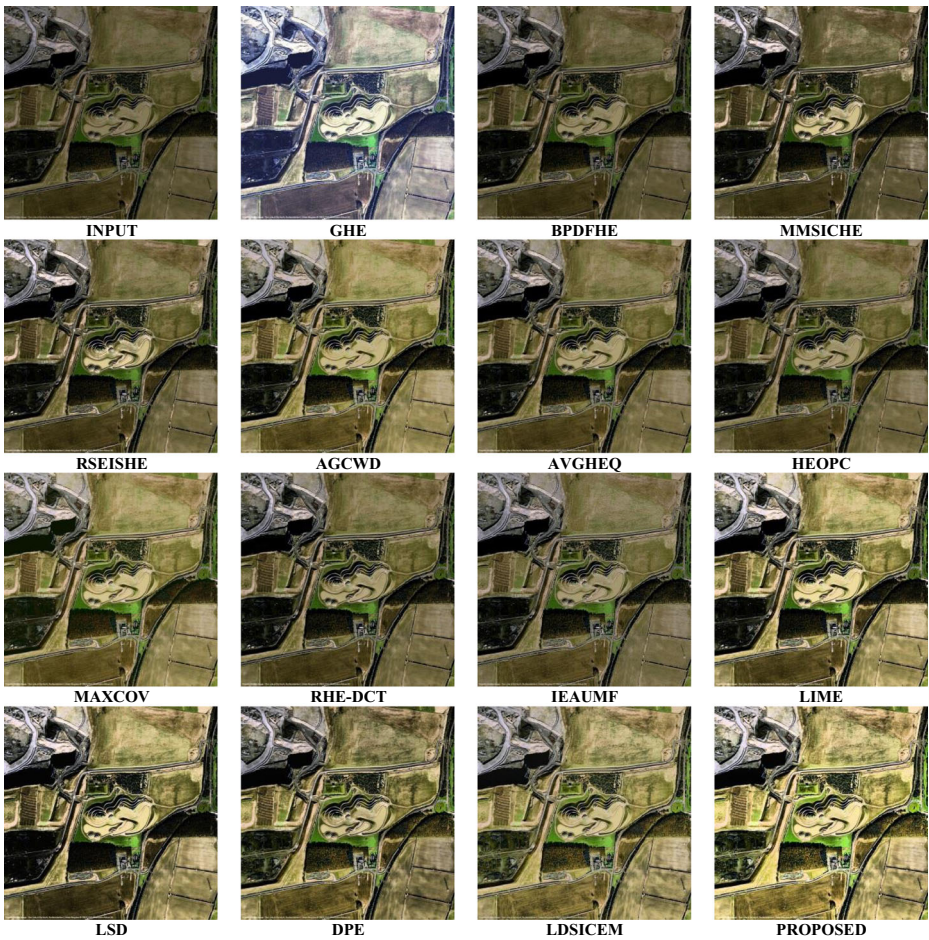


Fig. 16 Quality enhanced results of different algorithms for “Image 14”

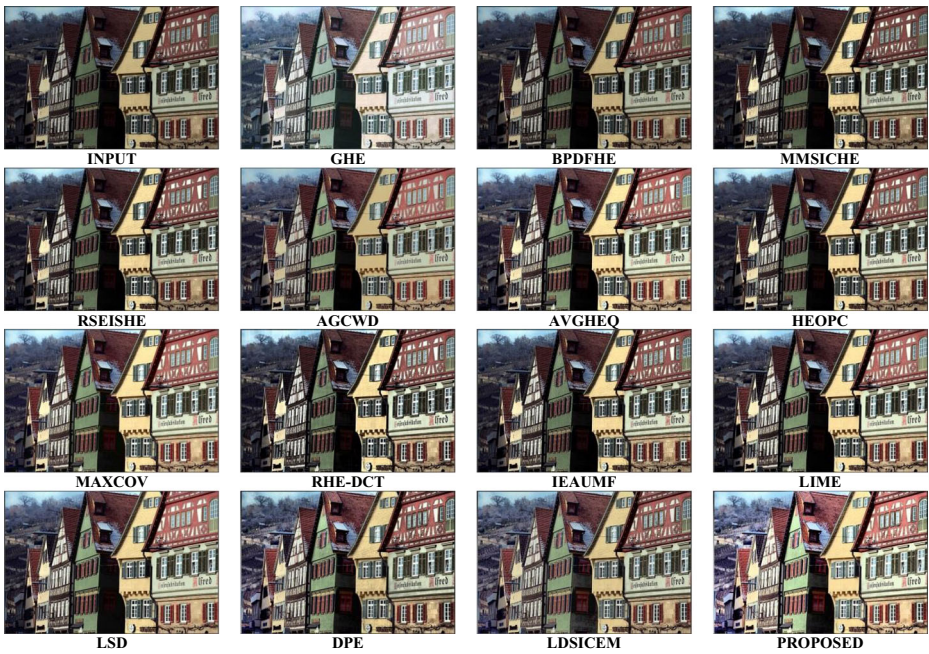


Fig. 17 Quality enhanced results of different algorithms for “Image 15”

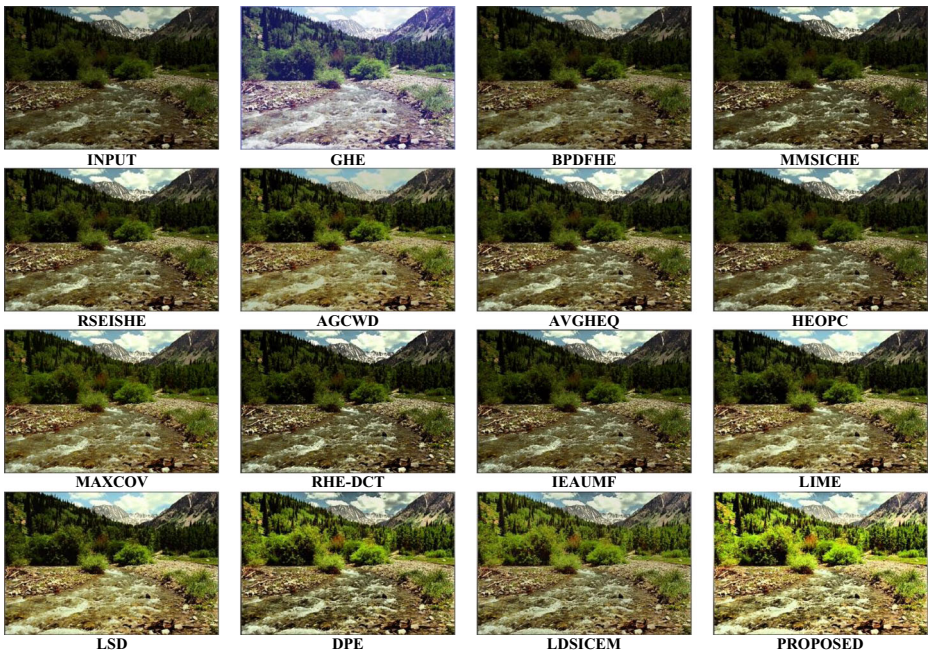


Fig. 18 Quality enhanced results of different algorithms for “Image 16”



Fig. 19 Quality enhanced results of different algorithms for “Image 17”



Fig. 20 Quality enhanced results of different algorithms for “Image 18”

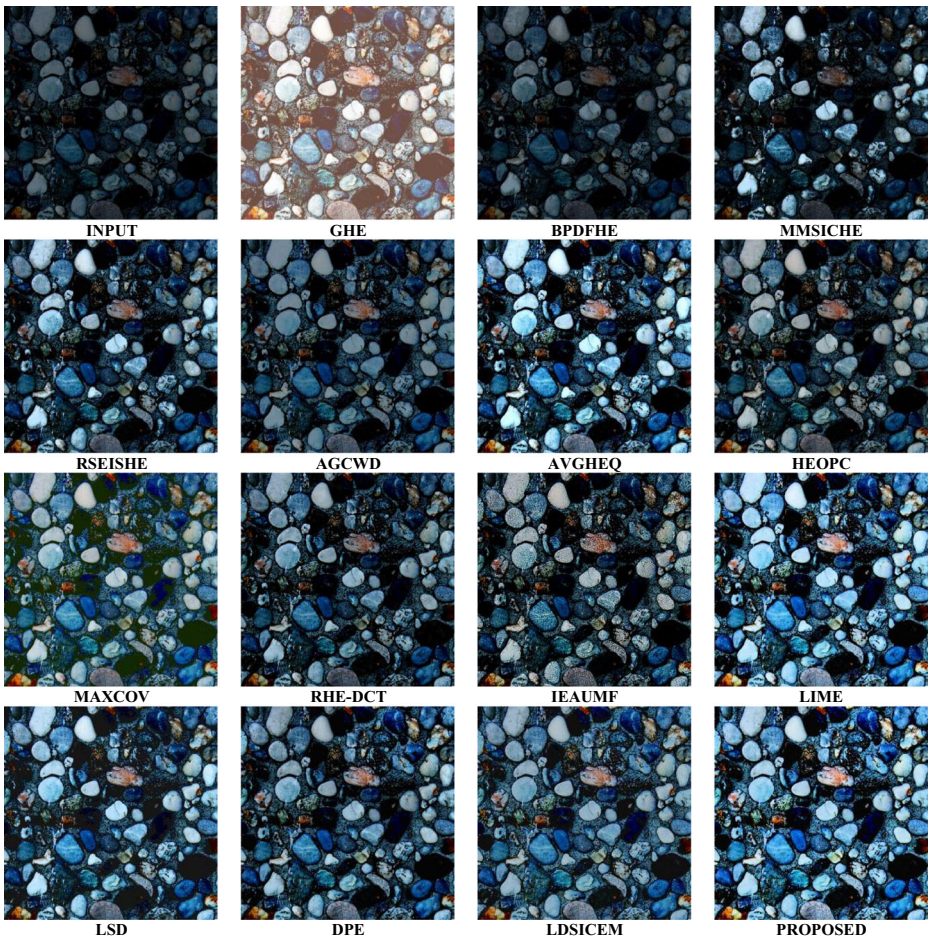


Fig. 21 Quality enhanced results of different algorithms for “Image 19”

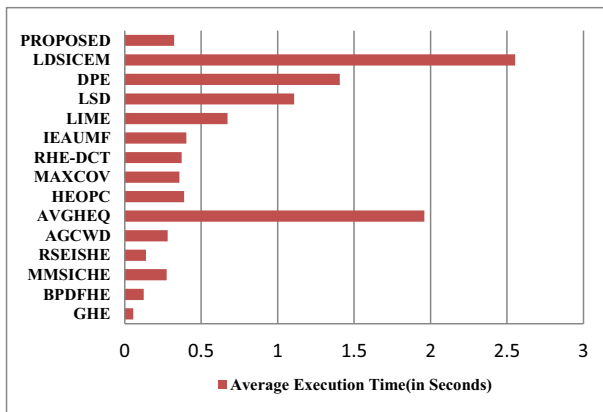


Fig. 22 Comparative analysis for execution times

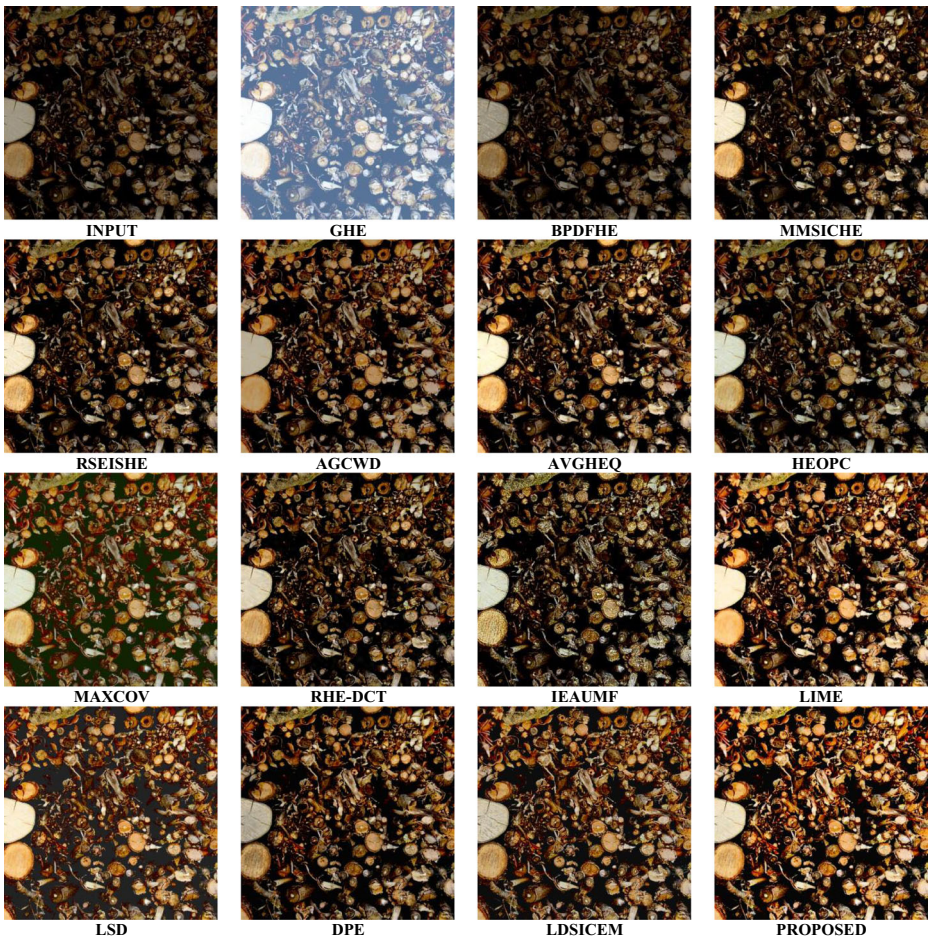


Fig. 23 Quality enhanced results of different algorithms for “Image 20”

Table 10 Average execution time (in seconds) for comparative quantitative evaluation among various algorithms

METHOD	GHE	BPDFHE	MMSICHE	RSEISHE	AGCWD
TIME (in Seconds)	0.057	0.124	0.275	0.139	0.282
METHOD	AVGHEQ	HEOPC	MAXCOV	RHE-DCT	IEAUMF
TIME (in Seconds)	1.959	0.389	0.358	0.373	0.404
METHOD	LIME	LSD	DPE	LDSICEM	PROPOSED
TIME (in Seconds)	0.673	1.109	1.407	2.553	0.324

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Cai J, Gu S, Zhang L (2018) Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans Image Process* 27(4):2049–2062
- Chen C, Chen Q, Xu J, Koltun V (2018) Learning to see in the dark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3291–3300
- Chen YS, Wang YC, Kao MH, Chuang YY (2018) Deep photo enhancer: unpaired learning for image enhancement from photographs with gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 6306–6314
- Fu X, Wang J, Zeng D, Huang Y, Ding X (2015) Remote sensing image enhancement using regularized-histogram equalization and DCT. *IEEE Geosci Remote Sens Lett* 12(11):2301–2305
- Gonzalez RC, Woods RE (2017) *Digital image processing*, 4th edn. Pearson/Prentice-Hall, New York
- Guo X, Li Y, Ling H (2017) LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans Image Process* 26(2):982–993
- Huang SC, Cheng FC, Chiu YS (2013) Efficient Contrast Enhancement Using Adaptive Gamma Correction with Weighting Distribution. *IEEE Trans Image Process* 22(3):1032–1041
- Huang SC, Yeh CH (2013) Image contrast enhancement for preserving mean brightness without losing image features. *Journal of Engg Applications of Artificial Intelligence* 26(5):1487–1492
- Kodak Lossless True Color Image Suite. <http://r0k.us/graphics/kodak/>. Accessed 02 June 2017
- Lin SCF, Wong CY, Jiang G, Rahman MA, Ren TR, Kwok N, Shi H, Yu YH, Wu T (2016) Intensity and edge based adaptive unsharp masking filter for color image enhancement. *Optik-Int J Light Electron Optics* 127(1):407–414
- Lin SCF, Wong CY, Rahman MA, Jiang G, Liu S, Kwok N, Shi H, Yu YH, Wu T (2015) Image enhancement using the averaging histogram equalization (AVHEQ) approach for contrast improvement and brightness Preservation. *Comput Electr* 46:356–370
- NASA Visible Earth. <https://visibleearth.nasa.gov>. Accessed 02 June 2017
- Pléiades Satellite Image. <https://intelligence-airbusds.com>. Accessed 02 June 2017
- Satellite Imagery and Geospatial Services | SATPALDA. <https://satpalda.com>. Accessed 02 June 2017
- Sheet D, Garud H, Suveer A, Mahadevappa M, Chatterjee J (2010) Brightness preserving dynamic fuzzy histogram equalization. *IEEE Trans Consum Electron* 56(4):2475–2480
- Singh H, Agrawal N, Kumar A, Singh GK, & Lee HN (2016) A novel gamma correction approach using optimally clipped sub-equalization for dark image enhancement. 21 *IEEE International Conference on Digital Signal Processing (DSP)*, Beijing, pp 497–501. <https://doi.org/10.1109/ICDSP.2016.7868607>
- Singh K, Kapoor R (2014) Image enhancement using exposure based sub image histogram equalization. *Pattern Recogn Lett* 36:10–14
- Singh K, Kapoor R (2014) Image enhancement via median-mean based sub-image-clipped histogram equalization. *Optik-Int J Light Electron Optics* 125(17):4646–4651
- Singh K, Kapoor R, Sinha SK (2015) Enhancement of low exposure images via recursive histogram equalization algorithms. *Optik* 126:2619–2625
- Singh H, Kumar A (2016) Satellite image enhancement using beta wavelet based gamma corrected adaptive knee transformation. 5th *IEEE International Conference on Communication and Signal Processing (ICCSP)*, Melmaruvathur, pp 128–132
- Singh H, Kumar A, Balyan LK, Singh GK (2018) Swarm intelligence optimized piecewise gamma corrected histogram equalization for dark image enhancement. *Comput Electr Eng* 70:462–475
- Singh, H., Kumar, A., & Balyan, L. K. (2017). Cuckoo search optimizer based piecewise gamma corrected auto-clipped tile-wise equalization for satellite image enhancement. In 14th *IEEE India Council International Conference (INDICON)*, Roorkee, India, 2017, pp 1–6. <https://doi.org/10.1109/INDICON.2017.8487901>
- Singh H, Kumar A, Balyan LK (2017) A levy flight firefly optimizer based piecewise gamma corrected unsharp masking framework for satellite image enhancement. In 14th *IEEE India Council International Conference (INDICON)*, Roorkee, India, 2017, pp 1–5. <https://doi.org/10.1109/INDICON.2017.8487501>
- Singh H, Kumar A, Balyan LK, Singh GK (2017) A novel optimally weighted framework of piecewise gamma corrected fractional order masking for satellite image enhancement. *Computers and Electrical Engineering*, (in press): 1–17. <https://doi.org/10.1016/j.compeleceng.2017.11.014>
- Singh H, Kumar A, Balyan LK, Singh GK (2017) A novel optimally gamma corrected intensity span maximization approach for dark image enhancement. In 22nd *IEEE International Conference on Digital Signal Processing (DSP)* 2017 (pp. 1–5). <https://doi.org/10.1109/ICDSP.2017.8096035>
- Singh H, Kumar A, Balyan LK, Lee HN (2018) Piecewise gamma corrected optimally framed Grunwald-Letnikov fractional differential masking for satellite image enhancement. In 7th *IEEE International*

- Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2018, pp 0129–0133. <https://doi.org/10.1109/ICCSP.2018.8524564>
27. Singh H, Kumar A, Balyan LK, Lee HN (2018) Fuzzified histogram equalization based gamma corrected cosine transformed energy redistribution for image enhancement. In 23rd IEEE International Conference on Digital Signal Processing (DSP), Shanghai, China, 2018, pp 1–5. <https://doi.org/10.1109/ICDSP.2018.8631612>
 28. Singh H, Kumar A, Balyan LK, Singh GK (2018) Slantlet filter-bank-based satellite image enhancement using gamma-corrected knee transformation. *Int J Electron* 105(10):1695–1715. <https://doi.org/10.1080/00207217.2018.1477199>
 29. Singh H, Kumar A, Balyan LK (2019) A sine-cosine optimizer-based gamma corrected adaptive fractional differential masking for satellite image enhancement. In *Harmony Search and Nature Inspired Optimization Algorithms*. *Advances in Intelligent Systems and Computing*, vol 741, pp 633–645 Springer, Singapore. https://doi.org/10.1007/978-981-13-0761-4_61.
 30. Wong CY, Jiang G, Rahman MA, Liu S, Lin SCF, Kwok N, Shi H, Yu YH, Wu T (2016) Histogram equalization and optimal profile compression based approach for colour image enhancement. *J Visual Commun and Image Represen* 38:802–813
 31. Wong CY, Liu S, Liu SC, Rahman MA, Lin SCF, Jiang G, Kwok N, Shi H (2016) Image contrast enhancement using histogram equalization with maximum intensity coverage. *J Mod Opt* 63(16):1618–1629



Himanshu Singh received the B. E. (Hons.) in Electronics and Communication Engineering from MITS Gwalior, India in 2010. He is currently pursuing Ph.D. (Dual Degree) in Electronics and Communication Engineering at Indian Institute of Information Technology Design and Manufacturing (IIITDM) Jabalpur, India. His research interests include satellite image enhancement, image segmentation, image denoising, image compression and optimization techniques.



Anil Kumar is an assistant professor in discipline of Electronics and Communication Engineering IIITDM, Jabalpur, India. He did his B.E. from Army Institute of Technology, Pune, India in Electronics and Telecommunication Engineering, and M.Tech. and Ph.D. from IIT Roorkee, India in 2002, 2006 and 2010, respectively. His research interests include Digital Filters, Multirate Filter Bank Designing, Signal and Image Processing.



L. K. Balyan is an assistant professor of mathematics in the discipline of Natural Science at IIITDM, Jabalpur. He did his Master of Science in applied mathematics from Indian Institute of Technology (IIT), Roorkee and Ph.D in applied mathematics from IIT Kanpur in 2009. His research interests include numerical solution of partial differential equations and spectral methods and their applications.



H. N. Lee received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He was with the HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member from 1999 to 2002. From 2002 to 2008, he was an Assistant Professor with the University of Pittsburgh, PA, USA. In 2009, he then moved to the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea, where he is currently affiliated. His areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain–computer interface.

Affiliations

Himanshu Singh¹ · Anil Kumar¹ · L. K. Balyan¹ · H. N. Lee²

¹ Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur, India

² Gwangju Institute of Science and Technology, Gwangju, South Korea

Multimodal Sparse Representation-Based Classification Scheme for RF Fingerprinting

Kiwon Yang, Jusung Kang, Jehyuk Jang, and Heung-No Lee, *Senior Member, IEEE*

Abstract—In this letter, we propose a multimodal method for improving radio frequency (RF) fingerprinting performance that uses multiple features cultivated from RF signals. Combining multiple features, including a falling transient feature that has not previously been used in RF fingerprinting studies, we aim to demonstrate that the proposed method results in improved accuracy. We show that a sparse representation-based classification (SRC) scheme can be a good platform for combining multiple features. Experimental results on RF signals acquired from eight walkie-talkies show that the RF fingerprinting accuracy of the proposed method improves significantly as the number of features increases.

Index Terms—Classification algorithm, Feature extraction, Multimodality, RF fingerprinting, Radio frequency identification

I. INTRODUCTION

CLASSIFYING radio frequency (RF) signals is useful in electronic warfare to identify the radio transmission signals of adversaries [1]. For the classification to work well, the availability of good features and a simple but robust technique are essential. A feature is a sample vector cultivated from the transmitted RF signals and bears unique information about the pertinent device. The identification of RF transmitters using such features is called RF fingerprinting. Features are known to arise from many sources, including tiny difference in device fabrication process and electronic components [1].

RF fingerprinting has attracted significant attention [2–6]: Patel *et al.* [2] used RF-DNA features which contain information on variance, skewness, and kurtosis, within a preamble response and showed that an ensemble method combining multiple classifiers performs well. Peng *et al.* [3] used four features—differential constellation trace figure, carrier frequency offset, modulation offset, and I/Q offset where classifications were done by calculating the minimum distance between test data and training data. Jia *et al.* [4] used the mean of the instantaneous amplitude of the received signal and the modulation symbol. They found an optimal dimension-reduced matrix that maximizes the quadratic mutual information between the low-dimensional features and the class, and minimizes the classification error. The same authors also

investigated an RF fingerprinting scheme based on the low-rank representation of the original data with the robust classifier parameter [5]. Merchant *et al.* [6] used a convolutional neural network (CNN) for seven commercial Zigbee devices. They collected 1,000 data per class. In [2–6], RF fingerprinting schemes with multiple features, which exhibited a high accuracy rate, were proposed for Zigbee devices and satellite terminals.

The contributions and novelties of this letter are as follows:

- We propose a new RF fingerprinting algorithm and a set of three RF features—*rising transient*, *falling transient*, and *sync*—and show the possibility that each feature can provide unique information through a real-life experiment. The falling transient feature has never been used in RF fingerprinting studies. Our results indicate that the performance of RF fingerprinting improves as each feature is additionally employed.
- Even though SRC is a common algorithm in classification [7], there are no studies on RF fingerprinting with a combination of SRC and multiple features. We show that SRC can be a good platform for RF fingerprinting.

The remainder of this letter is organized as follows. The experimental system is described in Section II. The proposed method is outlined in Section III. Results are presented and analyzed in Section IV. The conclusion is given in Section V.

II. EXPERIMENTAL SYSTEM

A. Walkie-talkie Signals

Our RF signals follow the digital mobile radio (DMR) standard. The DMR standard follows time-division multiple access (TDMA) and 4-level frequency-shift keying modulation [8]. A signal burst appears for 30 ms and disappears for 30 ms using the 2-slot TDMA method. This pattern is repeated in transmission.

The signal burst consists of rising transient, falling transient, and steady-state signals. The rising transient signal grows from zero to the designed level of the RF signal. Contrary to the rising transient signal, the falling transient signal decreases from the designed level to zero. The steady-state signal refers to the resting part between the rising and the falling transient signal. The steady-state signal is composed of data and a sync signal. The data have 216 bits and the sync signal has 48 bits. The bit rate of the DMR standard is 9,600 bits/s. The sync signal is used to synchronize between a transmitter and receiver.

Manuscript received February 28, 2019. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2018R1A2A1A19018665)

The authors are with the School of Electrical Engineering and Computer Science (EECS), Gwangju Institute of Science and Technology (GIST), South Korea (e-mail: heungno@gist.ac.kr).

From the pertinent signal part, a feature is obtained. Each feature is the main lobe of a spectrum of the pertinent signal part. We show how to extract each feature from the pertinent signal part in Section III. A. As we mentioned, the features are the result of the inherent nonlinear properties of radio transmitters in the manufacturing process [1]. Owing to the presence of such features, RF fingerprinting can be accomplished.

B. Signal-Acquisition Setup

For our experiment, two walkie-talkie models were used: the Motorola *SIIM* and Hytera *BD-358*. Each model follows the DMR standard. Four units of each type, eight walkie-talkies in total, were used in the experiment.

Signal was transmitted from the transmitter and acquired from an SMA male mini car-mounted antenna, the receiving frequency band of which is 400–470 MHz. We then down-converted 423.1875 MHz to 10 MHz using an XL-11-411 RF mixer and an E4438C ESG vector signal generator. Then, we filtered the signal bandwidth and sampled the signal using an IF recording system with the PX14400 operator functioning as a low-pass filter and analog-to-digital converter. Signals sampled at 96 MHz were saved to a computer and loaded to MATLAB. As we captured 50 signals per walkie-talkie, 400 signals were saved to the computer.

III. PROPOSED FEATURE EXTRACTION AND CLASSIFICATION

A. Signal Burst to Features

To cultivate features, we extracted the three signal parts from a single signal burst. Then, each feature was selected from the pertinent extracted signal parts.

Each signal part is extracted out from the first signal burst of the total received signal through time-windowing. To design the time-window for each signal, we used a thresholding method. For a rising transient signal $\mathbf{f}_R \in \mathbb{R}^{500,000 \times 1}$, the starting threshold is the first time point at which the amplitude of the signal burst exceeds 10% of its maximum; the ending threshold is the time point it exceeds 90%, respectively. Similarly, for the falling transient signal $\mathbf{f}_F \in \mathbb{R}^{500,000 \times 1}$, the starting threshold and the ending threshold are the latest time points at which the amplitude of the signal burst exceeds 90% and 10% of its maximum, respectively. Since the length of each transient signal fluctuates, we used zero padding method after the ending point of each transient signal to match the length. To design the time window for the sync signal, we referred to the DMR standard [8]. The sync signal $\mathbf{f}_S \in \mathbb{R}^{480,000 \times 1}$ is located at the center of a signal burst. We set the center of the time-window for the sync signal to the central time point between the ending time of the rising transient signal and the starting time of the falling transient signal. The width of the time window was set to 0.005 s, as per the DMR standard [8].

The extracted signal parts are transformed to the spectrum domain by fast Fourier transform (FFT) with the size of the time signal. Then, the operation of taking the absolute value of each element is executed to compare energy and frequency

information of the extracted signal part with those of the others. Since the main lobe occupies most of the energy of each signal part, the main lobe is taken from each spectrum. The main lobes, $F(\mathbf{f}_R)_{ML} \in \mathbb{R}^{2,000 \times 1}$, $F(\mathbf{f}_F)_{ML} \in \mathbb{R}^{2,000 \times 1}$, and $F(\mathbf{f}_S)_{ML} \in \mathbb{R}^{1,920 \times 1}$, are the unique features used for RF fingerprinting in our experiment, where $F(\cdot)$ is the FFT operation function and ML means the main lobe of the spectrum.

To extract the main lobe, we used a bandpass filter first. Then, the center frequency of the filtered spectrum was down-converted to zero. Finally, we decimated the signal to reduce the length of the sample sequence. The main lobe was set to occupy the most of energy of each signal part, considering the channel bandwidth.

B. Proposed SRC

SRC is a classification algorithm based on the compressed sensing theory [7] and is used to determine the class from the sparse solution of the representation equation

$$\mathbf{y} = \mathbf{D}\mathbf{s}, \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^{P \times 1}$ is a test data vector, $\mathbf{D} \in \mathbb{R}^{P \times NL}$ is a training data matrix composed of N training data vectors for each class label $l \in \{1, \dots, L\}$, and the column vector $\mathbf{s} \in \mathbb{R}^{NL \times 1}$ is the vector of sparse representation coefficients. The sparse signal recovery algorithm in [7] was used to solve (1) with $P < NL$. The obtained sparse unique solution \mathbf{s} is divided into L disjoint subvectors $\mathbf{s}^{(l)}$, for $l = 1, \dots, L$. Specifically, $\mathbf{s} = [(\mathbf{s}^{(1)})^T, (\mathbf{s}^{(2)})^T, \dots, (\mathbf{s}^{(L)})^T]^T$, where T is the transpose. Likewise, we divide \mathbf{D} into L submatrices corresponding to $\mathbf{s}^{(l)}$. To identify the class of test data, we solve

$$\text{class} = \arg \min_{l \in \{1, \dots, L\}} \left\| \mathbf{y} - \mathbf{D}^{(l)} \mathbf{s}^{(l)} \right\|_2. \quad (2)$$

If the column vectors in \mathbf{D} are less correlated, the solution \mathbf{s} of (1) is approximated to be sparse since the condition of having a sparse solution depends on the mutual correlation between the columns of \mathbf{D} [9]. Thus, the compressed sensing algorithms in [7] can be used to find a unique solution \mathbf{s} . However, when RF signals are taken directly to form the column vectors of \mathbf{D} , the solution \mathbf{s} cannot be sparse because they may be highly correlated. Then, the performance of SRC may be poor. Thus, RF signals must be processed to remove correlation to obtain high performance in SRC [9].

To remove correlation among RF signals, the proposed method applies principal components analysis (PCA) to the column vectors, each of which combines multiple features. PCA is known to be good at geometrically separating the features in the Euclidean domain, removing the mutual correlation [10]. This section aims to show that how three kinds of features are concatenated and how PCA is applied to features. For clarity in explanation, we first introduce the single modal method and move on to the proposed multimodal method.

1) *Single modal RF fingerprinting*: Consider that one of the rising transient, falling transient, and sync features is used as

the sole representative feature of a single transmitter. We first form a feature matrix in which the columns are the sample vectors of the feature of candidate RF transmitters. Mathematically, for L RF transmitters (classes) and N sample vectors of a feature of each RF transmitter, we construct feature matrix $\mathbf{A} \in \mathbb{R}^{M \times NL}$ as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{(1)}, \dots, \mathbf{a}_N^{(1)}, \mathbf{a}_1^{(2)}, \dots, \mathbf{a}_N^{(L)} \end{bmatrix}, \quad (3)$$

where column vector $\mathbf{a}_n^{(l)} \in \mathbb{R}^{M \times 1}$ is the n^{th} sample vector of a feature of the l^{th} RF transmitter for $n=1, \dots, N$ and $l=1, \dots, L$, and M is the length of $\mathbf{a}_n^{(l)}$. We denote a sample vector of a feature of an unknown RF transmitter as \mathbf{u} . From the PCA operation, (6) and (8), \mathbf{A} and \mathbf{u} are changed to a training data matrix \mathbf{D} and a test vector \mathbf{y} , respectively.

2) *Multimodal RF fingerprinting*: The proposed method is to concatenate the multiple features in the representation equation $\mathbf{u} = \mathbf{A}\mathbf{s}$, as shown in Fig. 1; the feature matrices are concatenated in a row-wise manner. On the columns of this combined matrix, the PCA is applied. Let us denote the n^{th} sample vector of the k^{th} feature of the l^{th} RF transmitter for $n=1, \dots, N$, $l=1, \dots, L$, and $k=1, \dots, K$ by $\mathbf{a}_{k,n}^{(l)} \in \mathbb{R}^{M \times 1}$, where K is the number of features to be combined. The feature matrices are concatenated as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1^T & \mathbf{A}_2^T & \dots & \mathbf{A}_K^T \end{bmatrix}^T, \quad (4)$$

where $\mathbf{a}_{k,n}^{(l)}$ forms the columns of feature matrix $\mathbf{A}_k \in \mathbb{R}^{M \times NL}$,

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{a}_{k,1}^{(1)}, \dots, \mathbf{a}_{k,N}^{(1)}, \mathbf{a}_{k,1}^{(2)}, \dots, \mathbf{a}_{k,N}^{(L)} \end{bmatrix}. \quad (5)$$

We obtain the training data matrix \mathbf{D} as follows:

$$\mathbf{D} = \mathbf{V}^T (\mathbf{A} - \mathbf{m}\mathbf{1}), \quad (6)$$

where $\mathbf{m} = \frac{1}{L \times N} \sum_{l=1}^L \sum_{n=1}^N \mathbf{a}_n^l \in \mathbb{R}^{M \times 1}$ is an average vector of columns of $\mathbf{A} \in \mathbb{R}^{M \times NL}$, $\mathbf{1} := [1 \ 1 \ \dots \ 1]$ is the 1 by NL vector of 1 s, $\mathbf{a}_n^l = \begin{bmatrix} (\mathbf{a}_{1,n}^l)^T & (\mathbf{a}_{2,n}^l)^T & \dots & (\mathbf{a}_{K,n}^l)^T \end{bmatrix}^T \in \mathbb{R}^{M \times 1}$ is a column vector which combines K features, and $\mathbf{V} \in \mathbb{R}^{M \times P}$ is a rearranged eigenvector matrix of the covariance matrix $(\mathbf{A} - \mathbf{m}\mathbf{1})(\mathbf{A} - \mathbf{m}\mathbf{1})^T \in \mathbb{R}^{M \times M}$. The eigenvectors of

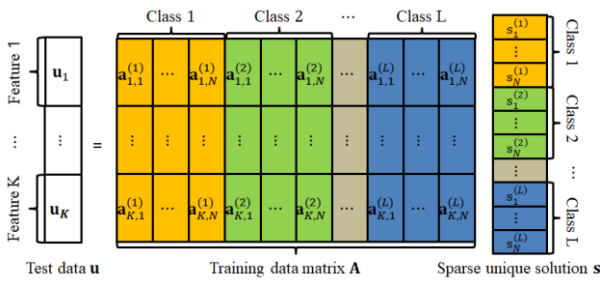


Fig. 1. Feature concatenation in the proposed method

$(\mathbf{A} - \mathbf{m}\mathbf{1})(\mathbf{A} - \mathbf{m}\mathbf{1})^T$ are arranged according to the eigenvalues

in descending order. Since the eigenvalue of the covariance matrix is proportional to the variance of the columns of \mathbf{A} and the eigenvectors of the covariance matrices are orthonormal, the column vector of \mathbf{V} becomes a basis of the new space on the variance of the columns of \mathbf{A} [10]. The dimension of \mathbf{V} can be selected by user as $P \in \{1, \dots, MK\}$. To obtain the test data vector \mathbf{y} of SRC, we first concatenate the sample vectors of different features of an unknown transmitter \mathbf{u}_k as follows:

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1^T & \mathbf{u}_2^T & \dots & \mathbf{u}_K^T \end{bmatrix}^T. \quad (7)$$

Finally, \mathbf{y} is obtained by mapping the difference between concatenated vector $\mathbf{u} \in \mathbb{R}^{MK \times 1}$ and \mathbf{m} , i.e., $\mathbf{u} - \mathbf{m}$, onto the space with the eigenvector matrix \mathbf{V} ,

$$\mathbf{y} = \mathbf{V}^T (\mathbf{u} - \mathbf{m}). \quad (8)$$

By using PCA, the equation in Fig. 1 is changed to (1), which has principal components as training and test data. The SRC solution in (1) was determined using the basis pursuit algorithm, which finds the unique sparse solution that has the minimum L1 norm [11].

IV. EXPERIMENTAL RESULT AND DISCUSSION

For our experiment, we set the decimation rate for all feature extractions to 250, considering the bandwidth of the RF signal following the DMR standard [8] and the sampling rate. To evaluate the performance of the proposed classifier, we used a five-fold cross validation technique. Fifty data were used per walkie-talkie, such that each test data was classified on the total of 320 training data. The experiment was performed in a line-of-sight environment. SNR was around 35-40 dB.

Table I shows the accuracy rate of the proposed method. The accuracy rate of the multimodal scheme is much better than that obtained from using only one feature. The minimum number of principal components is the minimum number of column vectors in the eigenvector matrix \mathbf{V} that yields the highest accuracy rate. Fig. 2 shows five results of classification using from 1 to 100 principal components, 1) on the rising transient feature, 2) on the falling transient feature, 3) on both transient

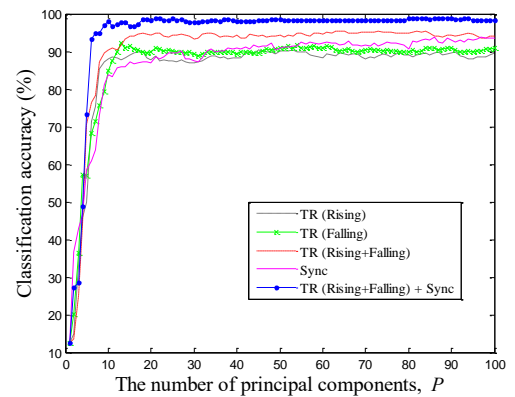
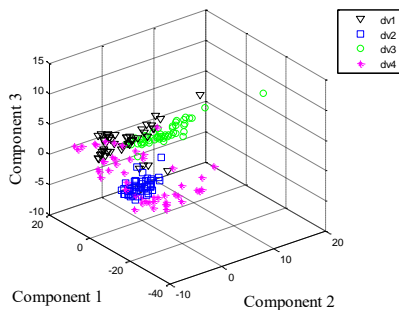


Fig. 2. Classification result of four BD-358 and four SL1M walkie-talkies

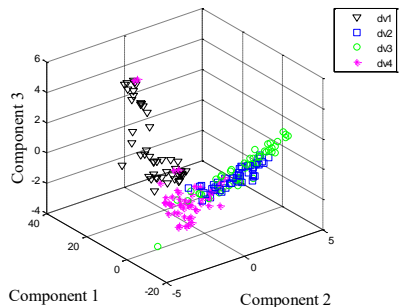
features combined, 4) on the sync feature, and 5) on all features combined. Improved performance with an increased number of

features indicates that each feature could contain unique information. Since the eigenvectors of the 21st and higher eigenvalues of $(\mathbf{A} - \mathbf{m}\mathbf{1})(\mathbf{A} - \mathbf{m}\mathbf{1})^T$ do not have enough information to represent the differences of data, the classification accuracy is not changed as $P > 21$. Fig. 3 shows a feature map in which the principal components of features of the Motorola *SLIM* are mapped onto a 3D plot. The label of each axis, such as Component 1 and Component 2, means the projection of $\mathbf{u} - \mathbf{m}$ to the P^{th} column vector of \mathbf{V} . The figure shows distinct cluster formation when a concatenated feature is used.

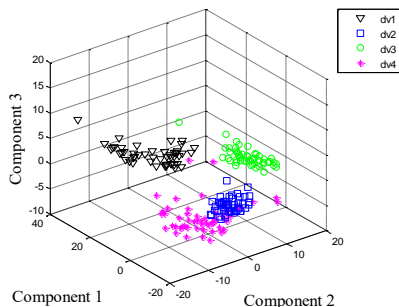
To compare the proposed method with convolutional neural network (CNN), we referred to the study [6]. For additional tests, we used a five-fold cross validation technique. Five training neural networks were constructed using the same training dataset with the proposed method. We used the concatenated features as input data. The average classification accuracy rate of the CNN was 90.75% which can be compared with 98.75% of our method. This comparison shows that SRC



(a) SLIM transient (rising + falling) features



(b) SLIM sync features



(c) SLIM transient (rising + falling) + sync features

Fig. 3. Feature map of a 3-D principal components space can perform RF fingerprinting well with fewer training data.

Table I. Accuracy rate of the proposed method

	4 BD-358	4 SL1M	4 BD-358 4 SL1M
	Accuracy rate (Minimum number of PC)		
TR(R)	88% (24)	82% (48)	90.5% (45)
TR(F)	87.5% (45)	90% (12)	92.25% (13)
TR(R + F)	93% (49)	92% (20)	95.5% (63)
Sync	99% (45)	83.5% (22)	93.75% (86)
TR(R + F) + Sync	99% (44)	98.5% (22)	98.75% (21)

R: Rising, F: Falling, PC: Principal components

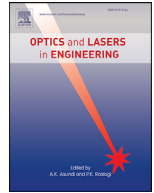
Because it is simple to add more training data and different kinds of features in SRC, the performance of the proposed method could be improved with additional training data and other kinds of features.

V. CONCLUSIONS

This letter proposed a multimodal RF fingerprinting scheme based on SRC. We showed that the proposed multimodal scheme, which concatenates multiple features in the row-wise manner and applies PCA to the concatenated dictionary matrix, improves accuracy significantly. We showed the possibility that the three signal features we have cultivated from RF signal samples could provide mutually independent information. The proposed scheme is efficient in the sense that improved RF fingerprinting accuracy is obtained. In addition, it is simple and easy in the proposed scheme to add more data and various kinds of features. The MATLAB source code for this study can be obtained at “Lab homepage address will be written”.

REFERENCES

- [1] Y. Jia, S. Zhu, and L. Gan, “Specific emitter identification based on the natural measure,” *Entropy*, vol. 19, pp. 117, 2017.
- [2] H. J. Patel, M. A. Temple, and R. O. Baldwin, “Improving ZigBee device network authentication using ensemble decision tree classifiers with radio frequency distinct native attribute fingerprinting,” *IEEE Trans. Rel.*, vol. 64, no. 1, pp. 221–233, Mar. 2015.
- [3] L. Peng, A. Hu, J. Zhang, Y. Jiang, J. Yu, and Y. Yan, “Design of a Hybrid RF Fingerprint Extraction and Device Classification Scheme,” *IEEE Trans. Internet Things.*, May, 2018 (in press)
- [4] Y. Jia, J. Ma, L. Gan, “Combined Optimization of Feature Reduction and Classification for Radiometric Identification,” *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 584–588, May, 2017.
- [5] Y. Jia, J. Ma, L. Gan, “Radiometric Identification Based on Low-Rank Representation and Minimum Prediction Error Regularization,” *IEEE Comm. Lett.*, vol. 21, no. 8, pp. 1847–1850, Aug, 2017.
- [6] K. Merchant, S. Revay, G. Stantchev, and B. Nousain, “Deep learning for Rf device fingerprinting in cognitive communication network,” *IEEE, J.Sel.Topics Signal Process.*, vol. 12, no. 1, pp. 160–167, Feb. 2018.
- [7] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 210–227, Feb. 2009.
- [8] ETSI TS 102 361-1 v2.4.1. “Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) Systems; Part 1: DMR Air Interface (AI) protocol,” European Telecommunications Standards Institute, 2016.
- [9] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.
- [10] H. Abdi and L. Williams, “Principal component analysis,” *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
- [11] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.



Fabrication of 2D thin-film filter-array for compressive sensing spectroscopy

Cheolsun Kim, Woong-Bi Lee, Soo Kyung Lee, Yong Tak Lee, Heung-No Lee*

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

ARTICLE INFO

Keywords:

Spectroscopy
Thin films
Inverse problems
Compressive sensing

ABSTRACT

We demonstrate 2D filter-array compressive sensing spectroscopy based on thin-film technology and a compressive sensing reconstruction algorithm. To obtain different spectral modulations, we fabricate a set of multilayer filters using alternating low- and high-index materials and reconstruct the input spectrum using a small number of measurements. Experimental results show that the fabricated filter-array provides compatible spectral resolution performance with a conventional spectrometer in monochromatic lights and LEDs. In addition, the fabricated filter-array covers a wide range of wavelengths with a single exposure.

1. Introduction

The demand for spectrum information is increasing not only in research and development but also in the private sector. In response to this demand, researchers are trying to make spectrometers that are both small and inexpensive. These spectrometers could be used in various fields, such as medical systems, mobile applications, and remote sensing [1–3]. In particular, optical filter-based spectrometers do not need motorized or dispersive elements, and their filter-array can be directly attached to the detectors so that they can be easily miniaturized. However, there is a trade-off between size (for integrating filters) and spectral resolution with miniaturized spectrometers.

Over the years, numerous approaches to applying compressive sensing (CS) techniques have been proposed to reduce the size of spectrometers without reducing spectral resolution, or potentially even improving it. These approaches [4–7] include the following: band pass filters [4], random transmittance filters [5], photonic crystal slabs [6], and liquid crystal phase retarders [7]. Recently, Fabry–Perot (FP)-based CS spectroscopy methods have been presented [8,9]. To acquire differently modulated spectral measurements, a 2D array of FP resonators with different cavity depths has been tried [8] as well as a piezo-actuated device that changes the distance between two FP mirrors has been tried [9]. A hundred FP resonators are used to recover the input spectrum in [8], and the operational range of the piezo-actuator imposes mechanical limitations in [9].

The CS framework [10–12] is an efficient sampling and reconstruction scheme that requires fewer samples to reconstruct the signal than that required by conventional sampling. The CS framework can be applied to filter-based spectroscopy, offering the advantage of reducing

the number of filters and detectors required and allowing the system to be miniaturized.

In spectroscopy, the relation between the spectral components of the input light source $\mathbf{x} \in \mathbb{R}^{N \times 1}$ and the modulated signal $\mathbf{y} \in \mathbb{R}^{M \times 1}$ can be expressed as follows:

$$\mathbf{y} = \mathbf{T}\mathbf{x}, \quad (1)$$

where $\mathbf{T} \in \mathbb{R}^{M \times N}$ is the sensing matrix. Each row of the sensing matrix is to represent the transmission function (TF) of i -th filter, $T_m \in \mathbb{R}^{1 \times N}$ for $m = 1, 2, \dots, M$. In order to achieve miniaturization of spectroscope without degradation of spectral resolution, the CS framework is utilized in spectroscopy, where the number of filters is set to be smaller than the number of spectral components ($M < N$). Then, Eq. (1) becomes an underdetermined linear system. A sparse signal reconstruction algorithm with $L1$ norm minimization can be used to solve Eq. (1), if the input spectrum is either naturally sparse or can be sparsely represented in some basis $\Phi \in \mathbb{R}^{N \times N}$, i.e., $\mathbf{x} = \Phi\mathbf{s}$, where $\mathbf{s} \in \mathbb{R}^{N \times 1}$ is a sparse vector. Then, Eq. (1) becomes

$$\mathbf{y} = \mathbf{T}\Phi\mathbf{s} \quad (2)$$

The sparse signal \mathbf{s} can be estimated by solving the following $L1$ norm minimization problem:

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{T}\Phi\mathbf{s}\|_2 \leq \epsilon \quad (3)$$

where ϵ is a small non-negative constant. The reconstructed input spectrum $\hat{\mathbf{x}}$ is then $\Phi\hat{\mathbf{s}}$.

In this paper, we demonstrate 2D filter-array CS spectroscopy. This uses a multilayer thin-film filter-array for spectral modulation, where each filter modulates the input spectrum using different sensing patterns. A CMOS image camera reads out the modulated signals with a

* Corresponding author.

E-mail address: heungno@gist.ac.kr (H.-N. Lee).

Table 1
Recursion for calculating reflection coefficients.

Input: $\lambda, \theta_1 = 0, \mathbf{n} = \{n_1, n_2, \dots, n_{l-1}, n_l\}, \mathbf{d} = \{d_2, d_3, \dots, d_{l-1}, d_l\}.$
Step 1: Obtain $\theta_k, \beta_k,$ and N_k $\theta_k = \sin^{-1} \left(\frac{n_{k-1}}{n_k} \sin \theta_{k-1} \right),$ for $k = 2, 3, \dots, l.$ $\beta_k = 2\pi \cos(\theta_k) n_k d_k / \lambda,$ for $k = 2, 3, \dots, l.$ $N_k = \begin{cases} n_k / \cos \theta_k & \text{for TE} \\ n_k \cos \theta_k & \text{for TM} \end{cases},$ for $k = 1, 2, \dots, l.$
Step 2: Set $\eta_l = N_l$
Step 3: Obtain η_2 Decrement k by 1 from $l-1$ to 2 $\eta_k = N_k \frac{\eta_{k+1} \cos \beta_k + j N_{k+1} \sin \beta_k}{N_{k+1} \cos \beta_k + j \eta_{k+1} \sin \beta_k}$
return η_2
Step 4: Compute $\rho = (N_1 - \eta_2) / (N_1 + \eta_2).$
Output: ρ

single exposure, and then a reconstruction algorithm is applied that depends on the modulated signals and the sensing matrix, allowing the input spectrum to be recovered.

The research focus has been given to fabrication of the multilayer thin-film filters for actual CS spectroscopy implementation and verification experiments. For fabricating as a 2D filter-array, we use commonly available materials SiNx and SiO₂ for high and low refractive index materials which are deposited alternately on the substrate with varying thicknesses. Furthermore, we come up with a practical way that set of filters can be deposited on a single substrate with different thicknesses of layers.

2. 2D filter-array

2.1. Multilayer thin-film filter

Thin films are a basic component that have been applied in a variety of areas, including semiconductor devices, optical coatings, and solar cells [13]. The theoretical TF of a multilayer thin-film filter is given by [14]

$$T(\lambda, \theta_1) = 1 - \frac{1}{2} \left(|\rho_{TE}|^2 + |\rho_{TM}|^2 \right), \quad (4)$$

where ρ_{TE} and ρ_{TM} are the reflection coefficients. Given a wavelength λ and the incident angle θ_1 , TF can be calculated using recursive routines shown in Table 1.

In Table 1, given the input of a wavelength λ , a vector of l refractive indices $\mathbf{n} = (n_1, n_2, \dots, n_{l-1}, n_l)$ and a vector of $l-1$ layer thicknesses $\mathbf{d} = (d_2, d_3, \dots, d_{l-1}, d_l)$, a reflection coefficient ρ is generated. Note that there are l layers considered in total. The first one is the layer of the air and the last one is the layer of the substrate. The light is assumed to be arriving from the air to the second layer in normal incidence. The first index n_1 in the vector \mathbf{n} represents the refractive index of the air. The last one n_l in the vector \mathbf{n} represents the refractive index of the substrate. The refractive indices of the intermediate thin-film layers are denoted by n_2 to n_{l-1} . The thickness of the air does not need to be considered. The thickness of the substrate is denoted by d_l .

The thicknesses of the intermediate thin-film layers are denoted by d_2 to d_{l-1} . The incidence angle of the light passing from the k th to the $k+1$ th layer is θ_k , and η_k is the effective complex-valued index of the k th layer. A TF for a single filter is obtained by considering all wavelengths in the range of interest. An array of TFs for the M filters can be obtained by repeating this process where each filter $T_m \in \mathbb{R}^{1 \times N}$ for $m = 1, 2, \dots, M$ in Eq. (1) is generated from a unique set of refractive index and thickness vectors.

2.2. Numerical design of 2D filter-array

To implement the proposed 2D filter-array, we numerically modeled the proposed spectroscopy method with reference to [14–16], and according to the following steps. (i) Generate the reference vector of layer

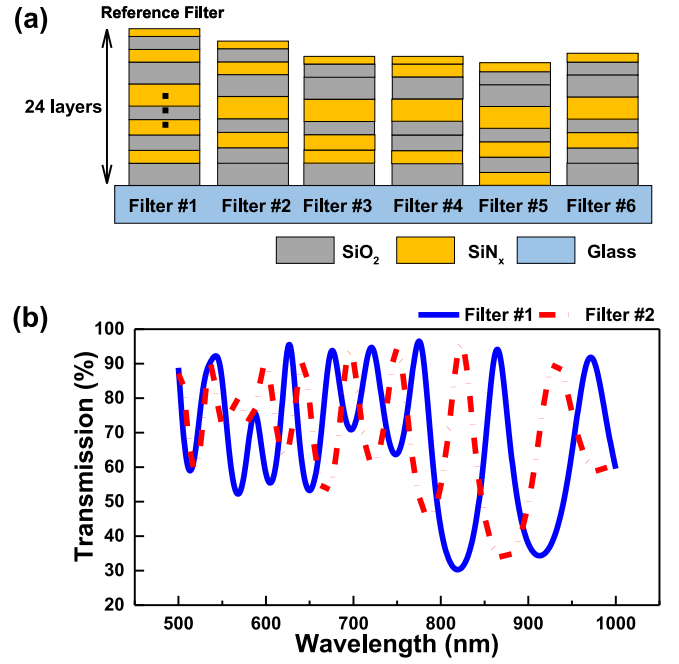


Fig. 1. (a) Schematic of the thin-film filter-array. (b) Example of two transmission functions for thin-film filters.

thicknesses, i.e. $\mathbf{d} = (d_2, d_3, \dots, d_{l-1}, d_l)$, for the reference filter. (ii) Generate a vector of thicknesses for the other filter by randomly removing one to five layer thicknesses from the reference vector. (iii) Repeat the step (ii) 35 times to create a total of 36 vectors of thicknesses. (iv) Use the recursion Table 1 and Eq. (4) to calculate the TFs for a new filter-array (sensing matrix). (v) Use the mutual coherence μ to quantify the goodness of the sensing matrix of the designed filter-array. Mutual coherence μ is defined as $\mu \triangleq \max_{i,j} |o_{ij}|$, where o_{ij} is the (i, j) th off-diagonal element of the Gram matrix, $\mathbf{T}^* \mathbf{T} \in \mathbb{R}^{N \times N}$. \mathbf{T}^* denotes the conjugate transpose of \mathbf{T} . With these steps, we can generate a single set of 36 filters. By repeating these steps, multiple sets of 36 filters can be obtained. Among these sets of filter-arrays, the set of filters with a smallest mutual coherence is selected.

In CS framework, a sensing matrix with a smaller mutual coherence is better than the one with a higher mutual coherence to capture the information of input signal to be reconstructed [5,17]. A schematic of the proposed filter-array is shown in Fig. 1(a). Each time a layer is removed, the layers above and those below come together to form a single layer with two thicknesses added up. We consider two materials, SiNx and SiO₂ for the high- and low-refractive index materials with refractive indices of 2.02 for SiNx and 1.45 for SiO₂. The thickness range of each layer is from 50 to 150 nm. Through the numerical design, we empirically found that removal of up to five layers from the 24-layer reference filter was possible to create a 6×6 filter-array with a low coherence.

Fig. 1(b) shows the TFs for two designed filters as examples. In conventional spectroscopy, the TFs with a large spectral depth and a narrow spectral peak are preferred in order to prevent interference among measurements. In compressive sensing spectroscopy, Each TF of the filter should be wide enough so that the set of the small number of filters fully senses the spectral information in the given wavelength range [9].

Each filter shows several spectral peaks and rapid changes of transmission value with respect to wavelength. Therefore, each filter has a high optical throughput that the energy (intensity) which passes through the filter is higher than that with the conventional bandpass filter approach. In addition, fewer filters can be used to cover the entire wavelength range with the proposed method. For example, suppose 250 bandpass filters are used to cover the wavelength range from 500 to

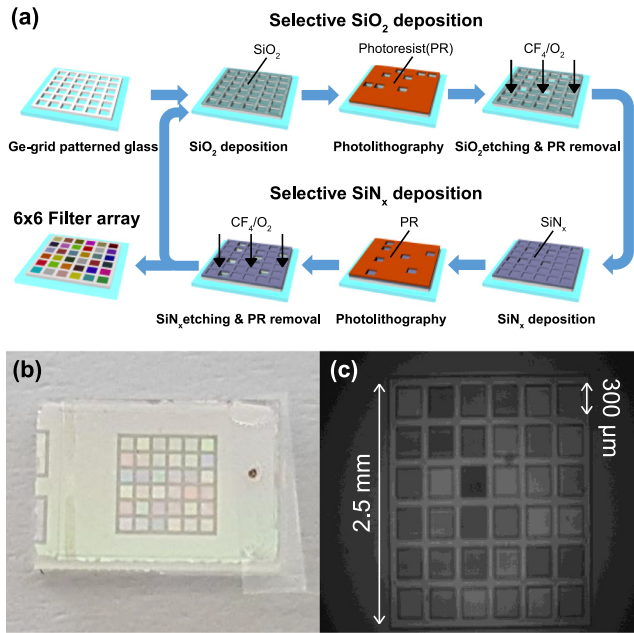


Fig. 2. (a) Schematic of the thin-film filter-array fabrication process. (b) Photograph of a fabricated thin-film filter-array. (c) Monochrome image of the thin-film filter-array taken at a wavelength of 700 nm.

1000 nm. Then, the bandwidth of TF is 2 nm, according to the conventional bandpass filter design. In the proposed approach, the same range of wavelength can be covered with only 36 proposed filters, subject to the use of a recovery algorithm present at the reconstruction end.

2.3. Filter-array fabrication

Fig. 2(a) shows the process in which a thin-film filter-array is fabricated. This comprises two main parts; one is SiO₂ film deposition and the other is SiN_x film deposition according to the specified thicknesses. Prior to depositing an SiO₂ film, a 6 × 6 germanium (Ge) grid with elements of size 300 μm and spacing 100 μm was formed on the glass using an e-beam evaporator to separate the filters. In this grid, SiO₂ and SiN_x layers were deposited with the width of 300 μm in each filter. Then, selective deposition was done as follow: An intentionally thick SiO₂ film was deposited on the glass patterned with the Ge grid using plasma-enhanced chemical vapor deposition. The regions where the film should not be deposited were then removed by conventional photolithography, namely CF₄/O₂ reactive ion etching. The process pressure and radio-frequency power were maintained at 50 mTorr and 50 W, respectively. The SiN_x film deposition process was performed in the same manner as for SiO₂. Finally, these two main steps, SiO₂ and SiN_x film deposition, were repeated 12 times each to lay down 24 layers. Fig. 2(b) and (c) show a photograph of a fabricated thin-film filter-array and a monochrome image of the filter-array, respectively. Each filter is composed of a different number of layers each with different thicknesses; therefore, each one has unique color due to its different TF, as shown in Fig. 1(b).

3. Experiments

3.1. Experimental setup

Optical setups for experimental verification of the proposed spectroscopy system are shown in Fig. 3. Fig. 3(a) depicts the optical setup for measuring TFs of a filter-array. The setup for testing the performance of the proposed system is shown in Fig. 3(b). The photographs of the optical setup and the CMOS image camera with the thin-film filter-array

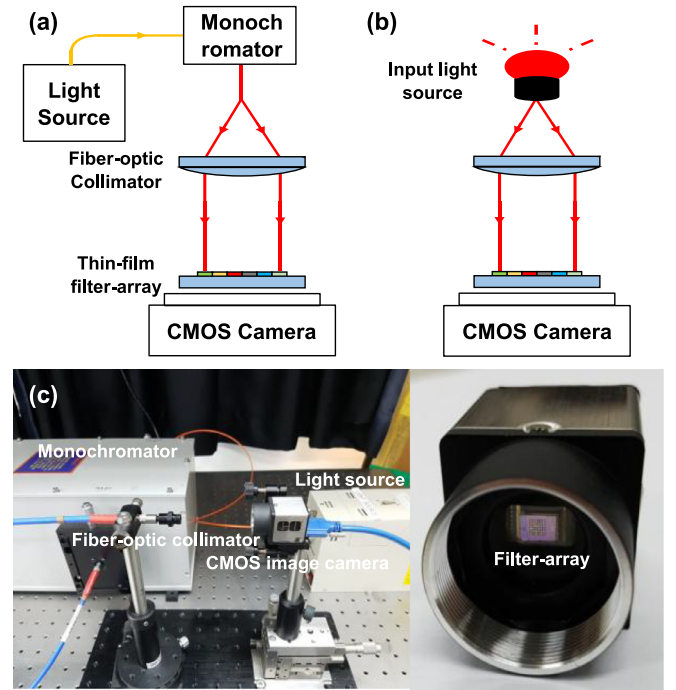


Fig. 3. (a) Schematic of the optical setup for measuring the sensing matrix. (b) Schematic of the optical setup for testing the performance of the proposed spectroscopy system. (c) Photographs of the optical setup and the CMOS image camera with the thin-film filter-array.

are shown in Fig. 3(c). During the optical experiments, we set the incident angle to filter-array as normal incidence. Using a linear stage, a rotational stage and optical mounting posts, we aligned the optical fiber with the CMOS image camera (E0-1312, Edmund Optics) for the normal incidence.

In Fig. 3(a), a halogen lamp (KLS-150H-LS-150D, Kwangwoo) was used to provide a continuous light spectrum. It was put into a monochromator (MMAC-200, Mi Optics) to produce a specific narrow wavelength band. Then, a fiber-optic collimator was used to form a beam of parallel light. The beam was fed into the CMOS image camera through the fabricated thin-film filter-array. With a single exposure, each filter modulated the light in a different pattern. The modulated light was read out by pixels of CMOS image camera, yielding $M = 36$ distinct output signals \mathbf{y} in Eq. (1). Each output signal was taken by summing up the modulated values of the pixels underneath the pertinent filter.

To apply CS reconstruction algorithms to the proposed system, the sensing matrix \mathbf{T} must be pre-determined. Let us denote the intensity which passes through the filter-array as $IF(m, \lambda)$ and the intensity without the filter-array as $IWF(m, \lambda)$, where m is the filter index and λ is the wavelength. The sensing matrix is then given by

$$T(m, \lambda) = \frac{IF(m, \lambda) - BI(m, \lambda)}{IWF(m, \lambda) - BI(m, \lambda)}, \quad (5)$$

where $BI(m, \lambda)$ is the background intensity. We took 500 wavelength samples, spaced 1 nm apart, in the range from 500 to 1000 nm. The measured sensing matrix $\mathbf{T} \in \mathbb{R}^{36 \times 500}$ obtained from the fabricated thin-film filter-array is shown in Fig. 4. Each TF of the filters, a row of the color map, is shown as a combination of colors, i.e., red (high transmission value) and blue (low transmission value). Different TFs show different places of high and low transmission values indicating mutual uncorrelation. As a set of 36 filters, the filter-array covers the entire wavelength range with high optical throughput.

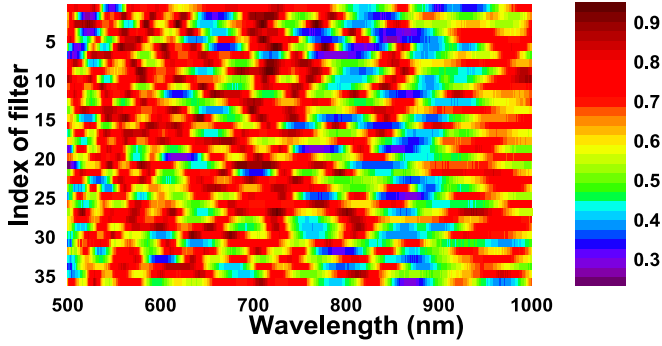


Fig. 4. Color map of the measured sensing matrix for the thin-film filter-array. Each row represents the TF of a filter with respect to wavelength.

3.2. Computational experiments

To quantify the performance and explore the two-point resolution of the fabricated filter-array, we conducted computational experiments. The two-point resolution is the ability to distinguish the spectral peaks which are closely spaced. For the experiments, we generated mono-peak spectra and two-peak spectra as input spectra using the Gaussian function. A generated input spectrum \mathbf{x} was numerically modulated by multiplying the measured sensing matrix \mathbf{T} as shown in Eq. (1). Then, using the M -modulated signals (measurements) and the sensing matrix \mathbf{T} , a reconstruction algorithm is used to recover the input spectrum. In the experiments, we considered that the input spectrum was a directly sparse signal. The mean-squared error (MSE) between the input spectrum \mathbf{x} and the reconstructed spectrum $\hat{\mathbf{x}}$ was calculated. The MSE is defined as $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 / N$.

We firstly tested the spectral reconstruction performance of the fabricated filter-array with changing the full width at half maximum (FWHM) of the generated input signals. We made three noisy environments by adding the additive noise \mathbf{n} to Eq. (1) as $\mathbf{y} = \mathbf{T}\mathbf{x} + \mathbf{n}$ whose the signal to noise ratios (SNRs) were 20, 25, 30 dB. The SNR in decibels is defined as $10 \cdot \log_{10}(\|\mathbf{x}\|_2^2 / N\sigma^2)$, where σ is the standard deviation of the noise.

The spectral reconstruction performances with respect to the FWHMs are shown in Fig. 5(a). For the two-peak spectrum, the distance between two peaks was determined as $[1.5 \cdot \text{FWHM}]$, where $[\cdot]$ is the nearest integer function. We averaged all the MSEs of the spectrum over the peak-locations from 500 to 999 nm in a given FWHM. As shown in Fig. 5(a), the mono-peak spectrum is reconstructed better than two-peak spectrum. As the FWHM increased, the performance of spectral reconstruction is degraded. This is due to the increased sparsity of the spectrum.

Second, we verified the stability of noise along the SNR conditions for the fabricated filter-array. As shown in Fig. 5(b), the reconstruction performance on mono-peak spectrum is better than that of the two-peak spectrum. In addition, when the FWHM is 1 nm, the reconstruction performance is better than the FWHM with 2 nm. Despite the additive noise, the results show that the fabricated filter-array is robust to the noisy environments.

As depicted in Fig. 5, the reconstruction performance of the fabricated filter-array depends on the FWHM and the SNR. For the two-point resolution, the MSE has the smallest value when the FWHM of the two-peak spectrum is 1 nm. The overall MSEs are small enough to use the fabricated filter-array to conduct the optical experiments.

3.3. Optical experiments

Optical experiments were then conducted to evaluate the performance of the proposed system, as shown in Fig. 3(b). Narrow-band monochromatic lights and LEDs were used as input light sources. To gen-

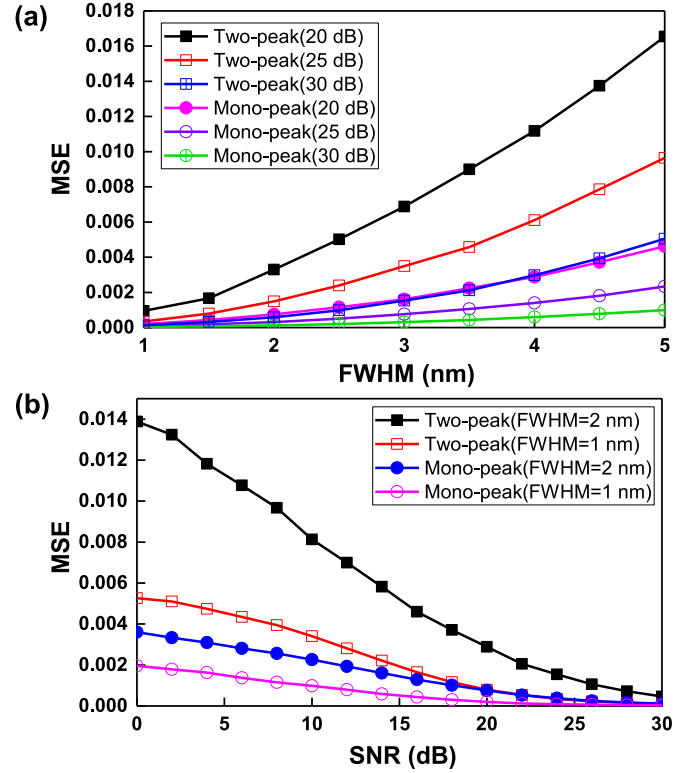


Fig. 5. (a) Computational reconstruction performance of the fabricated thin-film filter-array with respect to the FWHM. (b) Computational spectral reconstruction performance of the fabricated thin-film filter-array with respect to the SNR.

erate narrow-band light, a supercontinuum white light source (SuperK COMPACT, NKT Photonics) was placed in the monochromator, making a narrow band of light with a full width at half maximum (FWHM) of approximately 1 nm. These light sources were fed into the CMOS image camera through the filter-array, simultaneously capturing the M differently modulated signals. The M -modulated signals and the measured sensing matrix \mathbf{T} were then used to solve Eq. (3). We used a Gaussian kernel matrix as the sparsifying basis Φ . The spectral waveform can be represented as a linear combination of Gaussian kernels, and a Gaussian kernel can be easily generated with two parameters, namely the peak location and the FWHM value [4,18]. The *l1* *l*_s *noneg* algorithm [19] was used as a reconstruction algorithm to solve Eq. (3) with non-negativity constraints.

Fig. 6 shows the reconstruction results for monochromatic lights and LEDs. For comparison, the reference spectrum and the reconstructed spectrum were normalized to the range between zero and one.

The optical experimental results for monochromatic lights are shown in Fig. 6(a). In our optical experiment, depicted in Fig. 3(c), we use four different monochromatic spectra, with spectral peaks located at 600, 700, 800, and 900 nm, respectively. The reference spectra are measured using an optical spectrum analyzer (AQ-6315B, Ando) which indicate actual spectral peak locations at 598.7, 700.4, 800.5, and 900.4 nm, respectively. Using the fabricated filter-array CS spectroscopy with the reconstruction algorithm, the spectral peak locations are reconstructed at 599, 699, 799, and 901 nm, respectively. The mean FWHM of the reference spectra is approximately 1 nm, and the mean FWHM for the reconstructed spectra is approximately 1.4 nm.

Fig. 6(b) shows the spectral reconstructions of green (527 nm) and red (635 nm) LEDs. For the reference spectra, we measure the LEDs using a grating spectrometer (QE65000, Ocean Optics). The spectral peak locations for the reference LEDs are 527.6 nm (green LED) and 634.9 nm

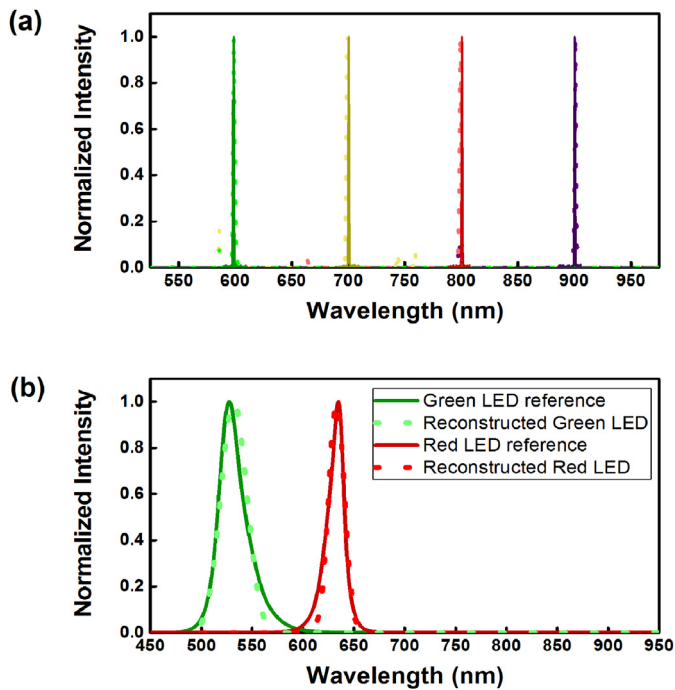


Fig. 6. Spectral reconstructions of several different input light sources. (a) Spectral reconstructions of monochromatic lights (dots) compared with reference spectra (solid lines): 600 nm (green), 700 nm (yellow), 800 nm (red), and 900 nm (purple). (b) Spectral reconstructions of LEDs (dots) compared with reference spectra (solid lines): green LED (527 nm), and red LED (635 nm).

(red LED), and the reconstructed spectral peak locations are 531 nm (green LED) and 633 nm (red LED). The peak signal-to-noise ratios are 28.3 dB (green LED) and 31.7 dB (red LED).

Discussing Fig. 6, the spectra of reconstructed monochromatic lights show several negligible spikes. This is probably due to background noise in the optical experiments. But overall, the reconstruction results of the proposed CS spectroscopy system for monochromatic lights and LEDs are similar to those of the grating spectrometer. Furthermore, the number of modulated signals is significantly small ($M=36$) that the measurement to wavelength sample ratio is 36:500 (ratio between M and N).

To further explore the performance of the proposed CS spectroscopy, we conducted the computational experiment on the fabricated filter-array using a continuous light source, halogen lamp. For the experiment, we used the measured sensing matrix T . The conventionally measured spectrum of the halogen lamp was used as the input spectrum x . The modulated signals were generated by numerically multiplying the sensing matrix and the input spectrum. By solving Eq. (3), we reconstructed the continuous spectrum of light. In Fig. 7, we present computational spectral reconstruction of the halogen lamp. The peak signal-to-noise ratio is 43.8 dB. Due to the limitations of our optical components to reject the spectrum of the halogen lamp except for the wavelength range from 500 to 1000 nm, we could not perform the optical experiment on the continuous source. However, the computational reconstruction result of the halogen lamp indicates that the fabricated filter-array can be utilized for recovering the various kinds of spectra in the given wavelength range without limitations of the optical components.

Fabricating the proposed filter-array can be more difficult than fabricating Fabry–Perot structure due to the large number of layers for the proposed filter-array. However, the proposed spectroscopy is compact and it does not need motorized components which were used with the Fabry–Perot structure [9]. In addition, thanks to the 2D array structure, the proposed spectroscopy captures all measurements in a single exposure. But the Fabry–Perot spectroscopy [9] required a number of

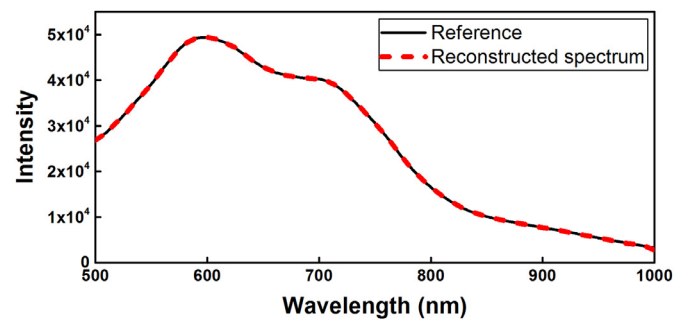


Fig. 7. Computational spectral reconstruction of a halogen lamp (red dash line) compared with the reference spectrum (black solid line) measured by a conventional spectrometer.

exposures as many times as the number of measurements. Compared to Fabry–Perot spectroscopy [8], the proposed spectroscopy utilizes 36 filters to cover the wavelength range from 500 to 1000 nm, but 100 filters were used in [8] to cover the range from 500 to 750 nm.

4. Conclusion

We have demonstrated a 2D array CS spectroscopy based on thin-film technology. A 2D thin-film filter-array is fabricated based on array processing. Using the fabricated filter-array, measurements are obtained to which the CS reconstruction algorithm is applied. Finally, demonstration of input spectrum reconstruction is successfully made. The proposed system is compact, portable, and obtains the necessary measurements in a single exposure thanks to its structural advantages. Moreover, it works over a wide spectral range, from the visible light region to the near-infrared region. Compared with conventional spectrometers (non-CS spectrometers), the proposed system has a high optical throughput and compatible spectral resolution performance in monochromatic lights and LEDs with significantly less number of measurements.

Acknowledgments




This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2018R1A2A1A19018665).

References

- [1] Bacon CP, Mattley Y, DePrece R. Miniature spectroscopic instrumentation: applications to biology and chemistry. *Rev Sci Instrum* 2004;75:1–16.
- [2] Kim S, Cho D, Kim J, Kim M, Youn S, Jang JE, et al. Smartphone based multispectral imaging: system development and potential for mobile skin diagnosis. *Biomed Opt Express* 2016;7:5294–307.
- [3] Clark RN, Roush TL. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *J Geophys Res Solid Earth* 1984;89:6329–40.
- [4] Oliver J, Lee W, Park S, Lee H-N. Improving resolution of miniature spectrometers by exploiting sparse nature of signals. *Opt Express* 2012;20:2613–25.
- [5] Oliver J, Lee W-B, Lee H-N. Filters with random transmittance for improving resolution in filter-array-based spectrometers. *Opt Express* 2013;21:3969–89.
- [6] Wang Z, Yu Z. Spectral analysis based on compressive sensing in nanophotonic structures. *Opt Express* 2014;22:25608–14.
- [7] August Y, Stern A. Compressive sensing spectrometry based on liquid crystal devices. *Opt Lett* 2013;38:4996–9.
- [8] Huang E, Ma Q, Liu Z. Etalon array reconstructive spectrometry. *Sci Rep* 2017;7.
- [9] Oiknine Y, August I, Blumberg DG, Stern A. Compressive sensing resonator spectroscopy. *Opt Lett* 2017;42:25–8.
- [10] Donoho DL. Compressed sensing. *IEEE Trans Inf Theory* 2006;52:1289–306.
- [11] Baraniuk RG. Compressive sensing [lecture notes]. *IEEE Signal Process Mag* 2007;24:118–21.
- [12] Qaisar S, Bilal RM, Iqbal W, Naureen M, Lee S. Compressive sensing: from theory to applications, a survey. *J Commun Netw* 2013;15:443–56.

- [13] Macleod HA. Thin-film optical filters. CRC press; 2001.
- [14] Barry JR, Kahn JM. Link design for nondirected wireless infrared communications. *Appl Opt* 1995;34:3764–76.
- [15] Pedrotti FL, Pedrotti LS. Introduction to optics. 2nd ed. Prentice Hall; 1993.
- [16] Topasna DM, Topasna GA. Numerical modeling of thin film optical filters. *Proc. SPIE* 2009;9666:96661P.
- [17] Candes EJ, Eldar YC, Needell D, Randall P. Compressed sensing with coherent and redundant dictionaries. *Appl Comput Harmon Anal* 2011;31:59–73.
- [18] Kurokawa U, Choi BI, Chang C-C. Filter-based miniature spectrometers: spectrum reconstruction using adaptive regularization. *IEEE Sens J* 2011;11:1556–63.
- [19] Koh K, Kim S-J, Boyd S. An interior-point method for large-scale l_1 -regularized logistic regression. *J Mach Learn Res* 2007;8:1519–55.

Dry Electrode-Based Fully Isolated EEG/fNIRS Hybrid Brain-Monitoring System

Seungchan Lee , Younghak Shin , Member, IEEE, Anil Kumar , Member, IEEE, Minhee Kim, and Heung-No Lee , Senior Member, IEEE

Abstract—A portable hybrid brain monitoring system is proposed to perform simultaneous 16-channel electroencephalogram (EEG) and 8-channel functional near-infrared spectroscopy (fNIRS) measurements. Architecture-optimized analog frontend integrated circuits (Texas Instruments ADS1299 and ADS8688A) were used to simultaneously achieve 24-bit EEG resolution and reliable latency-less ($<0.85 \mu\text{s}$) bio-optical measurements. Suppression of the noise and crosstalk generated by the digital circuit components and flashing NIR light sources was maximized through linear regulator-based fully isolated circuit design. Gel-less EEG measurements were enabled by using spring-loaded dry electrodes. Several evaluations were carried out by conducting an EEG phantom test and an arterial occlusion experiment. An alpha rhythm detection test (eye-closing task) and a mental arithmetic experiment (cumulative subtraction task) were conducted to determine whether the system is applicable to human subject studies. The evaluation results show that the proposed system is sufficiently capable of detecting microvoltage EEG signals and hemodynamic responses. The results of the studies on human subjects enabled us to verify that the proposed system is able to detect task-related EEG spectral features such as eye-closed event-related synchronization and mental-arithmetic event-related desynchronization in the alpha and beta rhythm ranges. An analysis of the fNIRS measurements with an arithmetic operation task also revealed a decreasing trend in oxyhemoglobin concentration.

Index Terms—Electroencephalogram (EEG), functional near-infrared spectroscopy (fNIRS), hybrid brain-computer interface, multimodal analysis, portable instrument, simultaneous measurement.

Manuscript received April 12, 2018; revised July 19, 2018; accepted August 10, 2018. Date of publication August 27, 2018; date of current version March 19, 2019. This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MISP) under Grants [NRF-2018R1A2A1A19018665] and [NRF-2015A2A1A05001826], and by the Brain Research Program through the NRF funded by the Ministry of Science, ICT & Future Planning under Grant [NRF-2016M3C7A1905475]. (Corresponding author: Heung-No Lee.)

S. Lee, A. Kumar and H.-N. Lee are with the Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: future-max7@gmail.com, heungno@gist.ac.kr).

Y. Shin is with the Department of Electronic Systems, Norwegian University of Science and Technology.

M. Kim is with the Department of Biomedical Science and Engineering, Institute of Integrated Technology, Gwangju Institute of Science and Technology.

Digital Object Identifier 10.1109/TBME.2018.2866550

I. INTRODUCTION

THE brain-computer interface (BCI) [1], [2] was originally developed to assist severely disabled people who cannot control their peripheral nerves and muscles, due to neurological and neuromuscular disorders such as amyotrophic lateral sclerosis, brainstem strokes, and spinal cord injuries. This technology is now advancing to provide a new communication channel that facilitates human-machine interaction. Presently, a number of new techniques based on wearable devices and the Internet of Things (IoT) are being applied to BCIs related to the fields of healthcare, telemedicine, and clinical care [3]. Current BCI technology, however, faces several challenges, such as its limited number of controllable functional-brain signals [4], the need for recalibration of the signal processing algorithms, and uncontrollability for a non-negligible proportion of the users, referred to as “BCI-illiteracy” [5].

Multimodal analysis of brain activities—the so-called hybrid BCI [6], which can be implemented by simultaneously acquiring and analyzing two or more brain signals, has been proposed as an alternative BCI technique capable of overcoming the above challenges. Two or more complementary neurological signals can be combined and shared to maximize the amount of exploitable information, thereby enhancing the robustness of control accuracy in real-world applications.

Hybrid BCI systems could be established by the fusion of two or more modalities amongst various brain imaging techniques, such as electroencephalogram (EEG), magnetoencephalogram (MEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS). Among these modalities, the disadvantages of MEG- and fMRI-based techniques is the need to install the machines in confined areas and the fact that they can only be used for short runtimes because of their high cost, large size, and the need for expert operators [7]. Contrary to this, EEG- and fNIRS-based brain-monitoring systems are electromechanically simple, making them easy to design for lightweight, compact and low-cost systems. EEG/fNIRS-combined hybrid systems could easily be built as portable or wearable devices and utilized in more dynamic applications, such as driver drowsiness detection [8] and seizure monitoring in epileptic patients [9].

An EEG is the electrical potential produced by the sum of the synchronous activation from the dendritic branches of a large number of neurons. Because EEG recording can be achieved noninvasively through the electrodes placed on the scalp and its

time resolution is relatively high in the millisecond range, it is widely used as an electrophysiological recording modality [2]. On the other hand, fNIRS measures the changes in the local concentration of oxygenated and deoxygenated hemoglobin in the cerebral cortex region by utilizing low-energy optical radiation from light sources of two different wavelengths in the near-infrared range (700–1000 nm). Although this technique demonstrates a slower response compared to EEG, it enables an investigation of metabolic and microcirculatory neuronal activation regardless of the electrically synchronized activation of neurons [10]. The simultaneous acquisition of EEG and fNIRS measurements could provide more comprehensive neurodynamic information regarding the accessible neuronal metabolism and neuroelectric activities. As such, several researchers have recently developed EEG–fNIRS hybrid systems for use in various applications [11].

A review of the available literature related to hybrid BCI systems indicates that a combination of individual EEG and fNIRS systems has been used in various experimental accomplishments regarding motor imageries [12]–[16], visual and auditory stimulations [17] and mental workloads [18], [19]. In such a setup, fully synchronized operation of the entire system is difficult, because each individual system contains its own controller that is operated at a predefined clock speed. Therefore, the measurements acquired from two systems may not be completely synchronized in the absence of a precise simultaneous control mechanism. Attempts to address this concern have resulted in the design of customized EEG–fNIRS hybrid acquisition instruments.

One of the first attempts to this end has been started with the design of a probe for simultaneous measurements of EEG and fNIRS data [20]. Lareau *et al.* [21] and Sawan *et al.* [22] have proposed a similar hybrid system that was capable of acquiring multi-channel EEG and fNIRS measurements. However, it was difficult to use it as an out-of-lab device because of its large size ($16 \times 13 \times 8.2 \text{ cm}^3$). In 2013, a field-programmable gate array (FPGA) and an EEG application-specific integrated circuit (ASIC) based compact, and advanced bimodal acquisition system was developed by Safaie *et al.* [23]. Recently, Luhmann *et al.* [24] developed a miniaturized modular hybrid system, wherein one module was capable of simultaneously monitoring four channels of bio-electrical and bio-optical measurements. However, these reported studies still have several limitations related to practical usability in daily-life monitoring. The conductive gel of conventional wet electrodes leads to user irritation and easily degrades the signal quality as it becomes dry, making long-term monitoring difficult. Efficient suppression of the crosstalk and noise characteristics in a mixed-signal system is another key challenge in designing a hybrid instrument.

This paper proposes a dry electrode-based portable hybrid brain monitoring (HBM) system that provides simultaneous monitoring of fully synchronized 16-channel EEG and 8-channel fNIRS. Aiming at a use of out-of-lab and clinical applications, the performance and availability of the instrument have been improved by integrating the following advanced features with the proposed system:

- 1) Dry electrode-based gel-less EEG acquisition [25]–[27] for easy to put on, non-degraded EEG quality, and significant reduction in wearing time to less than 10 minutes (refer to Section II-C);
- 2) Architecture-optimized frontend design for sufficient resolution EEG and timing-secured errorless bio-optical measurement, i.e., delta-sigma (Δ - Σ) architecture ADC-based 24-bit EEG resolution and successive approximation register (SAR) architecture ADC-based latency-less ($<0.85 \mu\text{s}$) bio-optical measurements (refer to Section II-A);
- 3) Linear regulator-based fully-isolated circuit design for maximization of noise and crosstalk suppression (refer to Section II-B);
- 4) Customizable EEG electrode-positioning structure (named as EEGCAP) to meet various experimental scenarios (refer to Section III-B-1)).

Several evaluation tests were performed to verify the hybrid data acquisition performance. The acquisition of EEG measurements using the dry electrodes was evaluated by performing an EEG phantom test. An arterial occlusion experiment was performed to verify the hemodynamic responses of the fNIRS measurements. Finally, human subject studies including an alpha rhythm detection test and an experiment to assess mental arithmetic operation were performed to verify the practical capabilities for EEG and fNIRS feature measurements.

The remainder of the paper is organized as follows: Section II and III provide detailed descriptions of the design methods and the implementation of the proposed system, respectively. The evaluation of the EEG/fNIRS measurements and human subject studies, including an alpha rhythm detection test and a mental arithmetic operation experiment, are presented in Section IV. Section V summarizes several results, including system implementation, acquisition capability evaluation, and offline analysis of human subject studies. The contributions of this study are discussed in Chapter VI in comparison with previous studies. Finally, concluding remarks with a summary of the system design and experimental results are given in Section VII.

II. SYSTEM DESIGN

This section describes the key design methods for implementing the proposed HBM system, namely *architecture-optimized frontend design*, *linear regulator-based fully-isolated circuit design*, and *dry electrode-based gel-less EEG acquisition*.

A. Architecture-Optimized Frontend Design

Physiological signals, such as EEG and fNIRS, possess small amplitudes and are highly susceptible to various types of noise. For this reason, the use of complicated signal-conditioning circuits becomes necessary to achieve high-precision measurements. State-of-the-art integrated analog frontend (AFE) integrated circuits (ICs) combined with high-resolution analog-to-digital converters (ADCs), signal-conditioning circuits, and associated built-in circuits and their design benefits were reported [28]. The integrated functions of these ICs assist to reduce the number of discrete components required in the design of a data

acquisition system, enabling miniaturized and low-cost designs with reliable performance.

The proposed design employs the ADS1299 AFE IC (Texas Instruments, USA) [29] for EEG measurements. It was integrated with 8-channel, 24-bit resolution Δ - Σ ADCs, programmable gain amplifiers (PGAs), and other built-in peripherals. A sufficiently small step size of the least significant bit (LSB) ($0.022 \mu\text{V}$ at a 24 PGA gain) and low peak-to-peak noise performance ($0.98 \mu\text{V}$ at a 250-SPS sampling rate and a 24 PGA gain) enables precise detection of EEG signals in the μV range. The integrated 8-channel ADCs allow simultaneous sampling of multiple input measurements, thus no sampling skew and glitch noise exist in the converted data without the need for sample-and-hold circuits.

Although the ADS1299 was used for EEG measurements in previous studies [24], this is the first time the ADS8688A (Texas Instruments, USA) [30] was used for the acquisition of bio-optical measurements. This device is a 16-bit successive SAR ADC-based AFE integrated with numerous built-in functions such as 8-channel input multiplexer, PGAs, and second-order low-pass filters.

Compared to the Δ - Σ architecture employed in ADS1299, the SAR ADC architecture [31] can provide the precise delay-less measurement required for bio-optical acquisition. The delta sigma architecture is advantageous for acquiring high-resolution measurements exceeding 20 bits; however, its operating mechanism requires the use of a digital decimation filter for noise-shaped representation of oversampled data, thereby resulting in conversion latency known as the settling time [32]. This latency represents the delay between the beginning of the input signal conversion and the end time at which fully settled output data are available. In the case of the ADS1299, this latency reaches 16 ms at a sampling rate of 250 SPS. Unlike delta-sigma ADCs, the SAR ADC architecture does not require the conversion latency because it repeatedly performs a zero-latency task, which compares the reference voltage and input measurements through a sample-and-hold circuit, a comparator, and a DAC. This zero-latency feature, which produces digitized data within $0.85 \mu\text{s}$ in case of the ADS8688A, leads to reliable delay-less measurement. Because the bio-optical measurement requires on-time acquisition within predefined timing bins (4 ms) when the NIR light source is in an active state, this delay-less characteristic is essential for accurate acquisition of bio-optical measurements. Therefore, the ADS8688A, instead of the ADS1299, which is Δ - Σ architecture ADC-based AFE IC, is employed for the bio-optical measurement.

B. Linear Regulator-Based Fully-Isolated Circuit Design

In mixed-signal systems in which analog and digital components are integrated into a single circuit, the crosstalk noise generated in digital circuits could be coupled to neighboring analog circuits via stray capacitances [33]. In the proposed HBM system, a periodical switching operation of the NIR light source is necessary to acquire bio-optical measurements. The oscillating noise in the digital circuits is unavoidable because of the instantaneously high current flow in the driving circuit of the

light source. Without careful consideration of the crosstalk, this noise may appear on the analog circuits associated with the AFE ICs and can easily distort the small EEG and bio-optical amplitudes.

The crosstalk rejection capability was maximized by implementing a fully isolated circuit design technique, such as a circuit design with separate ground planes and an isolated digital interface, in the power and control circuits of the proposed system. The design of the power supply circuit included the use of a dedicated lithium-polymer battery and an isolated DC-DC converter (DCP020509, Texas Instruments, USA) to separate the ground for the data acquisition circuits and the digital control circuit. This design results in a total of three completely separated ground planes. Since independent return current paths are created on each ground plane and these paths are completely isolated from each other, the switching noise generated in the control circuit cannot reach the data acquisition areas. Therefore, the EEG and fNIRS acquisition circuits are able to maintain flat and stable ground potentials. Two digital isolators (Silicon Labs Si8662) are also used for the isolated interface of the EEG and fNIRS acquisition circuits. Many advanced features, such as high data throughput, low propagation delay, and noise robustness of the isolator IC serve to provide a reliable and uncoupled data path in the digital interface.

The linear regulator-based power supply circuits were carefully designed by using a number of decoupling capacitors and ferrite beads to provide low-noise DC power to the data acquisition circuits. The linear regulators provide several advantages compared to DC-DC converters, such as highly regulated output voltage, low noise spectral density, and a high power supply rejection ratio (PSRR), thereby making them ideally suited for noise-sensitive applications. In addition to these low-noise power supply circuits, an optimized printed-circuit-board (PCB) layout and advanced circuit-design techniques, such as grounding, signal routing, and decoupling [34], were applied to maintain stable and regulated DC voltages and build a low-impedance return current path.

C. Dry Electrode-Based Gel-Less EEG Acquisition

Conventionally, disc-shaped Ag/AgCl electrodes have been employed in EEG measurements. These electrodes require the use of conductive gels and hair preparation during installation in order to reduce the electrical impedance to an acceptable level. These procedures are time consuming and cause irritation in most subjects, because conductive gels are sticky. Moreover, these electrodes are not suitable for long-term and ambulatory applications, because conductive gels dry over time and their adhesion is easily lost during motional vibrations. Therefore, the signal quality of the wet electrodes may be continuously degraded in ongoing experiments, thus the use of wet electrodes is to be limited in experiments requiring more than 30 minutes. To overcome these problems, dry electrodes, which do not require conductive gels, are used in the proposed system. These electrodes comprise spring-loaded probes that maintain a constant pressure on the surface of the uneven scalp regardless of its movement. Consequently, these electrodes are capable of more

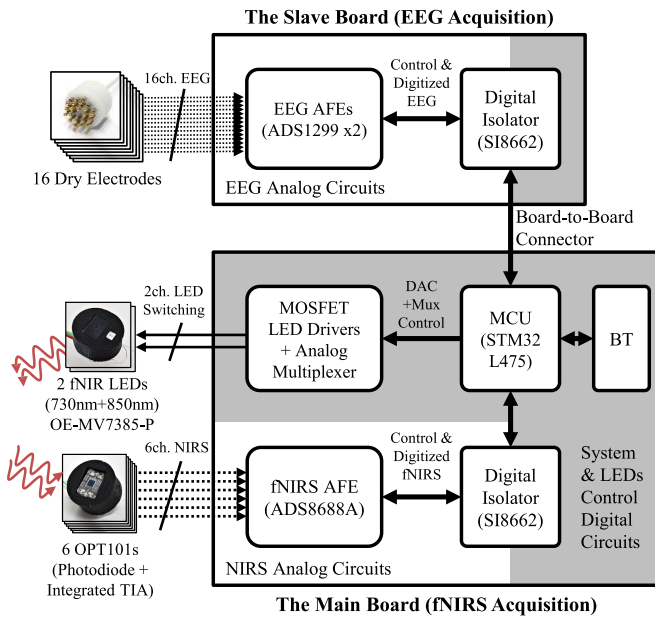


Fig. 1. Simplified schematic of the proposed HBM system. Solid and dotted arrows indicate the flow of digital logic signals and analog measurements, respectively. Likewise, the shaded and transparent regions indicate the digital and analog circuits, respectively. The boundary between the analog and digital circuits is isolated by a digital isolator and DC-DC converter. The dedicated EEG acquisition circuits is also isolated from the main board circuits.

stable EEG measurements even in out-of-lab environments. The dry-electrode structure is described in detail in Section III-B-1).

III. IMPLEMENTATION

A. Instrumentation

1) Data Acquisition Circuit: Fig. 1 depicts a schematic of the proposed system excluding the power supply circuits. The system comprises two boards—the main board and slave board. The main board is capable of performing 8-channel bio-optical measurements, and 4-channel dual-wavelength LED emissions. The slave board was designed to perform 16-channel EEG measurements. The two boards were connected using the Molex board-to-board connector, and all components were controlled by the STM32L475 low-power microcontroller (STMicroelectronics, USA) installed on the main board.

The following procedure was used to perform bio-optical measurements on the main board. Common-mode electromagnetic and radio-frequency interference noise is first filtered out from raw bio-optical measurements using a simple RC low-pass filter in the input stage. Inside the embedded ADS8688A AFE IC, the acquired bio-optical signal is amplified by the integrated PGA to pre-programmed input ranges (± 0.64 V) and subsequently filtered by an anti-aliasing low-pass filter with a 15-kHz cutoff frequency. Because the actual sampling rate of the bio-optical measurement reaches 20 kHz to obtain an averaged measurement from quick repeated samples, the built-in anti-aliasing filter is required for aliasing rejection. The filtered signal is then fed to the ADC driver and multiplexer circuits, and is finally sampled by a 16-bit SAR ADC. According to this

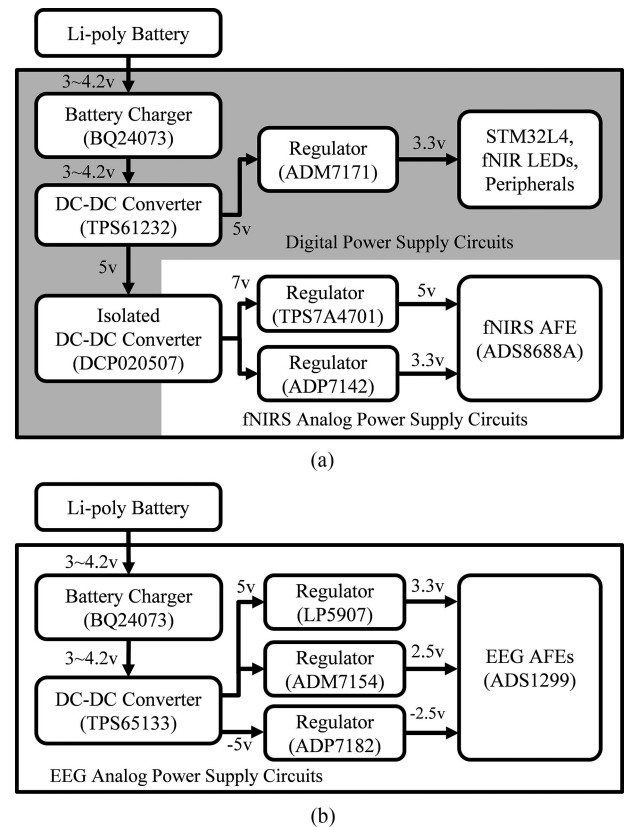


Fig. 2. Schematics of power-supply circuit for (a) main board, and (b) slave board. Two lithium-polymer batteries supply power to the main board and the slave board, respectively. In the main board, the isolated DC-DC converter separates the ground planes for the main control circuit (shaded digital power supply section) and the isolated NIRS acquisition circuit (fNIRS analog power supply section).

procedure, 8-channel bio-optical data can be finally obtained at a 5-SPS sampling rate.

The following procedure was also used to perform EEG measurements on the slave board. The EEG measurements acquired by the dry electrodes are filtered by the onboard input filter stage. X2Y type capacitors [35] were employed in this filter stage to facilitate higher attenuation of electromagnetic and radio-frequency noise, while reducing onboard space requirements. Inside the ADS1299 AFE IC, the filtered EEG measurements are amplified by a built-in low noise PGA with a 24 gain setting and digitized by a dedicated ADC for each channel over every sampling period (4 ms). The sampled EEG data are then transmitted to the microcontroller (MCU) via an SPI bus. With two ADS1299s in a daisy-chained configuration that allows multiple ICs to be controlled simultaneously using a single shared bus, 16-channel EEG measurements can be obtained at a 250-SPS sampling rate.

2) Power Supply Circuit: Fig. 2 depicts a schematic of the power-supply circuits of the proposed HBM system. The proposed system is powered by two lithium-polymer batteries—one each for the main and slave boards—which can be charged via the onboard battery management IC (Texas Instruments BQ24073) through a USB port. As the battery

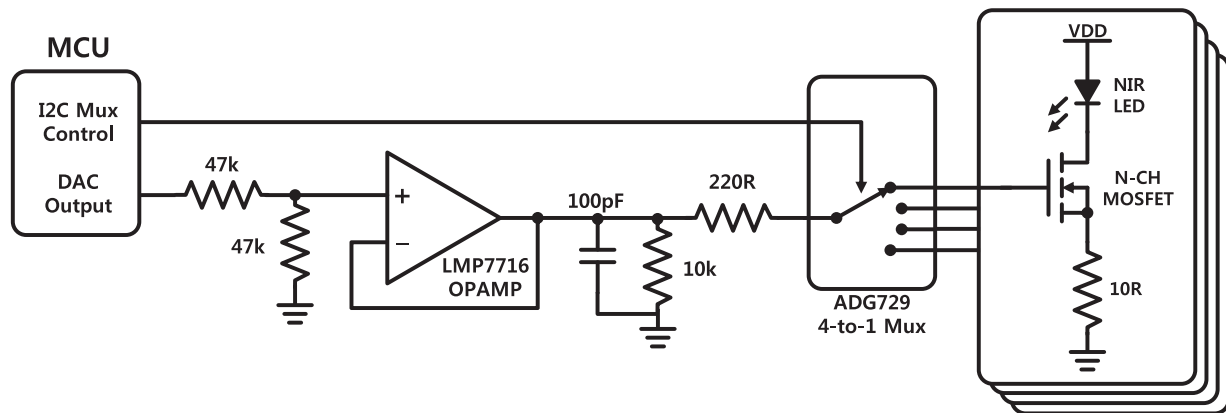


Fig. 3. Schematic of the MOSFET-based NIR LED driving circuit employed in the proposed HBM system. This circuit was combined with a DAC, analog multiplexer, and OPAMP-based buffering circuit to flexibly control the emission intensity of the four LEDs. By implementing two copies of this circuit, the proposed system can control up to eight LED emissions.

voltage decreases over time, boost and dual-output DC–DC converters (Texas Instruments TPS61232, TPS65133) are used to stabilize these output voltages. An isolated DC–DC converter (Texas Instruments DCP020507) is employed to supply fully isolated power for the fNIRS acquisition circuits on the main board. In the final stage of the power-supply circuits, low-noise DC voltage is lastly delivered to the AFE ICs, MCU, and other peripherals through six low-noise linear regulators (Analog Devices—ADM7154, ADP7182, ADP 7142, and ADM7171; Texas Instruments—TPS7A4701 and LP5907).

3) MOSFET-Based LED Driving Circuit: Fig. 3 illustrates the schematic of the MOSFET-based LED driving circuit. Because the number of NIR light sources required may vary depending on the configuration of the probe set layout and the experimental paradigm, a programmable control function for multi-channel emission is required for the LED driving circuitry. A calibration function for radiant intensity is also necessary because the NIR LED may exhibit radiant power mismatch even for the same current consumption. Thus, a programmable LED driving circuit was designed to flexibly control the radiant intensity of multi-channel NIR LEDs by combining a digital-to-analog converter (DAC), an analog multiplexer and MOSFET drivers. In operation, the MCU regulates the gate voltage of the MOSFET driver by controlling the output voltage of the built-in digital-to-analog converter (DAC) of the MCU. The regulated gate voltage is buffered with an OPAMP and then fed to the analog multiplexer (Analog Device ADG729) for controlling multi-LED emissions. The multiplexed gate voltage is lastly supplied to the N-channel MOSFET driver to modulate the LED current flow. This design provides flexibility to control as many as eight NIR LED emissions with fine-tuned radiant intensity in the proposed system. In the human subject studies described in this paper, the radiant intensity for all NIR LEDs was manually adjusted to 10 mW using an optical power meter and DAC output voltage control.

B. Sensors

Customized sensor units were designed for the EEG and bio-optical measurements to enhance the usability and re-

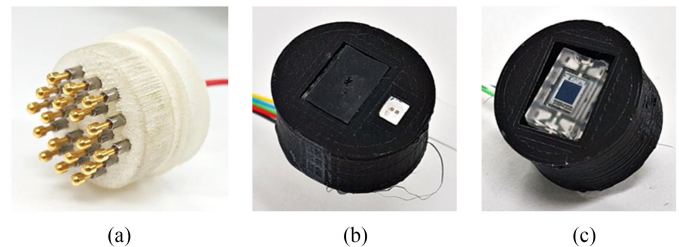


Fig. 4. (a) Dry electrode for EEG measurement, (b) dual wavelength LED-based NIR light source unit, and (c) silicon photodiode-based NIR detector unit for bio-optical measurement.

configurability of the proposed system. The sensor units comprise 16-channel dry electrodes, 2-channel NIR LEDs, and 6-channel photodiodes.

1) Spring-Loaded Dry Electrode and Customizable EEGCAP: The Fig. 4(a) depicts a prototype of the dry electrode for the EEG measurements, which comprises spring-loaded probes, a PCB, and a housing. The electrode unit, which is designed to remain in contact with the subject's scalp, acquires EEG potentials via the 18 spring-loaded probes (Leeno Industrial Inc., SK100R). Each probe comprises four components—the (1) plunger, (2) barrel, (3) spring, and (4) probe receptacles. The plunger is combined with a barrel and spring to constitute the spring-loaded structure. Each spring can withstand up to 54 g of pressure in its maximally compressed state. This enables each probe to maintain a suitable contact pressure on the uneven surface of the scalp. In terms of electrical specifications, the resistance of each probe is less than 50 mΩ, which is sufficiently low for conducting bioelectrical measurements. All probes are electrically connected to each other via the PCB embedded in the electrode housing and are thereby linked to a single electrode wire. The entire electrode assembly is enclosed by the 3D-printed plastic housing.

A helmet-like bracket (named the EEGCAP) was designed using flexible rubber materials to hold the dry electrodes in position in accordance with 10–20 systems. The mesh-type EEGCAP structure was equipped with as many as 58 holes to allow electrodes to be positioned on the scalp. Each electrode was

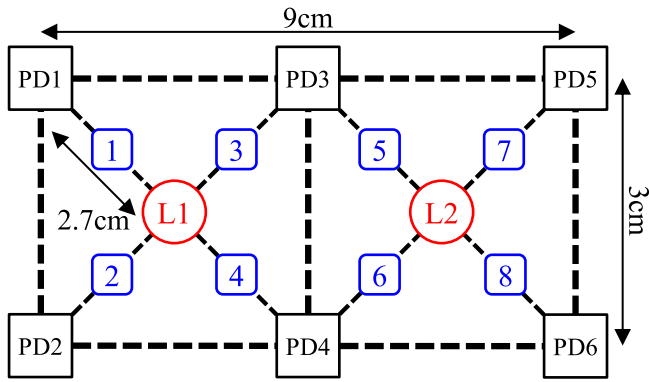


Fig. 5. Installation layout of NIR LEDs (L1 and L2) and photodiodes (PD1–PD6) for acquisition of the bio-optical measurements. To investigate hemodynamic changes at the frontal lobes, the light source and detector units are attached using a transparent double-sided tape. This layout produces 8-channel hemodynamic responses from the 1-to-8 bio-optical channels marked in blue color.

firmly engaged in the hole via an interlocking frame structure, and able to continuously push against the subject's scalp to maintain a constant pressure. This customizable structure could make a number of configuration choices available in terms of the electrode-positioning layout depending on the experimental paradigm.

2) NIR Light Source and Detector Units: Dual wavelength (730 and 850 nm) AlGaAs LEDs (Opto ENG OE-MV7385-P) were used for the NIR light source unit depicted in Fig. 4(b). Two LEDs packaged in a miniaturized plastic leaded chip carrier (PLCC) were soldered onto a light source PCB and covered by 3D printed materials. The spectral spread of the emitted radiation ($\Delta\lambda = 30\text{--}40$ nm) was broader compared to that of monochromatic laser diodes ($\Delta\lambda \approx 1$ nm). However, the incoherent and un-collimated characteristics of the LED light source achieve sufficient tissue penetration to enable the investigation of local hemodynamic changes. Owing to its suppressed heating and low risk of retinal damage, it can be used in direct contact with the human scalp [36].

The NIR detector unit depicted in Fig. 4(c) was based on a silicon photodiode device (Texas Instruments OPT101) integrated with an on-chip trans-impedance amplifier. Because the device exhibits high spectral sensitivity in the infrared spectrum (>0.5 A/W in the 730–850 nm wavelength), it is optimized for use in NIR detection applications. Owing to the built-in trans-impedance amplifier circuitry composed of an operational amplifier and an internal feedback network, the photodiodes provide direct voltage output with a sufficiently wide bandwidth (14 kHz) which is linearly proportional to the detected light intensity. The photodiode is soldered onto a detector PCB along with decoupling capacitors, and housed inside a 3D-printed-casing.

Fig. 5 illustrates the positioning layout for NIR LEDs and photodiodes for placement on the subject's forehead. The layout configuration occupies a $9\text{ cm} \times 3\text{ cm}$ area with two NIR LEDs and six photodiodes, and the distance between the light source and the detector unit was set at 2.7 cm. In operations using this layout, the NIR LEDs flicker alternately in accordance with the pre-programmed LED switching sequence and only the

photodiodes surrounding the turned-on LED are instantaneously activated. Each hemodynamic response is measured in the area between the pair of light sources and the detector unit and this area is defined as a bio-optical channel. To achieve the maximum number of bio-optical channels in the restricted forehead space, measurements for bio-optical channels 3 through 6 located between NIR LEDs L1 and L2 are all required. By exploiting a TDM-based channel-sharing scheme where one photodiode can provide multiple independent measurements in non-overlapped timing periods, the four independent measurements for these centrally located bio-optical channels can be provided from photodiodes PD3 and PD4; i.e., photodiode PD3 can provide measurements for the 3rd and 5th bio-optical channels and photodiode PD4 can provide measurements for the 4th and 6th optical channels in the same manner. Therefore, this channel sharing operation enables the proposed sensor layout to acquire 8-channel bio-optical measurements with only six photodiodes.

C. System Operation and Hybrid Data Acquisition

The ADC basically converts analog input signals into digitized signals with consistent intervals based on an internal or external reference clock. However, the clock may have its own tolerance and frequency drift characteristics. In heterogeneous data-acquisition systems employing two or more ADCs to produce a fully synchronized data stream, the clock tolerance of individual ADCs makes accurate synchronization difficult to achieve. This problem can be solved by using a reference system clock to which all ADCs could be universally referred.

Complete synchronization is achieved between the EEG and bio-optical measurements by using the data-ready signals (referred to as DRDY in the datasheet) generated by the ADS1299 AFE IC as the reference system clock. The DRDY signal represents the transition of a falling edge when the digitized EEG data stream becomes valid. It, therefore, generates a pulse signal of the same period as the sampling rate of EEG acquisition. By synchronizing the emission control of NIR LEDs and data acquisition of ADS8688A with the DRDY pulse cycle, the complete synchronization between EEG and bio-optical measurements can be preserved regardless of the occurrence of small timing errors in the reference clock of each AFE.

Fig. 6 depicts a single period of simultaneous EEG and bio-optical acquisition captured from the logic analyzer screen. Once ADS1299 begins to acquire EEG measurements at a pre-programmed sampling rate (250 SPS), the DRDY pulses begin to be generated with the same sampling period (4 ms) as EEG data generation. In accordance with the generation of the DRDY pulse, NIR radiation of dual wavelengths (730 and 850 nm) is alternately switched in the order—L1 (730 nm)—L2 (730 nm)—L1 (850 nm)—L2 (850 nm)—over the course of 50 EEG acquisition cycles (200 ms). Each time the NIR LED is turned on by the multiplexer switching, the radiation lasts for 4 ms, during which time the ADS8688A acquires NIR light intensities from the set of activated photodiodes surrounding the turned-on LED; i.e., when the L1–730 nm or L1–850 nm states are active, measurements from the photodiodes PD1–PD4 are sampled. This also applies to the two L2 states and sampling of photodiodes PD3–PD6. To

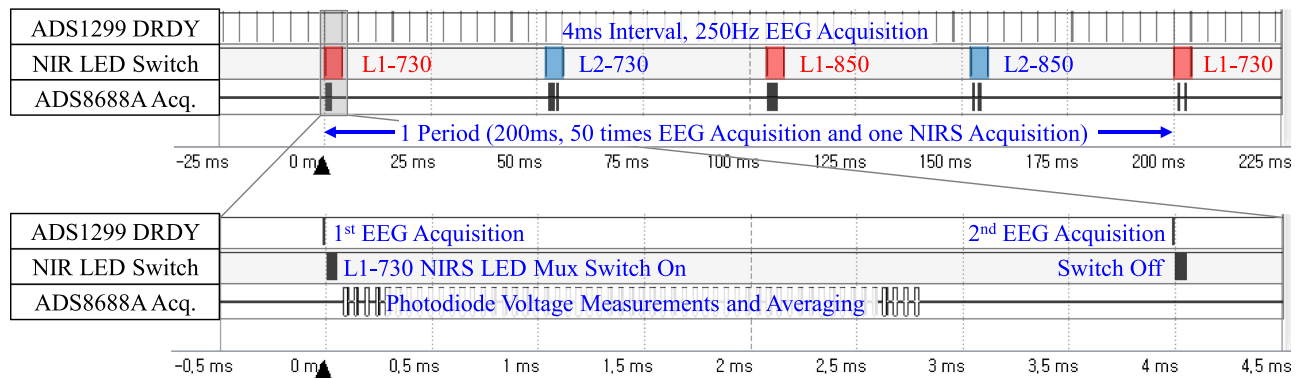


Fig. 6. Logic analyzer view of one period of simultaneous EEG and fNIRS acquisition and magnified view of the upper gray region (-0.5 – 4.5 ms). According to the DRDY pulse generated by the ADS1299, 16-channel EEG measurements are acquired, and the NIR light sources L1 and L2 are alternately activated for 4 ms. During NIR irradiation for 4 ms, each of the 4-channel photodiodes surrounding the light source measured the light intensity 14 times and averaged it. A total of 16-channel of bio-optical measurements are obtained over a 200-ms period, which is converted into 8-channels of fNIRS data during the fNIRS decoding process.

obtain stable measurements with minimized background noise, the light intensity measurement of each bio-optical channel is repeatedly acquired 14 times with a $50\text{-}\mu\text{s}$ interval and subsequently averaged. During the 4-ms period of LED radiation, a total of 56 optical measurements are then sequentially obtained within 2.8 ms from the four photodiodes surrounding the turned-on LED. While the four LEDs are flashing sequentially within a 200-ms period, a total of 16 bio-optical measurements can be obtained through a time-division multiplexing operation.

The aforementioned sequence allows fully synchronized 16-channel EEG and 16-channel bio-optical measurements to be acquired every 4 ms ($= 250$ SPS) and 200 ms ($= 5$ SPS), respectively. The acquired measurements are then packetized and successively transmitted to the host device via the SPBT3.0 DP2 Bluetooth module (STMicroelectronics, USA) with a header and timing information. The host device decodes the packets of EEG and bio-optical data using MATLAB 2014a (MathWorks, USA). Using the Modified Beer-Lambert Law [37], [38] in the decoding process, the 8-channel fNIRS data, including concentration changes in the oxy- (ΔHbO), deoxy- (ΔHbR), and total hemoglobin (ΔHbT), are also converted from the 16-channel bio-optical data.

The MCU system was programmed to perform the following operations:

- 1) Peripheral initialization—establishment of peripheral interfaces (SPI interface, general purpose input/output ports, and interrupt routine) and setting up registers of all AFE ICs;
- 2) Launching the data-retrieval loop upon detection of the start trigger;
- 3) Acquisition of EEG data from ADS1299, when a DRDY pulse is generated;
- 4) Control of NIR LED emission in accordance with the LED switching schedule and DRDY trigger;
- 5) Acquisition of bio-optical data of the predefined photodiode sets from ADS8688A in accordance with the LED control sequence;
- 6) Packetization of acquired EEG and bio-optical data along with header and timing indication and subsequent

transmission of data packets to the host device via the Bluetooth module;

- 7) Repeating steps 3 through 6 until the stop trigger is detected.

IV. EVALUATION AND EXPERIMENT

A. Evaluation of EEG and fNIRS Acquisition

1) **EEG Phantom Experiment using Dry Electrodes:** The proposed HBM system employs dry electrodes for EEG acquisition instead of the conventional wet electrodes for wide applicability and enhanced usability. Therefore, it is necessary to verify the acquisition capability of the dry electrodes at the level of micro-voltage amplitudes. In the proposed system, the fNIRS and EEG acquisition circuits operate simultaneously. Thus, EEG signal acquisition is subjected to interference from the electrical switching noise generated by the NIR LEDs and this effect must be examined. For this purpose, we devised an EEG phantom experiment.

First, EEG-like voltage signals were generated. Raw EEG data samples of 60-s duration were taken from the C3 channel of a BCI competition 3-IVa dataset (motor imagery task, down-sampled to 250 Hz) [39]. These EEG data samples were then inputted to an arbitrary waveform generator (Keysight 33220A) for reproduction of a EEG voltage waveform. The reproduced voltage waveform was then passed through a voltage divider circuit (of 10000:1 ratio) to create a microvolt-level EEG signal. This voltage waveform was finally fed to the EEG phantom.

Second, an EEG phantom was created using a conductive rubber pad ($10\text{ cm} \times 12\text{ cm} \times 5\text{ mm}$, $100\ \Omega/\text{cm}$) to simulate a real human scalp. An NIR LED unit is placed at the center of the rubber pad. Then, one dry and one wet electrode (with conductive gel) were attached around the LED unit on the rubber pad to emulate the NIR interference during EEG signal measurement. The two electrodes and the NIR LED unit were connected the EEG input port and NIR LED driving port of the HBM, respectively. The EEG reference input of the HBM system was connected to the ground potential of the waveform generator.

Third, the voltage waveform of 60-s duration prepared in the first step was reproduced in the EEG phantom. Measurement samples were recorded at a sampling speed of 250 SPS from the two electrodes during the 60-s period. The two acquired signals were compared with the prepared voltage waveform in terms of correlation coefficients. In offline analysis, three correlation coefficients were calculated and analyzed depending on the NIR LED ON/OFF state. The correlation coefficient between the acquired signal using a dry electrode and the prepared waveform is ρ_D ; the correlation coefficient between the acquired signal using a wet electrode and the prepared waveform is ρ_W ; and the correlation coefficient between two acquired signals obtained using a wet electrode and a dry electrode is ρ_{DW} . To ensure reliability of the analysis, this test was repeated thrice, and the averaged correlation coefficients were compared.

2) Arterial Occlusion Experiment: The hemodynamic response of the proposed system was verified by evaluating the fNIRS responsivity using an arterial occlusion experiment [22], [23]. The experiment was performed using an inflatable arm cuff and a sphygmomanometer. The arm cuff could be shrunk to block arterial blood flow to artificially change the concentration of oxy and deoxy hemoglobin in the bloodstream on the arm. This would enable us to verify the hemodynamic behavior of the proposed system by observing this occlusion through NIRS data acquisition and offline analysis.

For the experiment, NIR LEDs and photodiodes were attached to a subject's arm in the layout shown in Fig. 5. The experiment was carried out for 5 min. The first minute of the experiment was used as the baseline observation before constriction of the cuff. After 1 min, the pressure was increased to 200 mmHg for 6 s and maintained at this level for 2 min, and then, the contraction was released. Through offline analysis, recorded hemodynamic responses were filtered with a 4th order zero-phase Butterworth 0.2-Hz low-pass filter and normalized responsivities for all channel measurements were derived.

B. Human Subject Studies-Alpha Rhythm Detection Test and Mental Arithmetic Experiment

Although the evaluation and verification of the EEG and fNIRS acquisition system were conducted through the EEG phantom and fNIRS responsivity tests, an experiment involving a human subject also needed to be carried out to evaluate the practical applicability in hybrid EEG/fNIRS monitoring. To this end, an alpha rhythm detection test and a mental arithmetic experiment were carried out. The first is a basic level test to determine whether the proposed system is effective for EEG acquisition. The second is a more challenging experiment to establish whether the system can be used to discern the subtle difference in the EEG and fNIRS signal patterns when the brain engages in non-trivial mental activity, i.e., a mathematical *subtraction* operation.

The alpha rhythm is the most well-known EEG feature that can be easily detected when the user closes his or her eyes. When the eyes are closed, the spectral power of the alpha rhythm band (8–15 Hz) is amplified relative to the other spectral ranges. By comparing the spectral power when the eyes are closed and

when they are not, the detection capability of real EEG features can be verified. One subject participated in this test. Ten trials were performed and one trial consisted of maintaining the eye-open state for 12.5 ± 2.5 seconds and the eye-closed state for 10 seconds. In every transition of the command, a beep sound was used to alert the subject to the change of instruction.

The mental arithmetic experiment is designed to examine the functional brain activation that occurs when subjects are required to carry out non-trivial mathematical operations. During a *subtraction* operation, the brain activation can be observed in both EEG and fNIRS signals. In the EEG signals, the activation appears in the form of an event-related desynchronization (ERD) or event-related synchronization (ERS) [40], known as spectral and suppression and enhancement of the measured EEG signals. The activation in the fNIRS signals is also shown as a hemodynamic difference in oxy- and deoxy-hemoglobin concentration changes (ΔHbO , ΔHbR) [41]. We can investigate these distinctive responses through offline analysis, such as time-frequency analysis of the EEG measurements and time-course analysis of the fNIRS measurements.

Including the subject who participated in the alpha rhythm detection experiment, a total of three subjects voluntarily participated in the mental arithmetic experiment. All subjects (three males, average age: 26.3 ± 1.7 years old) were healthy and had no record of neurological and psychiatric disorders. Each subject was given a summary of the experiment and signed a consent form before their participation started.

The subjects were seated on a chair in front of a 24-inch LCD monitor. Prior to the experiment, pilot signal monitoring was performed to check the adhesion state of the probe set and baseline noise characteristics of the acquired EEG signals. The experiment consisted of two sessions, and each session consisted of 10 trials. In a trial, a white fixation cross was displayed while waiting for the next task period in the first 22.5 ± 2.5 s. In this resting state, the subjects were instructed to gaze at the center cross sign and to refrain from any thinking to maintain a low mental load. During the next task period, the subjects were instructed to *cumulatively* subtract a two-digit random prime number (ranging from 10 to 30) from a three-digit random number in the range 500 to 999 for 20 s. For example, the problem of subtracting 13 from 700 is presented to the subject via a computer screen, i.e., “700 – 13.” The subject had to solve this problem by subtraction inside his/her head. Once he/she arrived at the answer to the problem, $687 = 700 - 13$, they were required to memorize it and to continue to subtract another 13 from the answer, i.e., “687 – 13.” This continued until the end of the task period.

EEG measurements were conducted by attaching 16 dry electrodes to the scalp with the fabricated EEGCAP. To observe the task-related activation in the overall brain areas, 16 electrode positions covering the frontal (Fz, F3, F4, Fc1, Fc2, Fc5, and Fc6), motor/temporal (C3 and C4), and parietal (Pz, P3, P4, Cp1, Cp2, Cp5, and Cp6) regions were carefully chosen in accordance with the international 10–20 system. Reference and bias electrodes were also attached to the skin behind the left and right earlobes, respectively, using disposable wet electrodes. The EEGCAP equipped with dry electrodes was fastened to a

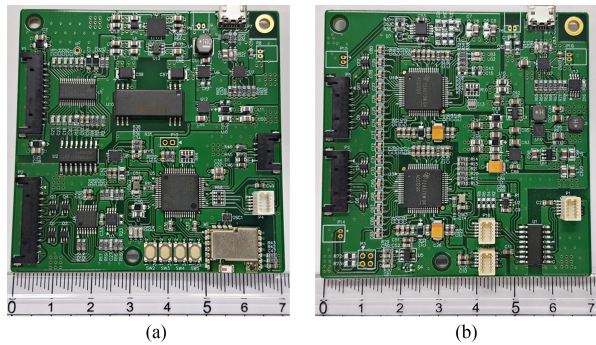


Fig. 7. Images of the (a) main board, and (b) slave board, of the proposed HBM system.

strap on the subject's chest. Two NIR LEDs and six photodiodes were also installed on the forehead using double-sided adhesive tape according to the probe layout in Fig. 5. These installation procedures may take less than 10 minutes, as there is no need for a series of additional preparation processes, such as hair arrangement and scalp abrasion. The EEG and fNIRS measurements acquired by the installed electrodes and photodiodes were simultaneously recorded with an event trigger in real time using MATLAB 2014a.

Offline analysis for the acquired EEG and fNIRS datasets was performed using MATLAB 2014a and EEGLAB toolbox [42]. The EEG datasets were obtained from both the alpha rhythm detection test (one subject participated) and mental arithmetic experiments (three subjects participated). Each EEG dataset was bandpass filtered with a 4th order zero-phase 0.5–40 Hz Butterworth filter. From the filtered dataset, each epoch before and after task onset (−10 to +10 s for the alpha rhythm detection dataset and −15 to +15 s for the mental arithmetic experiment dataset) was extracted based on the recorded event trigger. An EEGLAB built-in function is utilized to investigate ERD/ERS patterns for the time-frequency analysis of the EEG dataset. To visualize the grand-averaged ERD/ERS patterns for each experiment, we averaged the time-frequency decomposition outcomes for all sessions and all subjects who participated. The fNIRS datasets, which comprise the relative concentration changes of oxy-, deoxy- and total hemoglobin (ΔHbO , ΔHbR , and ΔHbT), were only obtained from the mental arithmetic experiments (three subjects participated). A 4th order zero-phase 0.01–0.2 Hz Butterworth bandpass filter was applied to the fNIRS datasets and each epoch was extracted similarly to the EEG pre-processing procedure. Baseline correction of the extracted epoch was performed by subtracting the averaged fNIRS data measured in the resting state between −5 s and 0 s. Identification of the grand-averaged hemodynamic trends during arithmetic operations was also obtained by averaging each of the hemodynamic time courses in the same manner the grand-averaged ERD/ERS patterns were derived.

V. RESULTS

A. System Implementation

Images of the circuit boards of the proposed HBM system for EEG and fNIRS acquisition are shown in Fig. 7. Two

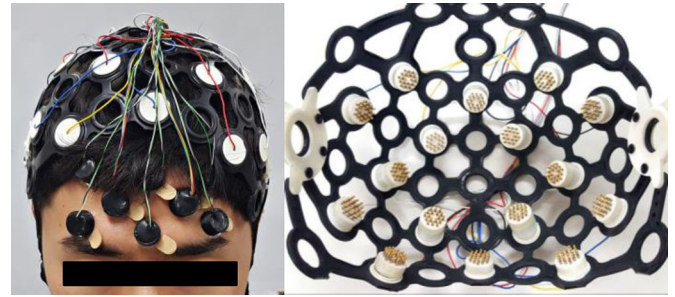


Fig. 8. Image of the complete system comprising the fNIRS probe set and rubber EEGCAP, including 16-channel dry electrodes. The dry electrodes were tightly engaged in the electrode-positioning holes for fixed electrode placement.

TABLE I
CORRELATION COMPARISON FOR ARTIFICIALLY GENERATED EEG RECORDING

NIR LED states	ρ_D	ρ_W	ρ_{DW}
On	0.9422	0.9423	0.9995
Off	0.9433	0.9437	0.9996

four-layered 70 × 70 mm PCBs were fabricated for 16-channel EEG and 8-channel fNIRS acquisition. These boards are connected to each other through the board-to-board connector and are powered by two 2,000 mAh lithium polymer batteries. Sixteen-channel dry electrodes with 18 spring-loaded probes were installed in the EEGCAP, as shown in Fig. 8. Six-channel NIR photodiodes and 2-channel NIR LEDs were also fabricated as depicted in Fig 4(b) and (c). In the experiment involving human subjects, installation of the dry electrodes and the fNIRS probe set was easily accomplished by attaching the set of NIR photodiodes and LED units to the subject's forehead and by requesting the subject to wear the EEGCAP equipped with dry electrodes.

B. Dry-Electrode Evaluation

The correlation coefficients for each electrode comparison set (dry electrode vs. raw signal, wet electrode vs. raw signal, and dry electrode vs. wet electrode) evaluated with the EEG phantom are summarized in Table I. A ρ_{DW} value close to one indicates that the dry and wet electrodes detect almost the same waveform regardless of the activation of the NIR LEDs. This confirms that the dry electrode is capable of obtaining EEG signals without the use of conductive gels and provides almost the same EEG measurement as the wet electrode. Values of ρ_D and ρ_W above 0.9 indicate that the phantom measurements through the dry and wet electrodes are not significantly different from the raw signal data. The slight decrease in the correlation coefficient, compared to ρ_{DW} , is considered to be caused by the error that occurred in the waveform-reduction process using the voltage-divider circuit during artificial EEG generation.

The waveforms recorded by the wet and dry electrodes on the EEG phantom, and the raw EEG signal are shown in Fig. 9(a). The signal recorded at the dry electrode looks like amplified

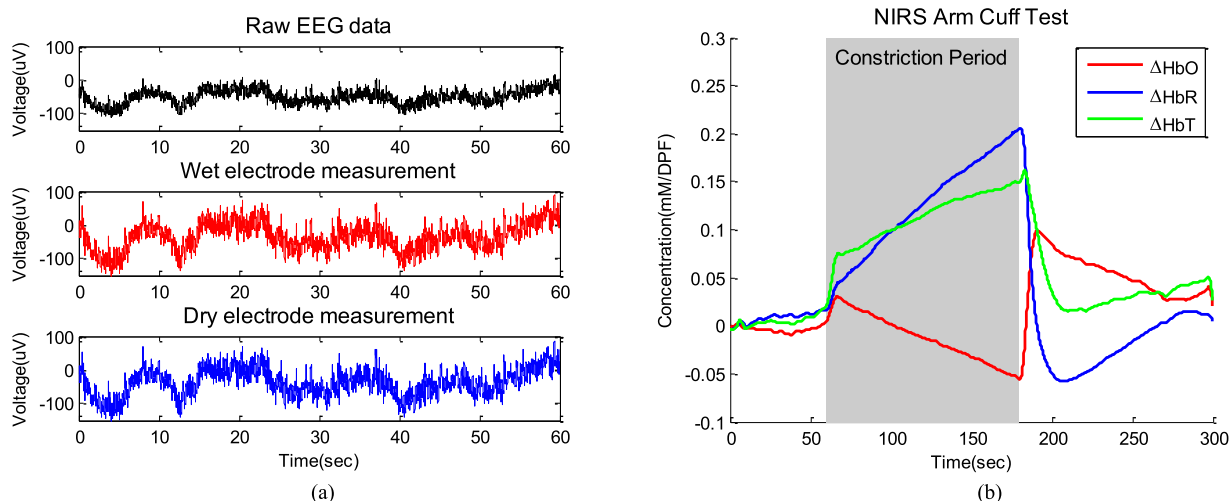


Fig. 9. (a) Comparison of the raw EEG signal and waveforms recorded by the wet and dry electrodes on the EEG phantom, (b) Normalized hemodynamic responses over the eight bio-optical channels with an arterial occlusion experiment.

version of the original signal; however, the overall trend of the waveform is not significantly different according to the correlation coefficient greater than 0.9.

C. fNIRS Response Evaluation

The 8-channel normalized ΔHbO , ΔHbR , and ΔHbT levels were obtained from the offline analysis of the data captured during the arterial occlusion experiment and these results are plotted in Fig. 9(b). All hemodynamic responses converge towards the baseline within ± 0.02 mM / DPF during the first 60 s before contraction of the cuff and increase rapidly over 6 s when the cuff is inflated. When the contraction is complete, the inflowing arterial blood is almost blocked and therefore, the ΔHbO and ΔHbR are linearly diverged until the moment the cuff is released. The slope of the ΔHbO and ΔHbR are $-0.7 \mu\text{M}/\text{DPF}\cdot\text{s}$ and $+1.4 \mu\text{M}/\text{DPF}\cdot\text{s}$, respectively. When the pressure on the cuff is released to allow the arterial blood flow to return, the ΔHbO and ΔHbR dramatically converge and overshooting occurs. After peaking to the opposite overshoot, all hemodynamic responses gradually converge to the steady state. Compared with previous studies [22], [23], in which the same experiment was conducted, the results of this experiment demonstrate that the proposed HBM system is sufficiently responsive to analyze the changes in the hemoglobin concentration.

D. Analysis of Human Subject Studies

The results of the grand-averaged time-frequency analysis results and a comparison of the normalized spectra of the alpha rhythm detection test are depicted in Fig. 10(a) and (c). The vertical dashed lines on the time-frequency analysis plot at 0 seconds denote the onset of the eye-closing task period.

In the alpha rhythm detection test, the event-related synchronization (ERS) pattern evoked by the instruction to close the eyes is clearly indicated with higher spectral power (red zones at Fig. 10(a)) in the alpha rhythm placed in the 8–13 Hz bands compared to the baseline spectral power of -7.5 to -2.5

s. The high spectral power of the beta rhythm in the range of 20–24-Hz at the beginning of the task is considered to be a harmonics related to the high spectral power of the alpha rhythm. The first and second maximum ERS intensities, i.e., 3.74 dB at 11.46 Hz and 2.13 dB at 21.16 Hz, were observed from the dB scale comparison of the normalized spectral graphs at Fig. 10(c). Based on these results, which show that the alpha rhythm associated with closure of the eyes can be detected by using spectral analysis, it is evident that the proposed system can appropriately acquire the general EEG feature signals.

The results of the grand-averaged time-frequency analysis and comparison of the normalized spectra recorded during the mental arithmetic experiments are depicted in Fig. 10(b) and (d). The spectral pattern of the time-frequency analysis was calculated based on the spectral power during the resting state (-15 to -5 s). Compared to the spectral pattern during the task period with those of the alpha rhythm detection test, it is evident that reversed patterns of the spectral perturbation are observed. First and second major event-related desynchronization (ERD) patterns are observed in the alpha rhythm at approximately 10 Hz and in the wide beta rhythm range 18–25 Hz, during the cumulative subtraction task period. The maximum ERD intensity of -2.62 dB at 10.79 Hz in the alpha rhythm range was observed from the dB scale comparison of the normalized spectral graphs in Fig. 10(d). The second highest ERD intensity is -2.10 dB at 19.49 Hz in the beta rhythm range.

The grand-averaged time courses of the concentration changes in oxy-, deoxy- and total hemoglobin (ΔHbO , ΔHbR , and ΔHbT) in the mental arithmetic experiments are plotted in Fig. 11. During the cumulative subtraction task, which is given to the subject to increase the workload level of the brain, we found a clear decreasing trend of ΔHbO . The diminished ΔHbO level is then rapidly restored again to the resting state after the task periods. In contrast, ΔHbR shows a weaker inverse pattern and more delayed response compared to the ΔHbO trend. The lowest ΔHbO is recorded just before the end of the task, whereas the ΔHbR trend continues to increase slightly after the task

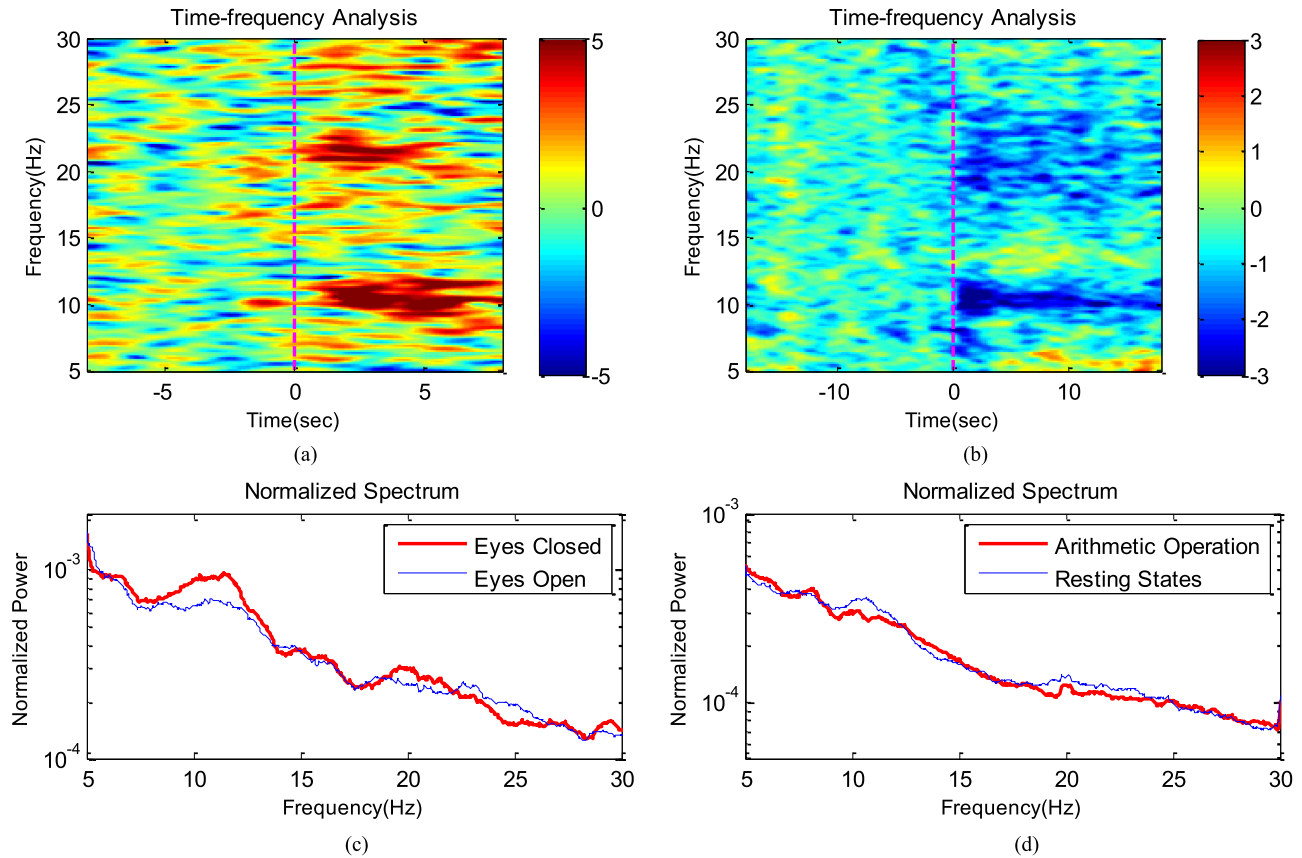


Fig. 10. Results of grand averaged time-frequency analysis (dB scale) for the alpha rhythm detection test (a) and mental arithmetic experiments (b). Vertical dashed lines indicate task onset. Red and blue zones mean the time and frequency ranges associated with high event-related synchronization (ERS) and desynchronization (ERD). Spectral comparisons ((c) eye open states vs. eye closed states, (d) arithmetic operating states vs. resting states).

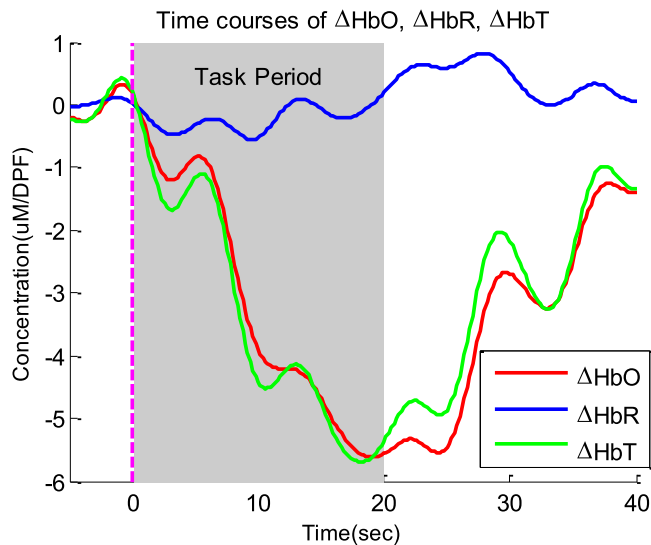


Fig. 11. Grand-averaged time courses of concentration changes in oxy-, deoxy- and total hemoglobin (ΔHbO , ΔHbR , and ΔHbT) for mental arithmetic experiments.

period. This ΔHbR trend begins to decrease belatedly at 8 s after the end of the task. These analysis results show that the ΔHbO pattern much more closely reflects the mental workload

level than the weaker ΔHbR response and the ΔHbT pattern also follows the more dominant ΔHbO trend.

The EEG and fNIRS responses in the mental arithmetic experiments provided the brain activation responses such as the ERD pattern on the alpha and beta rhythm bands and the decreasing trend of the ΔHbO response. These results were compared with those obtained in the previous study [43], in which similar experiments were conducted using commercial equipment. Based on our studies with human subjects, we can conclude that the proposed HBM system has sufficient capabilities to simultaneously monitor EEG and fNIRS signals.

VI. DISCUSSION

The system specifications and key differences compared with the previous studies are summarized in Table II.

[Electrodes] Compared to all previous studies, the proposed system is the first to apply the spring-loaded dry electrodes. More than one hour of continuous EEG monitoring using the conventional wet electrodes is difficult because the conductive gel needs to be replenished every time it becomes dry. Because the dry electrodes enable gel-less EEG acquisition, the quality of the measurement is not degraded and longer experimentation is possible for daily-life monitoring. In addition, it is easy to

TABLE II
COMPARISON OF SYSTEM SPECIFICATIONS AND CONTRIBUTIONS WITH PREVIOUS STUDIES

Comparison category		[21], [22]	[23]	[24]	Proposed
System Specification	# of EEG electrodes	8	16	4	16
	# of LED/PD	8/8	32/4	2/2	8/8
	EEG resolution, ADC architecture	16 bit, Undefined	16 bit, SAR	24 bit, Δ - Σ	24 bit, Δ - Σ
	fNIRS resolution, ADC architecture,	16 bit, Undefined	16 bit, Δ - Σ	24 bit, Δ - Σ	16 bit, SAR
	Volume efficiency	106.6 cm ³ /ch	1.4 cm ³ /ch	1.7 cm ³ /ch	2.0 cm ³ /ch
	Power efficiency, operation hour with 3.7 V 1 Ah battery	150 mW/ch 1.5 h	20 mW/ch 9.25 h	61.6 mW/ch 10 h	18.8 mW/ch 8.2 h
Dry electrode-based EEG acquisition		No	No	Not yet	Yes
Fully isolated circuit design		No	No	No	Yes
Linear regulator-based low-noise power supply		No	No	Yes	Yes
Customizable EEG electrode-positioning structure		Undefined	No	No	Yes

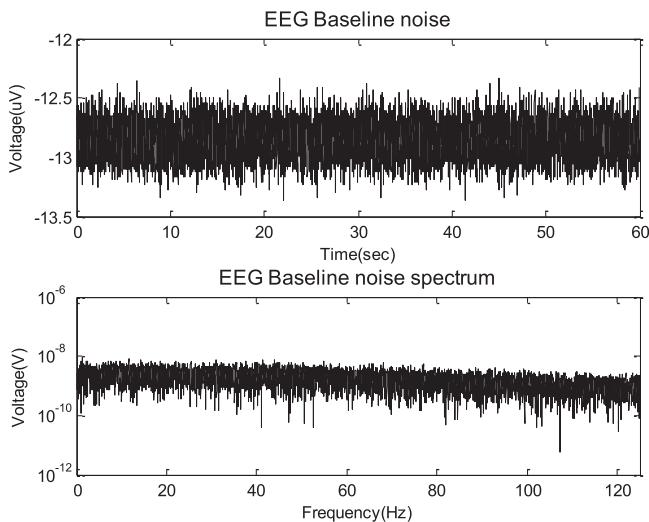


Fig. 12. EEG baseline noise measurements and their spectrum, under the NIR LED activated condition.

install without irritation, a shortened system setup time, and reduced complexity of the experiment.

[Isolated and low-noise circuit design] The implementation of an isolated circuit design is also a first attempt compared to previous studies. Owing to the complete separation of the EEG, fNIRS, and control circuitries with a linear regulator-based low-noise power supply, the proposed system is able to achieve excellent low-noise characteristics for EEG acquisition. During the EEG phantom test, the input-referred noise of the EEG acquisition circuit was evaluated using the built-in input-shorted function of an ADS1299 and its results are shown in Fig. 12. Even with the LED flashing condition, an input-referred noise of $0.141 \mu\text{V}_{\text{RMS}}$ and $1.066 \mu\text{V}_{\text{pp}}$ was measured and no crosstalk component was observed in the spectrum. These results verified

that the proposed system closely achieves the low-noise characteristics of $0.14 \mu\text{V}_{\text{RMS}}$ and $0.98 \mu\text{V}_{\text{pp}}$ (at a sampling rate of 250 SPS and a 24 PGA gain) as specified in the ADS1299 datasheet [29].

[Frontend design] Compared to previous studies on system specifications, the proposed system employs two different kinds of architecture-optimized AFE ICs to simultaneously provide superior EEG resolution and delay-less bio-optical measurement. Because high resolution and continuous sampling are required for EEG measurement, the conversion delay can be considered negligible and the 24-bit Δ - Σ ADC is ideal for use. However, in the case of bio-optical measurements, on-time data acquisition is more important than resolution performance because the sampling is required only for specific predefined time periods along the preprogrammed LED emission schedule. The Δ - Σ ADC-based ADS1299 has a conversion latency of 16 ms at a 250-SPS sampling rate, whereas the SAR ADC-based ADS8688A always maintains a data conversion time of up to $0.85 \mu\text{s}$, regardless of the sampling rate setting. Therefore, this instantaneous sampling characteristic prevents sampling errors in the bio-optical measurements caused by the phase transition of LED activation and ensures system reliability.

[System specifications] The positioning-customizable 16-channel EEG electrodes and 8-channel photodiode detectors indicate that the proposed system is ready for clinical applications for which sufficient spatial resolution is required. However, the estimated volume efficiency (system volume per number of EEG and PD channels) has been slightly reduced due to the implementation of advanced design techniques, such as isolation design and low-noise power supply. Nevertheless, the system size is such that it is still portable ($7 \times 7 \times 1 \text{ cm}^3$) and the power efficiency (power consumption per number of EEG and PD channels) is considerably improved, thus the operation time can be extended to more than 8 hours with a 1-Ah lithium polymer battery. This extended operation time adds the benefit

of a spring-loaded dry electrode that maintains good scalp contact without a conductive gel, facilitating hybrid brain monitoring in out-of-lab situations.

[Limitations and future development] One of the limitations is that it is difficult to obtain fNIRS measurements in various brain areas because the NIRS probes can only be attached to the hairless scalp. Overcoming this challenge necessitates the design of a probe structure that can be adhered to the scalp by applying pressure with a stretchable structure such as the spring-loaded structure of a dry electrode.

The achievement of stable EEG quality in an actual out-of-lab situation requires motion artifacts to be removed from EEG measurements. Therefore, a movement monitoring function is required, and it can be implemented by integrating a MEMS-based inertial sensor. A continuous impedance check function is also required to monitor the adhesion of the electrode in real time, because the adhesion pressure of the electrode has a significant effect on the quality of the acquired EEG signal. This function can be implemented by utilizing the built-in lead-off detection function with the ADS1299.

VII. CONCLUSION

In this study, a hybrid brain monitoring system for simultaneous acquisition of 16-channel EEG and 8-channel fNIRS has been proposed. A single low-power microcontroller unit synchronously controls two kinds of architecture-optimized AFE ICs to achieve fully synchronized data acquisition. Employing Δ - Σ ADC-based ADS1299 and SAR ADC-based ADS8688A simultaneously, the proposed system achieves 24-bit EEG resolution and delay-less ($<0.85 \mu\text{s}$) reliable fNIRS measurements. A fully isolated design, which completely separates the ground plane of each circuit section by using digital isolators and an isolated DC-DC converter, physically blocks inter-circuit interference. The isolated design applied with a linear regulator-based low-noise power supply improves system reliability and noise immunity for EEG/fNIRS measurements. Moreover, the use of spring-loaded dry electrodes and EEGCAP shortens system-wearing time and continuously provides stable EEG quality. It will allow longer experiments for out-of-lab applications. The acquisition of EEG and fNIRS measurements was evaluated by conducting an EEG phantom test using artificially generated EEG signals and an arterial occlusion experiment. Additionally, an alpha rhythm detection test and mental arithmetic experiments were performed to assess the practical capabilities of the proposed system for human subject studies. The grand-averaged results of the time-frequency analysis for EEG measurements and time courses for NIRS measurements verified that the proposed HBM systems are suitable for use in real BCI applications.

REFERENCES

- [1] J. R. Wolpaw *et al.*, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.
- [2] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors (Basel)*, vol. 12, no. 2, pp. 1211–1279, Jan. 2012.
- [3] J. D. R. Millán *et al.*, "Combining brain–computer interfaces and assistive technologies: State-of-the-art and challenges," *Frontiers Neurosci.*, vol. 4, Sep. 2010, Art. no. 161.
- [4] M. D. Murphy *et al.*, "Current challenges facing the translation of brain computer interfaces from preclinical trials to use in human patients," *Frontiers Cellular Neurosci.*, vol. 9, Jan. 2016, Art. no. 497.
- [5] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain Topography*, vol. 23, no. 2, pp. 194–198, 2010.
- [6] G. Pfurtscheller *et al.*, "The hybrid BCI," *Frontiers Neurosci.*, vol. 4, Apr. 2010, Art. no. 42.
- [7] N. Weiskopf, "Real-time fMRI and its application to neurofeedback," *NeuroImage*, vol. 62, no. 2, pp. 682–692, Aug. 2012.
- [8] T. Nguyen *et al.*, "Utilization of a combined EEG/fNIRS system to predict driver drowsiness," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 43933.
- [9] A. Machado *et al.*, "Detection of hemodynamic responses to epileptic activity using simultaneous Electro-Encephalography (EEG)/Near Infra Red Spectroscopy (NIRS) acquisitions," *NeuroImage*, vol. 56, no. 1, pp. 114–125, May 2011.
- [10] N. Naseer and K.-S. Hong, "fNIRS-based brain-computer interfaces: A review," *Frontiers Human Neurosci.*, vol. 9, Jan. 2015, Art. no. 3.
- [11] S. Amiri *et al.*, "A review of hybrid brain-computer interface systems," *Adv. Human-Comput. Interaction*, vol. 2013, 2013, Art. no. e187024.
- [12] S. Fazli *et al.*, "Enhanced performance by a hybrid NIRS–EEG brain computer interface," *NeuroImage*, vol. 59, no. 1, pp. 519–529, Jan. 2012.
- [13] A. P. Buccino *et al.*, "Hybrid EEG-fNIRS asynchronous brain-computer interface for multiple motor tasks," *PLOS ONE*, vol. 11, no. 1, 2016, Art. no. e0146610.
- [14] S. Ge *et al.*, "A brain-computer interface based on a few-channel EEG-fNIRS bimodal system," *IEEE Access*, vol. 5, pp. 208–218, 2017.
- [15] R. Li *et al.*, "Enhancing performance of a hybrid EEG-fNIRS system using channel selection and early temporal features," *Frontiers Human Neurosci.*, vol. 11, Sep. 2017, Art. no. 462.
- [16] A. M. Chiarelli *et al.*, "Deep learning for hybrid EEG-fNIRS brain-computer interface: application to motor imagery classification," *J. Neural Eng.*, vol. 15, no. 3, 2018, Art. no. 036028.
- [17] Y. Tomita *et al.*, "Bimodal BCI using simultaneously NIRS and EEG," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1274–1284, Apr. 2014.
- [18] M. J. Khan *et al.*, "Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface," *Frontiers Human Neurosci.*, vol. 8, Apr. 2014, Art. no. 244.
- [19] H. Aghajani *et al.*, "Measuring mental workload with EEG+fNIRS," *Frontiers Human Neurosci.*, vol. 11, Jul. 2017, Art. no. 359.
- [20] R. J. Cooper *et al.*, "Design and evaluation of a probe for simultaneous EEG and near-infrared imaging of cortical activation," *Phys. Med. Biol.*, vol. 54, no. 7, pp. 2093–2102, 2009.
- [21] E. Lareau *et al.*, "Multichannel wearable system dedicated for simultaneous electroencephalography /near-infrared spectroscopy real-time data acquisitions," *J. Biomed. Opt.*, vol. 16, no. 9, 2011, Art. no. 096014.
- [22] M. Sawan *et al.*, "Wireless recording systems: From noninvasive EEG-NIRS to invasive EEG devices," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 2, pp. 186–195, Apr. 2013.
- [23] J. Safaie *et al.*, "Toward a fully integrated wireless wearable EEG-NIRS bimodal acquisition system," *J. Neural Eng.*, vol. 10, no. 5, 2013, Art. no. 056001.
- [24] A. von Lüthmann *et al.*, "M3BA: A mobile, modular, multimodal biosignal acquisition architecture for miniaturized EEG-NIRS-based hybrid BCI and monitoring," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 6, pp. 1199–1210, Jun. 2017.
- [25] S. Lee *et al.*, "Dry electrode design and performance evaluation for EEG based BCI systems," in *Proc. Int. Winter Workshop Brain-Comput. Interf.*, 2013, pp. 52–53.
- [26] L.-D. Liao *et al.*, "Design, fabrication and experimental validation of a novel dry-contact sensor for measuring electroencephalography signals without skin preparation," *Sensors*, vol. 11, no. 6, pp. 5819–5834, May 2011.
- [27] Y. M. Chi *et al.*, "Dry-contact and noncontact biopotential electrodes: Methodological review," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 106–119, Oct. 2010.
- [28] S. Karthik and B. Mark, "Analog front-end design for ECG systems using delta-sigma ADCs," Texas Instrum., Dallas, TX, USA, Appl. Rep. SBAA160A, Mar. 2009.
- [29] ADS1299 Datasheet. Texas Instrum., Dallas, TX, USA, Jul. 2012.
- [30] ADS8688 Datasheet. Texas Instrum., Dallas, TX, USA, Jul. 2015.
- [31] W. Kester and Analog Devices, inc, Eds., *Data Conversion Handbook*. Amsterdam, The Netherlands: Elsevier, 2005.

- [32] B. C. Baker, "Conversion latency in delta-sigma converters," Texas Instrum., Dallas, TX, USA, 2007.
- [33] S. Pithadia and S. More, "Grounding in mixed-signal systems demystified, Part 1," Texas Instrum., Dallas, TX, USA, 2013.
- [34] W. Alexander and P. Alexander, "High-speed layout guidelines," Texas Instrum., Dallas, TX, USA, Appl. Rep. SCAA082, Nov. 2006.
- [35] H. Zumbahlen and Analog Devices, inc, Eds., *Linear Circuit Design Handbook*. Amsterdam, The Netherlands: Elsevier, 2008.
- [36] G. Strangman *et al.*, "Non-invasive neuroimaging using near-infrared light," *Biol. Psychiatry*, vol. 52, no. 7, pp. 679–693, Oct. 2002.
- [37] J. G. Kim and H. Liu, "Variation of haemoglobin extinction coefficients can cause errors in the determination of haemoglobin concentration measured by near-infrared spectroscopy," *Phys. Med. Biol.*, vol. 52, no. 20, pp. 6295–6322, 2007.
- [38] L. Kocsis *et al.*, "The modified beer–Lambert law revisited," *Phys. Med. Biol.*, vol. 51, no. 5, p. N91–N98, 2006.
- [39] B. Blankertz *et al.*, "The BCI competition III: Validating alternative approaches to actual BCI problems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, Jun. 2006.
- [40] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.
- [41] G. Pfurtscheller *et al.*, "Focal frontal (de)oxyhemoglobin responses during simple arithmetic," *Int. J. Psychophysiol.*, vol. 76, no. 3, pp. 186–192, Jun. 2010.
- [42] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.
- [43] J. Shin *et al.*, "Evaluation of a compact hybrid brain-computer interface system," *BioMed. Res. Int.*, vol. 2017, 2017, Art. no e6820482.

Received January 23, 2019, accepted February 18, 2019, date of current version April 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901292

Fractional-Order Integration Based Fusion Model for Piecewise Gamma Correction Along With Textural Improvement for Satellite Images

HIMANSHU SINGH¹, (Student Member, IEEE), ANIL KUMAR¹, (Member, IEEE),
L. K. BALYAN¹, AND HEUNG-NO LEE², (Senior Member, IEEE)

¹Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur 482005, India

²Gwangju Institute of Science and Technology, Gwangju 500712, South Korea

Corresponding author: Heung-No Lee (himanshu.iiitj@gmail.com)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MSIP) (NRF-2018R1A2A1A19018665).

ABSTRACT Fractional-order integration (FOI) and its beauty of optimally ordered adaptive filtering for image quality enhancement are latently too valuable to be casually dismissed. With this motivation, a new Riemann–Liouville fractional-order calculus-based spatial-masking methodology is proposed in this paper in association with counterbalanced piecewise gamma correction (PGC). A generalized FOI-based mask is also suggested. This mask is negatively augmented with the original image for harvesting texture-based benefits. PGC is just employed through a constructive association of both kinds of reciprocally dual gamma values ($\gamma_1 = \gamma$ and $\gamma_2 = 1/\gamma$, $\forall \gamma > 1$), which leads to optimally desired enhancement when applied in a weighted counter-correction manner. Efficiently improved and recently proposed opposition-based learning inspired sine–cosine algorithm is employed in this paper, along with a newly framed fitness function. This fitness function is devised in a novel manner by taking care of textural as well as non-textural details of the images. In this paper, especially for dark images, 130% increment is achieved over the input contrast along with the simultaneous 147% increment in the discrete entropy level and 22.8% increment in the sharpness content. Also, brightness and colorfulness are reported with 130% and 196.4% increased with respect to the input indices, respectively. In addition, the textural improvement is advocated in terms of desired comparative reduction of gray-level co-occurrence matrix-based metrics, namely, correlation, energy, and homogeneity, which are suppressed by 25.6%, 72.5%, and 21.8%, respectively. This performance evaluation underlines the excellence and robustness for imparting proper texture as well as edge preserved (or efficiently restored) image quality improvement.

INDEX TERMS Fractional-order (FO) masking filter, fractional-order integration (FOI), Riemann–Liouville (RL) definition, sine cosine algorithm (SCA), opposition-based learning (OBL), gray-level co-occurrence matrix (GLCM), quality enhancement, optimal mask designing, two-dimensional (2-D) adaptive filtering, piecewise-gamma correction (PGC).

I. INTRODUCTION

Remotely acquired digital imagery and its various forms keep on laying the core and firm foundation of today's technological era, and hence it is quite implicit that necessity and importance of quality enhancement and desired information restoration is having the prime concern. Various integers' based mathematical suggestions got appreciation in the last two decades, but fractional-order calculus-based

mathematical advancements in image processing applications are still fascinating in one form or the other. Fractional-order calculus (FOC) is a kind of eternal source of analytical processing power. FOC [1], [2], since the historical day of its paradoxical invention, evolved from a very popular Leibniz–L'Hôpital technological conversation, as on September 30, 1695; the theory is still gloriously blooming day-by-day for drawing significant application based consequences. FOC is now being more widely accepted in current technological trends. The involvement of classical, as well as integer-based calculus in anthropological scientific advancement,

The associate editor coordinating the review of this manuscript and approving it for publication was Sara Dadras.

is indispensable. Hence, it is very hard to say about “fractional-order or non-integral order calculus” even after 323 years of discussion and its consequent advancement, that in which domain of science and technology, FOC is incapable to facilitate the corresponding technological advancement. FOC has recently been applied in various areas of engineering, science, finance, applied mathematics, and bio-engineering with its remarkable success. Contemporary advancements in all technological spheres for social welfare cannot be imagined without signals and their “on-demand or application-specific “contextual processing, in one form or the other. A large amount of work has already been focused on image enhancement through FO differentiation based masking [3]. More robustness can be imparted for pre-existing methodologies by an organized and efficient involvement of swarm intelligence in association with the fractional-differential approach. Now through this paper, it’s the time of debut for fractional-order integral based adaptive filtering for image quality enhancement. With such an objective, an interesting approach is proposed by suggesting a fusion based framework by employing optimally ordered fractional-order integration for gamma corrected image enhancement application for almost all kind of images.

Fundamentally, image visualization characteristics can be categorized as spectral, textural and contextual features for almost all kinds of remotely sensed images. Spectral features account for general tonal variations corresponding to each band in the visible and/or infrared region of the electromagnetic spectrum. Structural variation and corresponding inter as well as intra-organization of the contextual surfaces of the images can be better identified as the texture of the image. Texture, in one form or the other, accounts for most of the information related to all kinds of preprocessing tasks associated with remotely sensed imagery. It is also responsible for identifying the objects or regions of interest in an image. Also, intensity levels and their sharpness also contribute to overall behavioral characteristics of the image. Hence, in general, quality enhancement can be seen as collective coordination of improved spatial, contextual, textural and edge-dependent image features. Initially, for various years, the histogram of the acquired image has been utilized for image enhancement.

Initially, researchers got highly fascinated by histogram equalization (HE) based approaches, especially, general histogram equalization (GHE). GHE flattens and enlarges the dynamic range of image’s histogram by remapping the gray but its performance is inadequate due to the issue of mean-shift and also, it is quite unable to preserve the local spatial features of an image [4]. Due to this, the sole attention of the researchers got shifted to histogram-distribution along with local histogram modifications, and also towards their corresponding advantages. Therefore, several algorithms were developed for image enhancement based on local histogram modifications. Later on, median-mean dependent sub-image-clipped HE (MMSICHE) [5] was proposed for image enhancement along with further bisecting both of sub-histograms on the basis of the median count of the

corresponding pixels in both sub-histograms. Contrast limited adaptive histogram equalization (CLAHE or ADAPHE [6]), has been also proposed as a very efficient variant of GHE. ADAPHE operates on small regions in the image in a tile-wise manner rather than the entire image. Later all neighboring tiles are combined using 2-D bilinear interpolation to eliminate artificially induced boundaries.

Blind-reliability over HE (local and global both) based enhancement approaches is not advisable because of their tendency to impart uniform intensity distribution without knowing the behavior of input image; and hence, gamma correction based methods were suggested by various researchers for contrast enhancement. Gamma correction was initially applied directly on image pixels in its original domain. Afterward, transform domain gamma correction based on wavelets and filter-banks based transformations were suggested with a common problem of manual tuning for the desired and relevant gamma value which is a very tedious and a trivial kind of task. It was found that applying gamma correction through histogram is easier and quite efficient rather than applying gamma correction directly on the corresponding image itself. Later on, adaptive gamma correction with weighting distribution (AGCWD) [8] and its various effectively improved versions like [9]; were better performed for this objective, where desired contrast enhancement is imparted by utilizing a gamma value-set of size 2^{L-1} for a corresponding L bit image, which itself is derived by using cumulative distribution of the intensity values present in low contrast input image.

The averaging histogram equalization (AVHEQ) approach [7] was proposed by combining linear color channel stretching, histogram averaging, which is followed by consequent one to one mapping of the intensity levels. In the same context, HE based optimal profile compression (HEOPC) [10] and HE with maximum intensity coverage (HEMIC) [11] were proposed, where the objective is to harvest more and more intensity levels in an exhaustive manner. Afterward, the intensity and edge-based adaptive unsharp masking filter (IEAUMF) [12] based enhancement have been also proposed by employing the unsharp masking filter for edge augmentation. Although, these approaches are focused over the harvesting of more and more intensity levels which generally leads to kind of smoothing and hence, improvement of textural details is not focused on it, and sometimes leads to the unbalanced exposure also.

Most of the methodologies are based on parallel pipelined methods by framing collective coordination of various efficient operations. Most of the conventional methodologies are based on histogram redistribution mechanism, in one way or the other. Usually, histogram-redistribution doesn’t care much about textural details. Also, the spatial pixel-wise orientations are not considered while (one-dimensional) histogram based processing. A second serious issue is the appearance of saturation effects in case of gamma correction which leads to over-enhancement and under-exposed patches in the enhanced version of the image. Both of these issues are tried

to be rectified, as in the proposed approach, gamma correction is employed in a counter-correction based balanced manner by involving both gamma expanded and gamma compressed channels. Piecewise gamma correction (PGC) is proposed in an effective manner by suggesting a fusion based framework by the constructive involvement of reciprocal gamma values which leads to interim compressed as well as interim expanded images. This is initially proposed as a dual correction for GHE and termed as PGC based HE (PGCHE) where the results are not so enriched due to lack of textural treatment. So, it was further improved for PGC based textural enhancement by taking the benefits of positive augmentation through fractional order differentiation based adaptive filtering. Later, it is found more effective than in place of FOD based masking, if unsharp masking is modified by applying fractional order integration based adaptive filtering in a negative augmentation manner, the outcomes are much better.

In this manuscript, idea is to introduce the benefits of FOI based adaptive two-dimensional filtering for overall textural and non-textural benefits. In addition, parallel “FOI based textural improvement and edge restoration” is also suggested. The prime contribution in this manuscript is an inclusion of highly adaptive, multidimensional optimally weighted framework through association of piecewise gamma correction along with the fractional-order integration based negatively augmented unsharp masking for texture highlighted, overall quality enhancement for remotely sensed imagery. An opposition based learning inspired sine-cosine algorithm based optimization mechanism termed as OBL-SCA model is framed in accordance with the concerned tuning problem for free parameters; utilizing a newly framed objective function which is designed by using intensity-dependent and GLCM based fidelity parameters. The fitness function is specially framed by considering textural and illumination improvement for dark images. As, FO masking is itself not sufficient for dark images, hence, for proper exposure shift, counter-balanced piecewise gamma correction is also entangled with it. Hence, idea is to employ gamma correction in a dual manner by constructive fusion of PGC through employing gamma correction in a dual manner by a constructive amalgamation of interim channels.

In this context, reciprocally dual values of gamma are utilized for employing the desired enhancement. The closed form pre-existing approaches are usually unable to serve the purpose efficiently. Hence, a constructive combination of artificial intelligence or machine learning inspired optimization principles along with classical optimization approaches can be applied. Particle Swarm Optimization (PSO) [13], Artificial Bee Colony (ABC) optimization [14], Moth-Flame Optimization (MFO) [15], Sine Cosine Algorithm (SCA) [16], etc. have been also developed by imitating various nature-inspired analogies. In general, most of such kinds of optimization strategies are suffering from common issues of local trapping, which can be eradicated efficiently if governed by some kind of artificial learning based intelligence.

Considering the above key-points in mind, sine-cosine optimizer using opposition-based learning [17] is integrated with the proposed framework. In this manner, better converging behavior is proposed by planning a gradient-based exploration and exploitation.

The core contribution for achieving overall quality enhancement, in this manuscript can be point-wise identified as:

- A novel framework of FOI based masking is proposed for adaptive image filtering which leads to analogous FO unsharp masking.
- Collective benefits of newly identified PGC and FO integral are justified and proposed in this paper for textural improvement along with on-demand dual gamma-correction.
- In one manner, textural and non-textural content of the image is separately processed as per the “on-demand” basis and later an optimally framed fusion is employed for collective contribution.
- A newly framed fitness function or the cost function is formulated for deciding the tuning parameters to make the approach highly adaptive for a diverse blend of images.
- This newly proposed cost function is designed by blending the significant fidelity parameters in an effective and robust manner for highlighting the effective convergence of the algorithm.

The remaining manuscript is drafted as follows: after brief literature survey and basic introduction in section 1; section 2 explains the problem formulation; section 3 explains the proposed Piecewise Gamma Corrected (PGC) RL-FOI based Masking (PGCRLFOIM) algorithm followed by its stepwise framework. Later, section 4 deals with the experimentation followed by corresponding results and discussion; and in section 5, conclusions are drawn.

II. PROBLEM FORMULATION

The central idea is to impose both of these corrections on the fractionally masked and negatively augmented image which is just an analogy of negative augmentation of the fractionally derived low pass filtered/smoothened intensity channel. Thus, sharpness behavior of the texture and edge content can be highlighted and finally corrected through a reciprocal set of gamma values in a dual balanced manner. Hence, both of these individually account for keeping the processed intensity channels in the permissible intensity ranges in a constructive manner and hence, fruitful exploration for all of the intensity values can be done, by proper avoidance for the saturation due to an accumulation of the pixel values in extreme-end bins of the histogram. Hence, information loss can be minimized by avoidance of over-saturated patches as well as improperly exposed regions. Thus, an attempt is made for the proposal of a complete framework is by collective contribution of textural, spectral and contextual base quality improvement. Hence, successful harvesting of the optimal masking which is analogs to low pass filtering of fractional order and its

negative augmentation leads to a kind of adaptive enhancement based masking which is later counter corrected by both gamma values by collective piecewise contribution of both values of reciprocal gamma set ($\gamma_1 = \gamma$ and $\gamma_2 = 1/\gamma$, $\forall \gamma > 1$), and thus leads to interim enhanced and interim compressed gamma value set.

III. PROPOSED METHODOLOGY

Fundamental instinct behind this approach is to design an efficient and highly adaptive behavior-dependent end-to-end optimal framework for overall image quality improvement. Both the textural as well as non-textural image details are managed along with proper edge restoration. A generalized N-ordered FOI based mask framing strategy is suggested using rotation-invariant uniformity principle; so that, whenever it is assimilated along with piece-wise gamma correction, it leads to better quality improvement for the texture of the poorly-illuminated digital image. Usually, remotely sensed, imperfectly illuminated satellite images come under this category; however, this proposed framework is also applicable for other similar kinds of rich-grained images.

A. PIECEWISE GAMMA CORRECTION

Piecewise gamma correction (PGC) is applied by an optimal evaluation of gamma-compressed and corresponding gamma-expanded channel. It should be notified that both of these counter-gamma values are reciprocal of each other and hence, advantageous for lower-end as well as higher-end intensity correction. This, in turn, results into restriction of oversaturated as well as sub-enhanced patches and hence, the objective of more and more information exploration get fulfilled. When intensity values of the input image (I_{in}) are normalized from zero to unity, if they are exponentially employed by a gamma value more than unity, it leads to intensity compression and hence, can be termed as gamma compressed intensity channel (I_{gcp}) as [2], [3]:

$$I_{gcp} = (I_{in})^\gamma, \quad \gamma > 1, \quad (1)$$

Correspondingly, its dual can be evaluated by the reciprocal of the gamma value as in Eq. (1). It leads to shifting towards the dark end of the histogram's abscissa, and hence, leads to a compressed kind of intensity distribution. Such interim channel is utilized basically for imparting counter correction for over-enhanced or saturated bright patches. Complimentary to this, as per the Eq. (2), the intensity values of the input channel get shifted towards the bright end of the histogram's abscissa. Hence, it leads to expansion of intensity distribution towards the bright end. This leads to counter correction for dark patches and also for boosting up for less illuminated image-patches. The interim gamma corrected expanded (I_{gex}) channel can be evaluated as [2], [3]:

$$I_{gex} = (I_{in})^{1/\gamma}, \quad \gamma > 1, \quad (2)$$

B. RL FRACTIONAL ORDER INTEGRATION

For adaptive kind of isolation of non-textural ingredients from the image in highly optimal fashion, benefits of the FOI are derived by following the Euclidean domain based RL definition for FOI. This is analogously derived from integer based fundamentals. Originally inspired by the Cauchy integration for the concerned analytic function, primarily for any complex plane; symbolically it can be justified as [1]:

$$D^n f(t) = \frac{1}{(n-1)!} \int_a^t (t-u)^{n-1} \cdot f(u) du, \quad n \in \mathbb{N} \quad (3)$$

Thus, according to the RL definition, as suggested, the Cauchy integral formula can be directly extended to enter the fractional-order calculus domain. By convention, it is required that $f(t)$ must be a causal function. In other words, $f(t)$ must identically be vanishing for $t < 0$, which occurs intuitively, by default, in case of digital images, since negative pixel intensities are insignificant. From the standard formalization of ν -ordered RLFOI for any function $f(t)$ in the interval $[0, t]$ and $[-\infty, t]$ as follows:

$$J^\nu f(t) = \frac{1}{\Gamma(\nu)} \int_{-\infty}^t (t-u)^{\nu-1} \cdot f(u) du, \quad \nu > 0, \quad (4)$$

The duration of signal $f(t)$ is $[0, t]$ and $\Gamma(\cdot)$ is the Gamma Function defined as:

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad (5)$$

Discrete-time equivalent of J^ν can be derived by framing a discrete time FO kernel $I^\nu(n)$ as:

$$I^\nu(n) = \begin{cases} \frac{n^{\nu-1}}{\Gamma(\nu)}, & n > 0 \\ 0, & n \leq 0 \end{cases}, \quad (6)$$

$$A = I^\nu(1); B = I^\nu(2); C = I^\nu(3); D = I^\nu(4); E = I^\nu(5); \quad (7)$$

$$\begin{aligned} & [A, B, C, D, E, \dots] \\ & = \left[\left(\frac{1^{\nu-1}}{\Gamma(\nu)} \right), \left(\frac{2^{\nu-1}}{\Gamma(\nu)} \right), \left(\frac{3^{\nu-1}}{\Gamma(\nu)} \right), \left(\frac{4^{\nu-1}}{\Gamma(\nu)} \right), \dots \right], \end{aligned} \quad (8)$$

C. PROPOSED PGCRIFOI FRAMEWORK

Rather than following the image-dependent FO differential masking based edge-augmentation concept (as earlier proposed by the same authors), the idea is to suppress the non-textural details first, which is followed by overall boost up of the processed image. Thus, the local spatial intensity saturation, as well as over-brightness enhancement, can be counter-attacked. For this purpose, the RL-FOI based odd-dimensional symmetric mask is framed in a generalized manner. To avoid the further higher order complexity, a 5×5 sized mask is derived by following the fundamental FO integral

calculus. In addition, the mask coefficients are arranged to maintain the proposed mask rotational invariant. The conceptual analysis is to suppress the non-textural content of the image first; and then, enhancement for the resulted image in a fashion so that PGC can be imposed in an optimal manner. Fractional-order 1-D filtering can be extended to a 2-D image matrix, and hence, a set of fractional-order partial differential equations w.r.t. x and y -direction can be expressed as [3]:

$$\begin{aligned}
 I_{x^v}^v f(x, y) &\approx \left(\frac{1^{v-1}}{\Gamma(v)}\right) \cdot f(x, y) + \left(\frac{2^{v-1}}{\Gamma(v)}\right) \cdot f(x-1, y) \\
 &+ \left(\frac{3^{v-1}}{\Gamma(v)}\right) \cdot f(x-2, y) + \dots + \left(\frac{(n+1)^{v-1}}{\Gamma(v)}\right) \\
 &\cdot f(x-n, y), \tag{9}
 \end{aligned}$$

$$\begin{aligned}
 I_{y^v}^v f(x, y) &\approx \left(\frac{1^{v-1}}{\Gamma(v)}\right) \cdot f(x, y) + \left(\frac{2^{v-1}}{\Gamma(v)}\right) \cdot f(x, y-1) \\
 &+ \left(\frac{3^{v-1}}{\Gamma(v)}\right) \cdot f(x, y-2) + \dots + \left(\frac{(n+1)^{v-1}}{\Gamma(v)}\right) \\
 &\cdot f(x, y-n), \tag{10}
 \end{aligned}$$

An RL definition based FO 5x5 mask is created by maintaining a similar kind of gradient behavior in almost all eight directions. These directions can be viewed w.r.t. the center pixel based balanced orientation at angles of 0, 45°, 90°, 135°, 180°, 225°, 270°, 315° and 360°, respectively. In this work, the behavior of smoothing filter or blurring image filter is extended by framing this FO low pass filter kind of mask. The elements of this 5x5 mask are normalized so that sum of all elements remains unity. Masks of 3x3 and 7x7 size have been also tested for this, but a trade-off is settled for 5x5 size. Integer-ordered masks are usually employed for detection of the smooth content of images by complete exclusion of major and minor edges. This extent of inclusion or exclusion depends on the order of FOI mask which itself acts as a 2-D adaptive filter to extract the low-frequency content of the image under consideration. Idea is to extract the non-edge content of the image optimally based on the adaptively decided order, and finally, deduct this information content from the input image. The later emphasis on the whole image leads to a highlighted textural content of the image. Masking is employed through a symmetric mask using only the first three coefficients as shown in Eq. (11). Fundamentally, 2-D linear filtering is done by convolving these filters from left to right for all rows individually and then by convolving these filters from top to bottom for all columns, similarly.

$$H_x = H_y = \frac{0.125}{(A+B+C)} \begin{pmatrix} C & 0 & C & 0 & C \\ 0 & B & B & B & 0 \\ C & B & 8A & B & C \\ 0 & B & B & B & 0 \\ C & 0 & C & 0 & C \end{pmatrix}, \tag{11}$$

It has to be taken care that, size of the convolved product must be the same as the size of input image channel matrix. FO changes and correspondingly adaptive nature of these masks can be identified through the spectral behavior for these masks more precisely. Next step is to compute the fractional-order integration based negatively augmented masking for deriving a third interim channel, which is a texture improved version of the input channel. This analogy is inspired from unsharp masking mechanism, with its negatively augmented version and hence, resembles the suppressed low pass filtered version of the image by optimally ordered version of the FOI by following its RL definition. This interim channel is evaluated as:

$$I_{fimf} = I_{in} + k \cdot \lambda \cdot (I_{in} - I_v), \tag{12}$$

2-D convolutional filtering of the input channel (I_{in}) by employing the RL FOI mask (H) can be understood as:

$$I_v = H \otimes I_{in}, \tag{13}$$

Later on, k which is the scaling factor for adaptive augmentation (can be assumed 0.5). In order to accomplish the design objective, a hyperbolic profile, λ is adopted, as:

$$\lambda = 0.5 [1 + \tanh(3 - 6(|I_v| - 0.5))], \tag{14}$$

The profile is decided by an adaptive contribution of the magnitude of negatively augmented; FOI based filtered channel's pixels at the corresponding image coordinate. Here, $\tanh(\pm 3) = \pm 0.995 \approx 1$ is sufficient to approximate a unity scale. The rationale in defining the input based profile also applies here. Since the corresponding magnitudes are bounded between ± 1 , a modification is made as $6 \times (|I_v| - 0.5)$ in order to bind the profile coefficients within ± 3 . The multiplication by 6 is used as a normalization accounting for the doubled edge magnitudes due to the possible two edge or textural polarities. An enhanced version of the image can be evaluated by weighted and collective fusion of all three interim enhanced images. The parameters p and q are solely responsible for the weighted amalgamation of these channels, as:

$$I_{en} = \left(\frac{p}{1+q}\right) \cdot I_{gcp} + \left(\frac{1-p}{1+q}\right) \cdot I_{gex} + \left(\frac{q}{1+q}\right) \cdot I_{foimf}, \tag{15}$$

Balanced benefits of both gamma compression and gamma expansion over the image under consideration in a reciprocally framed counter-correction manner; and hence, for weighted involvement in a pixel-wise augmentation manner, the parameter p is introduced for imparting intensity exposure based enhancement. Later on, for textural and edge improvement, a novel FOI based negative augmentation scheme is suggested and hence, in accordance with it, third interim channel is also framed. Further, its weighted involvement is also a matter of prime concern, and hence, the parameter q is involved. The role of the q must be framed in such a manner that a balanced involvement for gamma corrected interim channels along with the proposed

negatively augmented RL-FOI mask based filtered interim-channel I_{foimf} can be framed. Thus, collective involvement of texture enhanced version along with rest of the channels is premier concern and hence, a highly balanced provision is made for it as follows. Here, the gamma value is kept more than unity and hence, its reciprocal leads to value in the range (0, 1]. Balanced and weighted framework contribution of all interim channels can be maintained by keeping the value of p in the range of [0, 1] and q can be varied throughout the positive range. Assuming a case when $q \rightarrow 0$ it leads to the ignorance or absence of texture improved channel and such type of situation arises when the acquired image is of smooth nature. Contrary to this, a larger value of q stands for highly textured images. When, q is unity, it indicates half of the contribution is due to the presence of I_{foimf} and remaining both gamma based channels are confined to rest of the half share. According to above justification, q must be varied in the range [0, ∞), but while looking for its practical aspect, it can be noticed that beyond the range of [0, 4), there is a significant drop in contribution as it will be below 20%, and hence, ignored in this work. Now, the third parameter i.e. gamma (γ) value must be positively varied starting from the unity. Hence, when it is simply unity, it leads to simple unity scaled one-to-one mapping. When raised above unity value, it leads to counter related gamma-compressed and gamma-expanded interim channels. Very high gamma value obviously leads to unnatural artifacts and hence, just for avoiding such kind of scenario, range of the gamma parameter is confined to (1,3), which is experimentally found suitable w.r.t. both gamma-corrected and gamma-expanded channels. Next, to frame the influence of the fractional-order for corresponding fractional-order integration, ν is varied in the range of (0,1) so that accordingly adaptive mask can be framed. In this way, now this problem can be easily identified as an optimization problem for search in a four-dimensional search space for positive exploration and exploitation. Finally, it can be simply identified that all four parameters (p , q , γ , and ν) have to be varied in the range [(0,1),[0,4),[1,3),(0,1)], respectively.

D. FITNESS FUNCTION FORMULATION

Mostly the objective functions have been framed by considering only the entropy of the content in the mind. In this formulation, both kinds of discrete entropy contents (intensity based as well as GLCM based entropy values (i.e., DE_O and DE_{GLCM}) for the processing image is considered. To better highlight the effectiveness of the edges as well as the texture content of the images, special considerations are framed by proposing a novel objective function in this context. Magnitude gradient matrix for the output image (GM_O) is evaluated for the processed image by employing the Sobel-Feldman operator along the rows as well as columns of the image. Summing up all of the pixel-wise gradient's magnitude implies the sharpness or the edgy content of the image. This kind of summing-up may lead to a higher order magnitude. For this purpose, a logarithmic operation is

imparted twice over this sum to make it up to the comparable order. Along with it, normalized image contrast measure is also amalgamated by considering its exponential treatment for making it in a comparable order. Cube root for the product of these three quantities made this expression equivalent to the order for the other term in the expression namely, colorfulness metrics' enhancement factor which is just a ratio of colorfulness measure for the output image to the colorfulness measure for the input image. Also, to introduce sufficient dominance of entropy content, the exponential of the framed fraction is implied. As a whole, a newly introduced fitness-function in the proposed enhancement algorithm, is designed as (16), as shown at the bottom of the next page:

E. REINFORCEMENT LEARNING BASED SINE-COSINE OPTIMIZER

Remembering the well-known and remarkable applicability along with the keen understanding for No-free lunch theory, it can never be obvious that, "which evolutionary and/or metaheuristic algorithm will perform best, when associated with the proposed framework". It's again a very challenging to decide that, which optimization mechanism is more effective when associated with the proposed framework and hence, various algorithms (like, PSO, ABC, CSO, SCA, MFO, OBL-SCA, etc.) have been tested for this purpose. Finally, OBL-SCA is found performing very efficiently. Initially, a trigonometrically inspired stochastic population-based optimization, termed as Sine-Cosine Algorithm was suggested. For more optimal behavior, this mechanism is efficiently modified through reinforcement learning. Thus, the opposition based learning inspired Sine-Cosine Algorithm (OBL-SCA) came into existence. Here, OBL-SCA is incorporated in a well-framed manner for obtaining the desired level of quality enhancement in association with the proposed framework. In this context, both the fine as well as coarse texture should be restored and enhanced along with their contrast and entropy enhancement. Fundamentally, SCA is based on sine and cosine based trigonometric functions those are responsible for exploration and exploitation in the search space. A machine learning strategy termed as opposition based learning (OBL) is incorporated along with SCA, so that its intelligence for exploration as well as exploitation mechanism can be utilized, and the search-space can be explored in a more appropriate manner. While evaluating the cost function, the OBL leads to the best solution-set among the original as well as its opposite position data-set, collectively; and hence, this intelligent step-wise learning mechanism finally leads to early convergence. Similar to most of the population-based optimization approaches, initially, a set of randomly evaluated solutions is created in this approach. On repeated consecutive evaluations in a similar fashion, this random set is improved by imparting a set of certain sine-cosine based trigonometric rules, which is the core of this optimization approach. Obviously, the optimal solution hunt is never guaranteed in a single run execution for population-based techniques. Nevertheless, with appropriate population

size of search agents in fixed iteration counts, the probability of attaining global optimum solution can be increased. Irrespective of the behavior of stochastic population-based optimization approaches, the major effort distribution can be categorized into two phases, like exploration versus exploitation. Eventually, in the exploration phase, a high-level abrupt randomness is imparted to find the more and more promising regions of the search space as a set of random solutions. Contrary to this, in the exploitation phase, steady changes are framed for various random solutions, and consequently, the random variations are made noticeably less. The above-mentioned phases can be inherently characterized by following a set of position updating expressions, as follows:

$$X_i^{t+1} = X_i^t + r_1 \cdot \sin(r_2) \times |r_3 \cdot P_i^t - X_i^t|, \quad (17)$$

$$X_i^{t+1} = X_i^t + r_1 \cdot \cos(r_2) \times |r_3 \cdot P_i^t - X_i^t|, \quad (18)$$

where, X_i^t stands for the current position of the solution at t^{th} iteration correspondingly in the i^{th} dimension. The randomness of the exploration phase is maintained by employing four random variables, namely, r_1, r_2, r_3 and r_4 . The usage selection for both of the above equations, is separately governed by r_4 , and hence, it is random and equi-probable. P_i^t stands for i^{th} dimension's destination point for the corresponding t^{th} iteration. Above expressions behave somehow in conjugate fashion, merged using a uniformly distributed random variable r_4 in the range of $[0,1]$.

$$X_i^{t+1} = \begin{cases} X_i^t + r_1 \cdot \sin(r_2) \cdot |r_3 \cdot P_i^t - X_i^t|, & r_4 < 0.5, \\ X_i^t + r_1 \cdot \cos(r_2) \cdot |r_3 \cdot P_i^t - X_i^t|, & r_4 \geq 0.5, \end{cases} \quad (19)$$

The above equation expresses the core updating mechanism ensuring both exploration as well as exploitation collectively. The random parameter r_1 governs the direction of movement. The direction can be inside or outside the region covering the solution to destination intermediate distance. r_2 symbolizes the magnitude of the corresponding to and fro shift. Also, the random variable r_3 is associated for imparting weight randomly for the drift towards the destination and hence, stochastic behavior can be introduced. The corresponding net effect is emphasized if $r_3 > 1$, and deemphasized if $r_3 < 1$. At last, the parameter r_4 is solely responsible for switching exchange in-between the sine and cosine based conjugate equations. The engagement of sine and cosine expressions while framing the behavior of position updating equations, leads to the term ‘‘Sine Cosine Algorithm’’ (SCA) for this approach. It can be easily understood that the entire interim space between two solutions can be defined through the above equations. It should be easily noticed that higher dimensional equivalent system can also be realized, similarly in the corresponding

Algorithm 1 Sine Cosine Algorithm

- 1: **Initialize** a set of search agents (solutions)
 - 2: **Repeat**
 - 3: Compute the values for all search agents by employing the proposed cost function
 - 4: Update solution set by using the best achieved values so far ($P = X^*$)
 - 5: Update the random variable vector r_i ($\forall i \in 1$ to 4)
 - 6: Update the current search agents' position through Eq. (19)
 - 7: **until** ($t <$ maximum iteration count)
 - 8: **Return** the best found solution till the maximum count of iterations and consider it global optimum solution
-

hyper-plane. The conjugate behavior along with the cyclic or periodic nature for sine and cosine functions influences the re-allocation of intermediate or local solutions and hence, following it, better exploration can be easily achieved. Also, for exploring the outside region i.e. between the corresponding destinations can be done by just changing the range of these trigonometric expressions. The relative as well as absolute change of range of sine and cosine expressions leads to relative updating of the position outside/inside the interim region in-between itself and another solution. The above mentioned random position is characterized by r_2 defined in the range $[0, 2]$ employed through updating the equation. Hence, this mechanism ensures the collective effect of exploration as well as exploitation of the entire search space for corresponding dimension iteratively. It is the promising intellectual behavior to maintain the balance of exploration and exploitation, which highlights the outperformance of this optimizer. In this context, assuming a constant positive integral damping factor (a), for the t^{th} iteration, the random variable r_1 is defined in a linearly reducing fashion using the follow expression as:

$$r_1 = a(1 - t/T), \quad (20)$$

where, t signifies the current iteration, and T stands for the total iteration count. The resultant damping or range-reduction during the consecutive course of iterations can be easily understood as an effect of r_1 over the employed updating equation. Also, it can be noticed that the Sine Cosine optimizer explores efficiently, when sine and cosine function ranges in $(1, 2]$ and $[-2, 1)$. Correspondingly, the search space is exploited efficiently when sine and cosine ranges in $[-1, 1]$. The pseudo-code can be understood as a sequence of updating equations employed iteratively by evaluating all four kinds of random parameters. The core algorithm preserves the best achieved solution so far, and this solution is further identified as a destination point in the next step for

$$J \triangleq \left(\frac{CM_O}{CM_I} \right) + \sqrt[3]{e^{SDO} \cdot (DE_O + DE_{GLCM}) \cdot \left(\log \log \left(\sum_{m=1}^M \sum_{n=1}^N (GM_O(m, n)) \right) \right)}, \quad (16)$$

TABLE 1. Information regarding test satellite images.

Image S. No.	Geo-Spatial Location	Spatial Dimensions	Satellite Sensor	Pixel Resolution
1.	Brussels, Belgium [18]	1655 X 1128	QuickBird	0.65m
2.	Himalaya Range [19]	2000 X 1137	Pleiades-1A	0.5m
3.	Millau Viaduct, France [20]	2000 X 2000	Pleiades-1B	0.5m
4.	Guam - Mangilao Golf Resort [18]	2430 X 2488	QuickBird	0.65m
5.	Riyadh, Saudi Arabia [18]	1461 X 1352	QuickBird	0.65m
6.	Los Angeles, California [18]	3500 X 3341	QuickBird	0.65m

Algorithm 2 OBL-SCA Approach

- 1: Define variables
1. Defining upper and lower bounds
2. Initialize a set of search agents X (solutions)
3. Evaluate the opposite ensemble X' as: $\bar{X} = \{\bar{x}_{ij}\} = \{u_i + l_i - x_{ij}\}$, $i = 1, 2, 3, \dots, N$. Where, x_{ij} and \bar{x}_{ij} denote the i^{th} point of the j^{th} solution of x and its corresponding
4. Choose the best N solutions from combined population-set ($X \cup \bar{X}$).
5. Identify this solution set as input for further steps.
6. Repeat
7. Compute the values for all search agents by employing the proposed cost function
8. Update the solution set by using the best achieved values so far ($P = X*$)
9. Update the random variable vector r_i ($\forall i \in 1$ to 4)
10. Update the current search agents' position
11. **until** ($t <$ maximum iteration count)
12. **Return** the best found solution till the maximum count of iterations and consider it global optimum solution

updating other variables w.r.t. it. With the increasing iteration count, range of sine and cosine functions is updated to emphasize the better exploration. The termination of sine-cosine optimizer is executed as the maximum iteration count is achieved. The smooth transition switching in-between exploration and exploitation phases is the best intellectual feature of this approach due to adaptive range selection for sine and cosine functions. Also, the best global optimal approximation achieved till the current iteration is considered as the destination for drifting; and hence, the chances of getting lost of the search agents during optimization, is efficiently suppressed. SCA also leads to the abrupt changes initially and gradual changes in later stages. Step-wise, proposed approach and hereby employed algorithms are as follows:

IV. EXPERIMENTATION AND DISCUSSION

Experimental validation along with comparative performance evaluation is done by reimplementing of various state-of-the-art methods such as GHE, MMSICHE [5], ADAPHE [6], AVHEQ [7], AGCWD [8], HEOPC [10], HEMIC [11], IEAUMF [12], PGCHE [4] and PGCDFM [3]. Qualitative

Algorithm 3 Proposed PGCRLFOIM Framework

- 1: **INPUT (1)** : Input image (I_{in})
- 2: **INPUT (2)**: $X = \{p, q, \gamma, \nu\}$ as the input parametric vector consisting of weighting parameters (p, q) gamma value (γ) order of fractional-order integral mask (ν) scaling parameter for negative augmentation (k).
- 3: **OUTPUT**: GLCM based cost function (J) and Quality Improved Output Image (I_{en})
- 4: Evaluation of the tile-wise equalized (I_{teq}) input channel.
- 5: Computation of gamma compressed interim intensity channel
- 6: Computation of gamma expanded interim intensity channel
- 7: Evaluation of ν -ordered FOI mask (H) as suggested in eq. (11).
- 8: Fractionally ordered filtering is imparted through these masks to extract non-textured information as, $I_\nu = I_{in} \otimes H$
- 9: Computation of partially texture enhanced interim image by employing: $I_{foim} = I_{in} + 0.5.\lambda (I_{in} - I_\nu)$.
- 10: Computation of the enhanced image using Eq. (15),
- 11: Evaluation of the cost function, as shown in Eq. (16),
- 12: **RETURN**: Magnitude of the cost function J .

and quantitative outcomes for the enhanced images for various methods are also presented in this paper. For quantitative assessment and corresponding qualitative analysis, visually improved and resultant images obtained by employing a variety of state of the art methodologies can be collectively evaluated in Fig. 1-6. In the similar manner, corresponding tabular evaluation is also presented in Table 2, by assembling most significant eight fundamental performance measures for all of the state-of the art methodologies published recently. Various satellite images acquired from standard databases (as listed in Table 1) are tested. Especially, to make the images dark and low contrasted, a fixed intensity value is deducted from these test images and hence, substantial information content is initially loosed. Later, through proper experimentation the information regain is attained. In this manner, a kind of assurance/confidence is achieved that, the proposed framework will be well suited for the required enhancement of the remotely sensed, poorly illuminated satellite images.

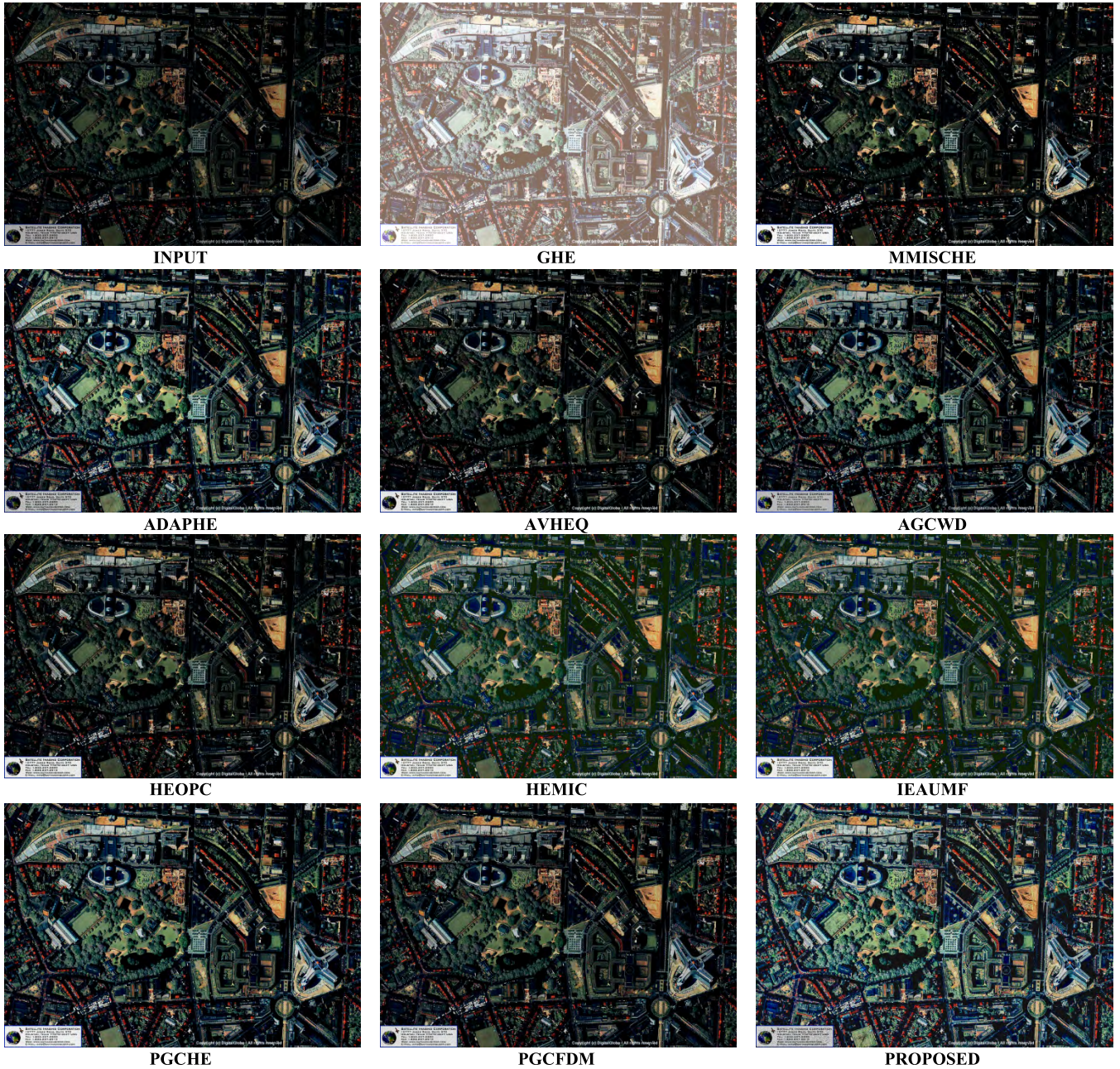


FIGURE 1. Visual presentation/ qualitative evaluation with comparison among input images; GHE; MMSICHE; ADAPHE; AVHEQ; AGCWD; HEOPC; HEMIC; IEAUMF; PGCHE; PGCDFM and the proposed approach for Image 1 (i.e., Brussels, Belgium).

A. BASIS FOR COMPARATIVE EVALUATION

Eight highly reliable and fundamentally identified measures, namely brightness (B), contrast (V), discrete Shannon entropy (H), sharpness (S), colorfulness (C), GLCM-correlation (GC), GLCM-energy (GE), and GLCM-homogeneity (GH) are employed here for explicit measurement for the excellence over various methodologies. Quality improvement can be justified by relative increment in B, V, H, S, and C along with relative decrement in GC, GE, and GH.

B. ONE-DIMENSIONAL HISTOGRAM BASED PERFORMANCE INDICES

Brightness (or mean, B) value for M by N sized image matrix $I(m,n)$ is evaluated as an averaged summation, as:

$$Brightness(B) = \frac{1}{M \times N} \sum_{m=1}^M \sum_{n=1}^N I(m,n), \quad (21)$$

Average intensity spread or the variance (V) accounts for the image contrast, responsible for a naturally pleasant look, can

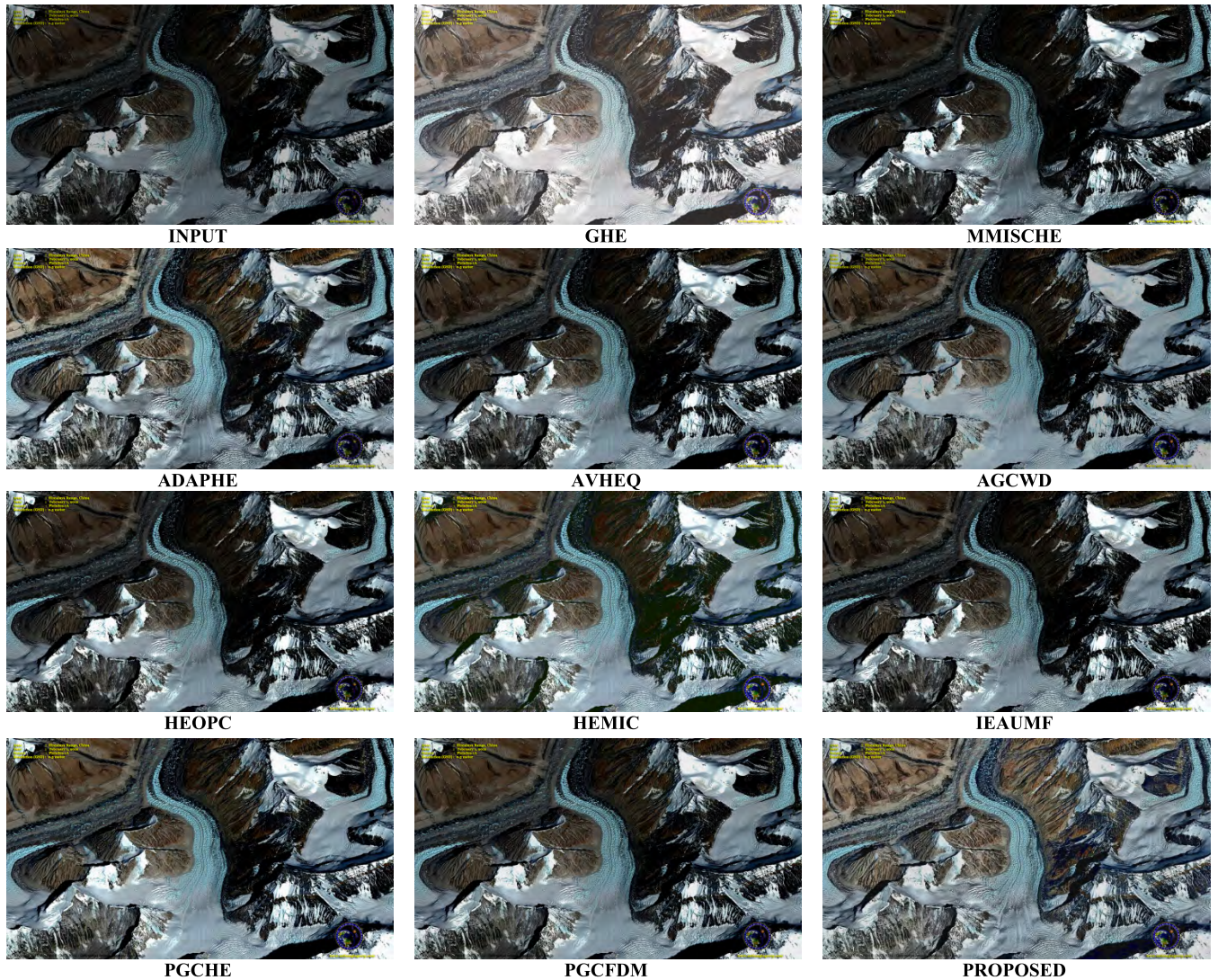


FIGURE 2. Visual presentation/ qualitative evaluation with comparison among input images; GHE; MMSICHE; ADAPHE; AVHEQ; AGCWD; HEOPC; HEMIC; IEAUMF; PGCHE; PGCDFM and the proposed approach for Image 2 (i.e., Himalaya Range).

be evaluated as:

$$Contrast (V) = \frac{1}{M \times N} \sum_{m,n} I(m, n)^2 - \left(\frac{1}{M \times N} \sum_{m,n} I(m, n) \right)^2, \quad (22)$$

Information content is quantified by Shannon entropy of the image and hence, bounded probability calculation using normalized image histogram, as:

$$Entropy (H) = - \sum_{i=0}^{I_{max}} p_i \log_2(p_i), \quad (23)$$

where, $p_i = n_i / (M * N)$ accounts for the intensity level-wise possibility of occurrence and, maximum intensity level is symbolized by I_{max} . Accountability for the presence of the edge content can be easily identified through the sharpness content of the image that can be also identified as the gradient

of the image, evaluated as:

$$Sharpness (S) = \frac{1}{M \times N} \sum_{m,n} \left(\sqrt{\Delta m^2 + \Delta n^2} \right), \quad (24)$$

where, $\Delta m = I_{enh}(m, n) - I_{enh}(m + 1, n)$ and $\Delta n = I_{enh}(m, n) - I_{enh}(m, n + 1)$ symbolizes for the accountability of the local values of gradient content of the image. For color images, color channel's coordination is also significant. Thus, utilizing relative colors' variance and relative colors' mean value, coordination of different color channels can be identified as 'colorfulness' of the image, which can be evaluated as,

$$Colorfulness (C) = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \quad (25)$$

$$\Delta rg = R - G; \Delta yb = 0.5 (R + G) - B; \quad (26)$$

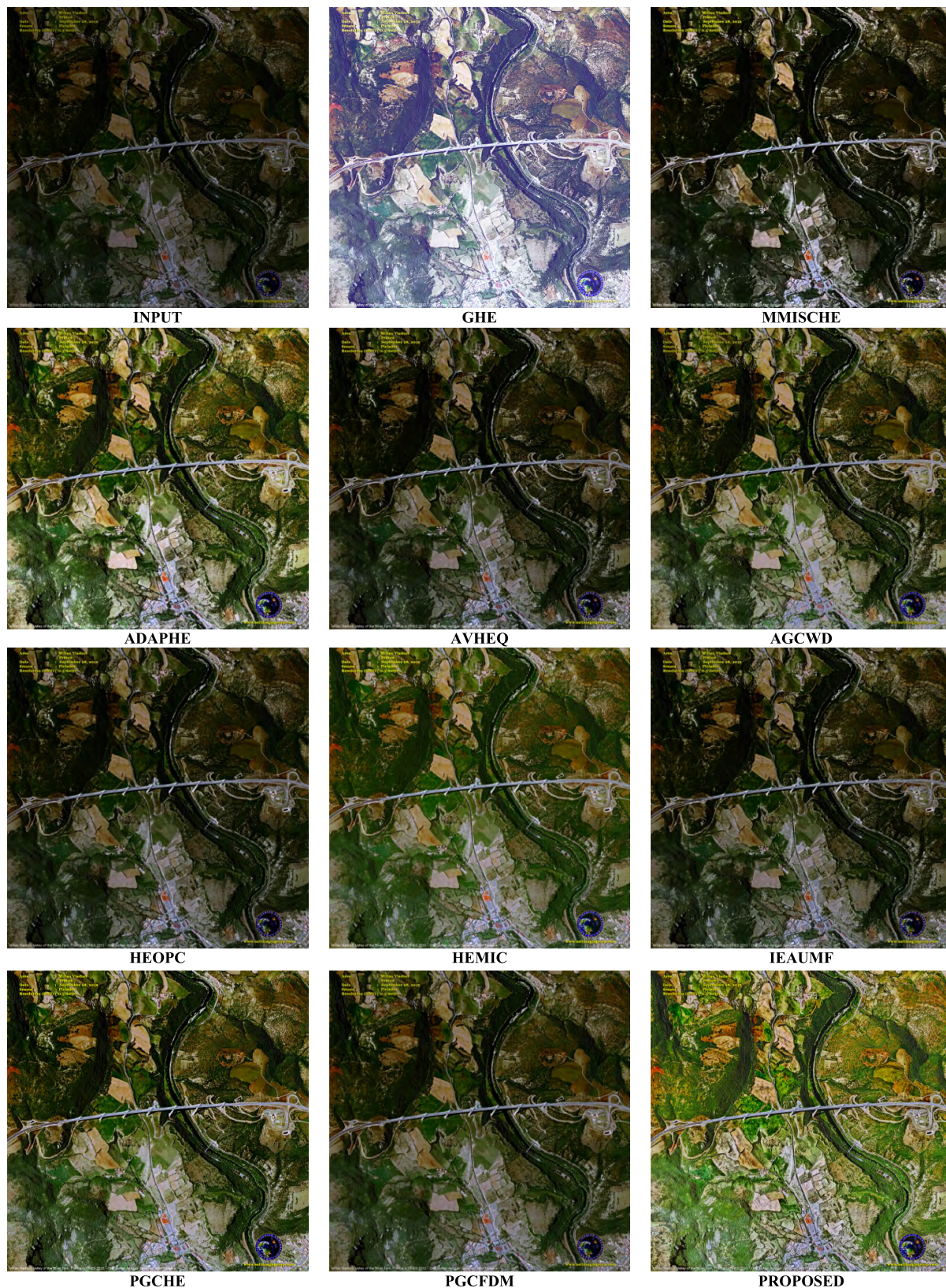


FIGURE 3. Visual presentation/ qualitative evaluation with comparison among input images; GHE; MMSICHE; ADAPHE; AVHEQ; AGCWD; HEOPC; HEMIC; IEAUMF; PGCHE; PGCFDMF and the proposed approach for Image 3 (i.e., Millau Viaduct, France).

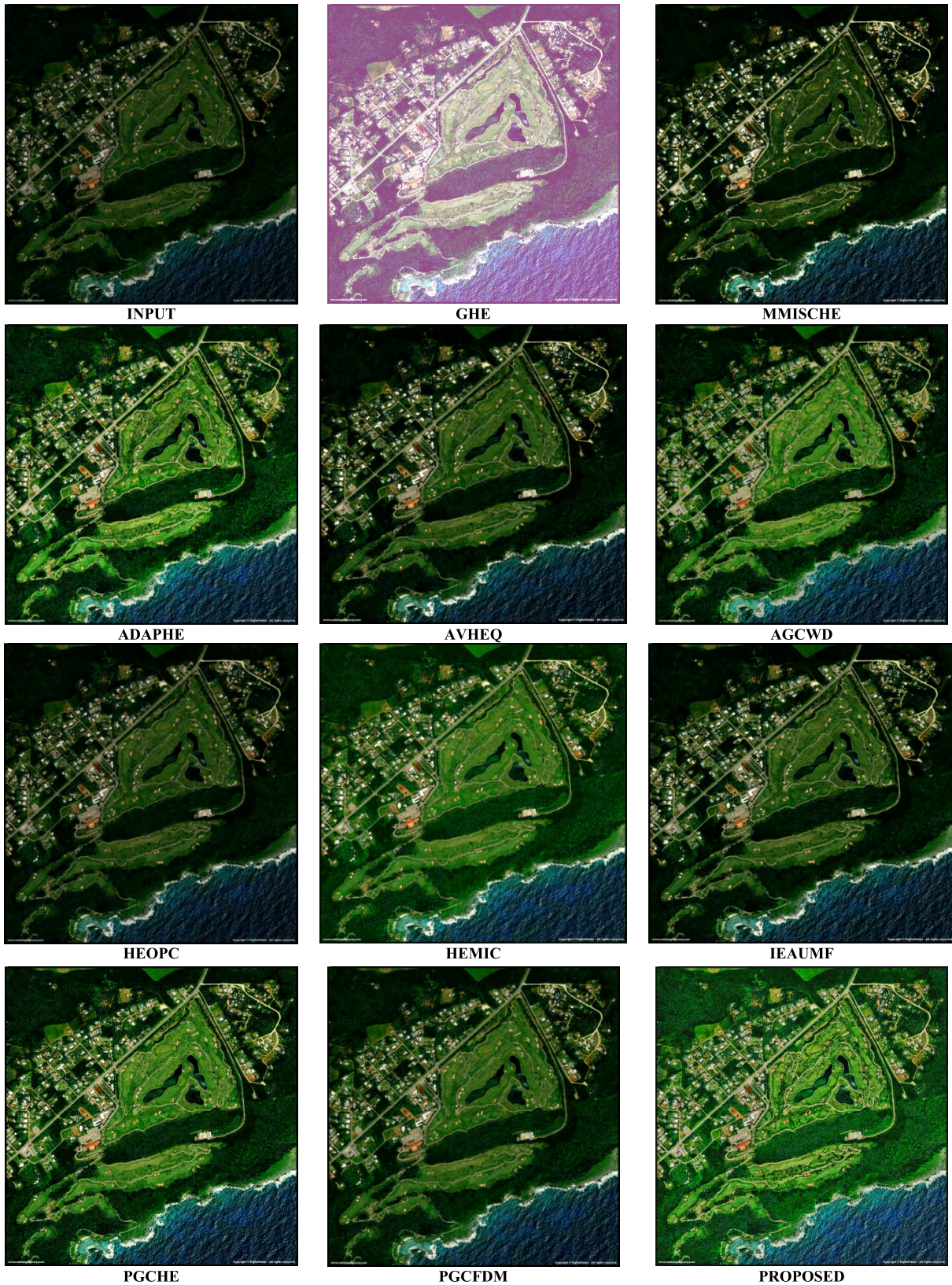


FIGURE 4. Visual presentation/ qualitative evaluation with comparison among input images; GHE; MMSICHE; ADAPHE; AVHEQ; AGCWD; HEOPC; HEMIC; IEAUMF; PGCHE; PGCDFM and the proposed approach for Image 4 (i.e., Guam - Mangilao Golf Resort).

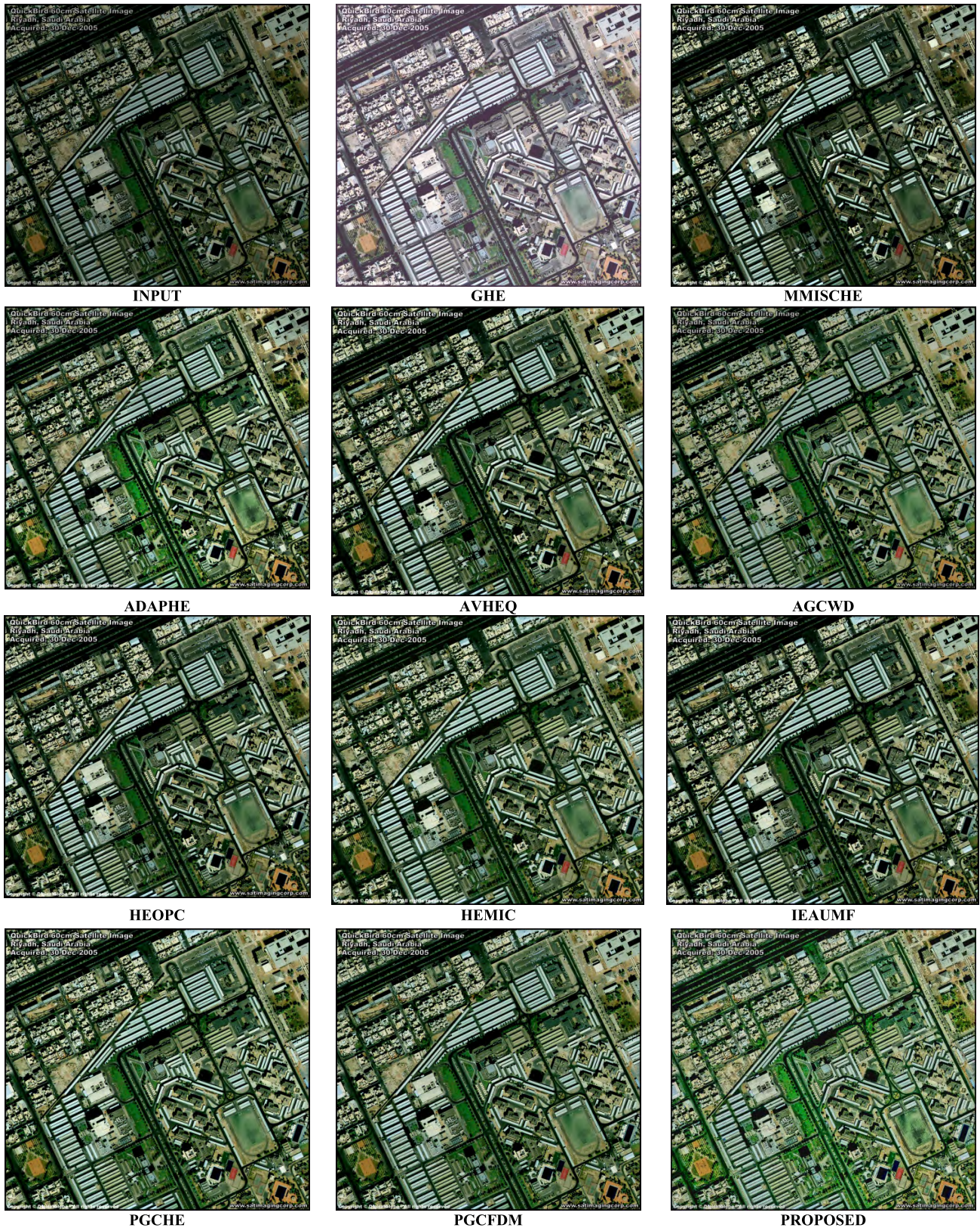


FIGURE 5. Visual presentation/ qualitative evaluation with comparison among input images; GHE; MMSICHE; ADAPHE; AVHEQ; AGCWD; HEOPC; HEMIC; IEAUMF; PGCHE; PGCDFM and the proposed approach for the Image 5 (i.e., Riyadh, Saudi Arabia).

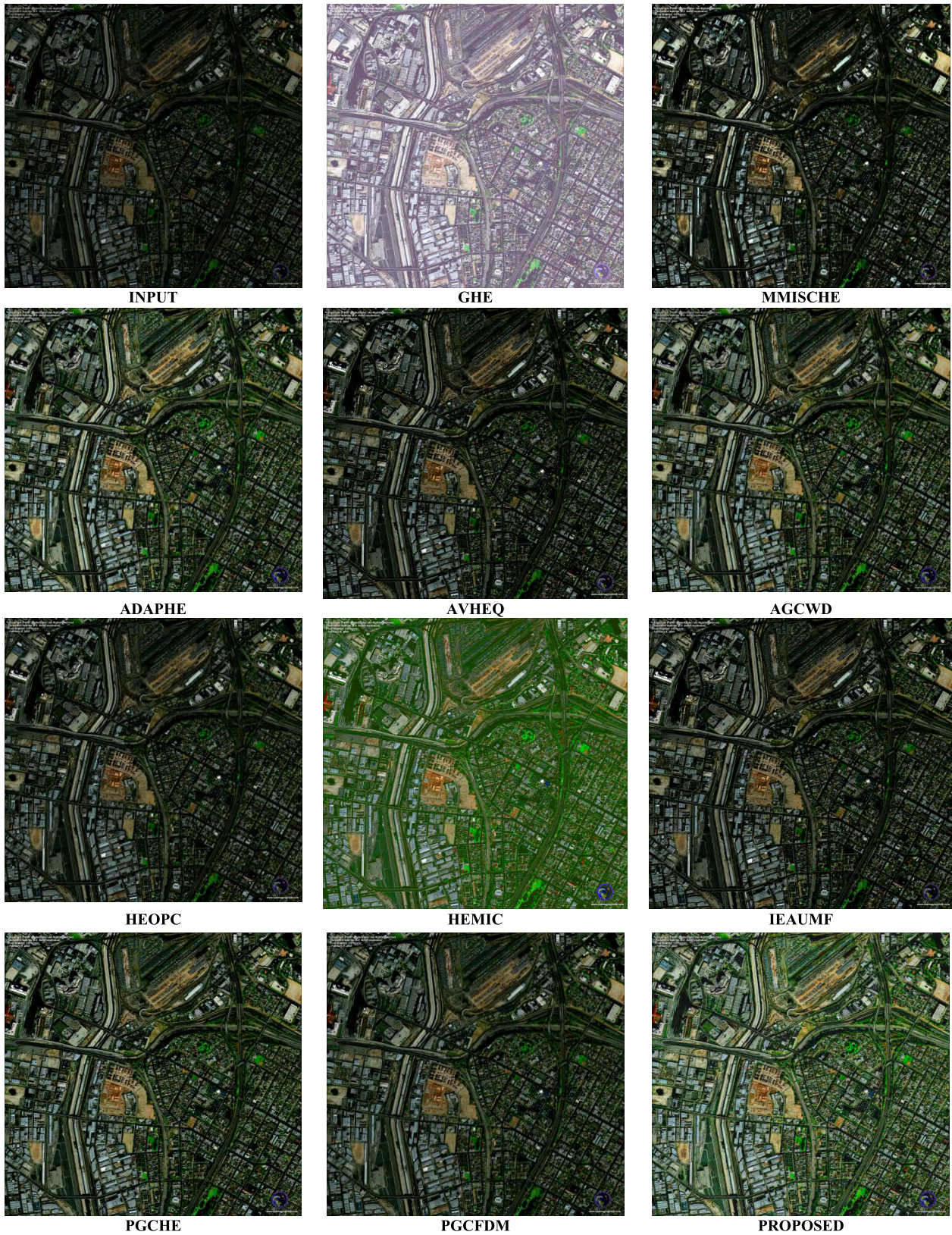


FIGURE 6. Visual presentation/ qualitative evaluation with comparison among input images; GHE; MMSICHE; ADAPHE; AVHEQ; AGCWD; HEOPC; HEMIC; IEUMF; PGCHE; PGCDFM and the proposed approach for the image 6 (i.e., los angeles, california).

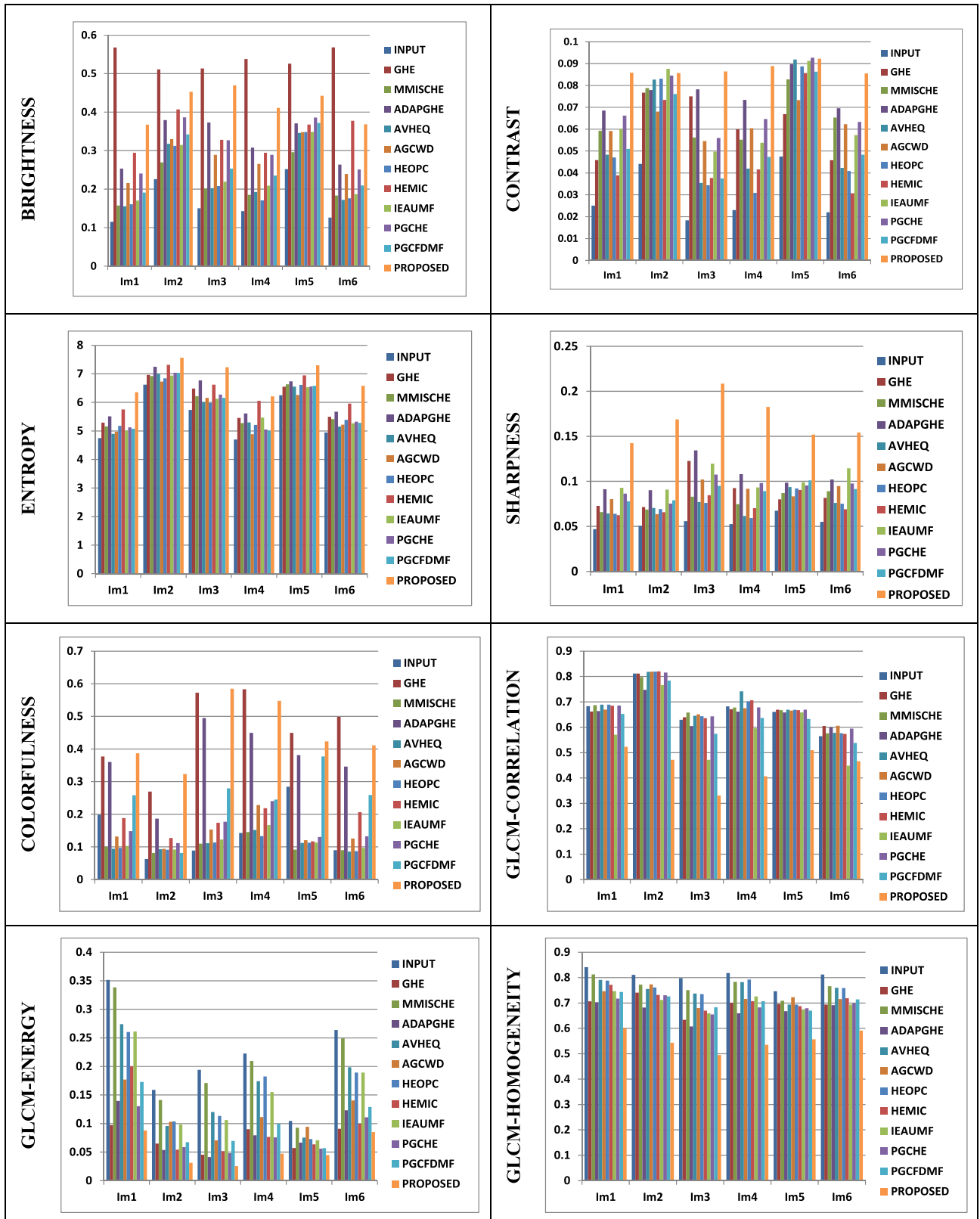


FIGURE 7. Comparative evaluation for performance indices for different test images (Im1 to Im6).

TABLE 2. Numerical results for comparative evaluation among input, GHE, MMSICHE, ADAPHE, AVHEQ, AGCWD, HEOPC, HEMIC, IEAUMF and the proposed approach by evaluating measures like brightness (B), CONTRAST (V), ENTROPY (H), SHARPNESS (S), COLORFULNESS (C), GLCM-CORRELATION (GC), GLCM-ENERGY (GE), GLCM-HOMOGENEITY (GH).

IMAGE No.	INDICES	INPUT	GHE	MMSICHE	ADAPHE	AVHEQ	AGCWD	HEOPC	HEMIC	IEAUMF	PGCHE	PGCFDM	Proposed
1.	B	0.1153	0.5681	0.1575	0.2534	0.1555	0.2164	0.1610	0.2947	0.1701	0.2407	0.1912	0.3675
	V	0.0251	0.0459	0.0593	0.0685	0.0483	0.0592	0.0471	0.0389	0.0601	0.0662	0.0510	0.0858
	H	0.0469	0.0728	0.0659	0.0913	0.0643	0.0802	0.0640	0.0625	0.0929	0.0864	0.0778	0.1424
	S	4.7471	5.2937	5.1580	5.5126	4.8985	4.9708	5.1778	5.7478	5.0058	5.1239	5.0781	6.3526
	C	0.1986	0.3770	0.1003	0.3602	0.0946	0.1318	0.0967	0.1882	0.1028	0.1481	0.2586	0.3870
	GC	0.8407	0.7066	0.8126	0.7029	0.7912	0.7461	0.7881	0.7714	0.7466	0.7174	0.7436	0.6000
	GE	0.3516	0.0972	0.3386	0.1397	0.2739	0.1770	0.2604	0.2001	0.2612	0.1302	0.1728	0.0878
GH	0.6831	0.6618	0.6864	0.6637	0.6891	0.6700	0.6891	0.6848	0.5706	0.6851	0.6523	0.5224	
2.	B	0.2261	0.5111	0.2693	0.3797	0.3179	0.3302	0.3126	0.4072	0.3146	0.3866	0.3421	0.4530
	V	0.0441	0.0767	0.0788	0.0779	0.0827	0.0681	0.0831	0.0734	0.0876	0.0845	0.0761	0.0857
	H	0.0506	0.0715	0.0686	0.0901	0.0704	0.0635	0.0692	0.0657	0.0908	0.0753	0.0790	0.1688
	S	6.6207	6.9598	6.9166	7.2433	7.0060	6.7278	6.8395	7.3162	6.9370	7.0363	7.0302	7.5627
	C	0.0627	0.2697	0.0815	0.1863	0.0932	0.0942	0.0913	0.1270	0.0922	0.1117	0.0812	0.3235
	GC	0.8108	0.7408	0.7724	0.6816	0.7552	0.7732	0.7608	0.7315	0.7119	0.7305	0.7255	0.5428
	GE	0.1589	0.0650	0.1413	0.0538	0.0956	0.1030	0.1037	0.0543	0.0982	0.0589	0.0672	0.0314
GH	0.8114	0.8113	0.7975	0.7473	0.8182	0.8195	0.8193	0.8201	0.7661	0.8156	0.7839	0.4713	
3.	B	0.1502	0.5135	0.2022	0.3731	0.2023	0.2892	0.2079	0.3281	0.2195	0.3271	0.2533	0.4699
	V	0.0184	0.0750	0.0562	0.0782	0.0354	0.0546	0.0344	0.0377	0.0498	0.0560	0.0375	0.0864
	H	0.0558	0.1227	0.0829	0.1345	0.0771	0.1021	0.0761	0.0846	0.1195	0.1075	0.0950	0.2085
	S	5.7321	6.4839	6.2171	6.7679	6.0040	6.1546	5.9817	6.6186	6.1216	6.2741	6.1538	7.2297
	C	0.0888	0.5724	0.1097	0.4949	0.1114	0.1534	0.1134	0.1739	0.1224	0.1773	0.2796	0.5848
	GC	0.7978	0.6336	0.7510	0.6072	0.7367	0.6802	0.7345	0.6700	0.6593	0.6547	0.6828	0.4952
	GE	0.1942	0.0454	0.1709	0.0412	0.1201	0.0704	0.1133	0.0515	0.1060	0.0480	0.0694	0.0254
GH	0.6292	0.6392	0.6578	0.6045	0.6451	0.6510	0.6434	0.6352	0.4714	0.6426	0.5743	0.3309	
4.	B	0.1430	0.5382	0.1852	0.3079	0.1925	0.2657	0.1708	0.2944	0.2093	0.2894	0.2354	0.4112
	V	0.0230	0.0600	0.0552	0.0734	0.0420	0.0604	0.0309	0.0416	0.0538	0.0646	0.0473	0.0889
	H	0.0525	0.0925	0.0745	0.1080	0.0616	0.0917	0.0592	0.0703	0.0932	0.0981	0.0892	0.1827
	S	4.6992	5.4557	5.2727	5.6113	5.2968	4.8846	5.2115	6.0543	5.4682	5.0509	5.0158	6.2101
	C	0.1426	0.5829	0.1452	0.4494	0.1519	0.2287	0.1332	0.2189	0.1670	0.2401	0.2448	0.5474
	GC	0.8180	0.7005	0.7830	0.6598	0.7822	0.7161	0.7924	0.7070	0.7253	0.6822	0.7069	0.5351
	GE	0.2227	0.0901	0.2097	0.0793	0.1742	0.1114	0.1824	0.0767	0.1550	0.0758	0.0995	0.0472
GH	0.6823	0.6708	0.6777	0.6614	0.7414	0.6749	0.7024	0.7064	0.5959	0.6781	0.6364	0.4065	
5.	B	0.2520	0.5259	0.2959	0.3711	0.3461	0.3480	0.3486	0.3679	0.3490	0.3859	0.3721	0.4426
	V	0.0475	0.0669	0.0828	0.0897	0.0918	0.0733	0.0887	0.0857	0.0912	0.0927	0.0863	0.0922
	H	0.0673	0.0800	0.0871	0.0984	0.0935	0.0832	0.0921	0.0905	0.0990	0.0953	0.1010	0.1520
	S	6.2499	6.5434	6.6297	6.7393	6.5505	6.2553	6.6110	6.9469	6.5278	6.5582	6.5771	7.3002
	C	0.2839	0.4493	0.0915	0.3813	0.1122	0.1206	0.1126	0.1167	0.1134	0.1305	0.3772	0.4234
	GC	0.7462	0.6960	0.7088	0.6675	0.6935	0.7225	0.6938	0.6874	0.6748	0.6791	0.6700	0.5569
	GE	0.1044	0.0572	0.0926	0.0664	0.0754	0.0942	0.0727	0.0635	0.0709	0.0561	0.0570	0.0447
GH	0.6606	0.6690	0.6672	0.6584	0.6691	0.6654	0.6683	0.6676	0.6584	0.6690	0.6322	0.5091	
6.	B	0.1265	0.5684	0.1833	0.2637	0.1719	0.2393	0.1761	0.3776	0.1871	0.2512	0.2095	0.3682
	V	0.0220	0.0458	0.0653	0.0696	0.0423	0.0623	0.0409	0.0307	0.0573	0.0633	0.0483	0.0855
	H	0.0551	0.0816	0.0890	0.1020	0.0761	0.0949	0.0751	0.0690	0.1146	0.0976	0.0914	0.1542
	S	4.9446	5.4939	5.4127	5.6706	5.1541	5.2182	5.3898	5.9595	5.2632	5.3252	5.2787	6.5802
	C	0.0900	0.4987	0.0896	0.3462	0.0855	0.1259	0.0872	0.2064	0.0966	0.1322	0.2589	0.4109
	GC	0.8119	0.6938	0.7655	0.6913	0.7598	0.7155	0.7583	0.7182	0.6939	0.6988	0.7140	0.5910
	GE	0.2637	0.0909	0.2491	0.1232	0.1984	0.1405	0.1894	0.0995	0.1893	0.1108	0.1291	0.0849
GH	0.5649	0.6051	0.5762	0.5993	0.5779	0.6061	0.5769	0.5737	0.4489	0.5948	0.5381	0.4663	

where, $\Delta_{rg}, \Delta_{yb}, \mu_{rg}, \mu_{yb}, \sigma_{rg}, \sigma_{yb}$, symbolizes corresponding differential values, differential mean and corresponding differential standard deviation values, respectively.

C. GLCM BASED PERFORMANCE INDICES

Spatial co-occurrence of the image pixels are usually avoided while evaluating the intensity based indices, and hence, to resolve it, Gray-Level Co-occurrence Matrix based performance indices also plays a significant role for texture

and other spatially influenced properties. Overall statistical and spatial behavior w.r.t. reference pixel can be derived by calculating the pixel-wise average for all four directional matrices:

$$GLCM = 0.25 \left(GLCM_0 + GLCM_{\pi/4} + GLCM_{\pi/2} + GLCM_{3\pi/4} \right); \tag{27}$$

TABLE 3. Statistical improvement achieved for various quality measures (averaged over 30 images).

	Brightness	Contrast	Entropy	Sharpness	Colorfulness	Correlation	Energy	Homogeneity
Input	0.1689	0.0300	0.0547	5.4989	0.1444	0.8042	0.2159	0.6719
Output	0.3890	0.0757	0.1353	6.7521	0.4281	0.5982	0.0594	0.5257
Increment	130.4%	152.4%	147.3%	22.8%	196.4%	-25.6%	-72.5%	-21.8%

In this paper, three well known GLCM based indices, i.e. GLCM-Correlation, GLCM-Energy and GLCM-Homogeneity are evaluated. Any element of the GLCM matrix $\Psi(m, n)$, is usually evaluated by considering the n^{th} neighboring pixel w.r.t. m^{th} pixel, and later on, by calculating the μ_m, μ_n, σ_m , and σ_n as the corresponding mean values and standard deviation values respectively. GLCM-correlation (GC) stands for the interdependency for the corresponding neighborhood of the pixels w.r.t. reference pixels, expressed as:

$$GC = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \frac{(m - \mu_m)(n - \mu_n) \Psi(m, n)}{\sigma_m \cdot \sigma_n}, \quad (28)$$

GLCM-Energy (GE) can be characterized by normalized count of repeated pairs. Intuitively, these are responsible for uniformity of texture, and hence, expressed as:

$$GLCM - Energy (GE) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \Psi(m, n)^2, \quad (29)$$

GLCM-homogeneity (GH) can be characterized by the closeness of neighboring pixels with reference pixels. Intuitively, these are also responsible for uniformity of texture, and hence, expressed as:

$$GH = - \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \Psi(m, n) \log_2 \Psi(m, n), \quad (30)$$

Ideally, all these values should be as low as possible for better texture visualization of the content. Hence, relatively lower values for these parameters are usually appreciated.

D. COMPARATIVE ASSESSMENT OF THE PROPOSED PGC-RLFOIM

Overall assessment of the proposed approach is done by rigorous experimentation by evaluating all above-mentioned performance indices for some of the earlier, state-of-the-art proposals by various researchers. The tabular results along with the resulted images and their corresponding bar statistics are produced here for portraying the general excellence of the proposed approach. Individual results can be easily analyzed by considering the core objective of simultaneous contrast and entropy improvement along with sharpness enhancement. In addition to it, some amount of brightness improvement is also desired in case of the dark images, and in this way more scope can be explored for further contrast improvement by exploring more span of intensity levels. For accounting GLCM-based assessment, as discussed

above, the least values of indices like GLCM correlation (GC), GLCM energy (GE) and GLCM-homogeneity (GH) are appreciated for a quality enhanced image. The theoretical excellence due to the presence of OBL is quite implacable, and hence can be accepted easily for the outperformance of the proposed method due to its self-ignited, exhaustive learning mechanism. It can be proved experimentally by integrating various optimization methods along with hereby proposed PGCRLFOIM framework by utilizing the comparable resources like the same iteration-count along with same population size; so that the unbiased comparison can be illustrated for the various algorithms' co-existence compatibility. OBL-SCA is utilized for final modeling, which itself has been derived by imparting intelligence based on opposition based exhaustive machine learning approach; and hence, termed as OBL-SCA.

Clustered bar-graphs as presented in Fig. 7 also draw a very clear sketch for the outperformance of the proposed approach. The performance bar graphs for six different test-images are plotted. The different colors of the bar columns represent different test images. Better quality improvement can be easily advocated by increased value for brightness (B), contrast (V), entropy (H), sharpness (S) and colorfulness (C) as shown in Fig 7. (a-e). Contrary to this, decreased values of GLCM-based indices namely correlation (R), energy (E) and homogeneity (M) collectively advocate the better texture based quality improvement, as shown in Fig 7(f-h). As listed in Table 3, an averaged analysis is also presented over various test images. Accordingly, 152.4% increment (2.52 times) is achieved over the input contrast along with the simultaneous 147.3% (2.47 times) increment in the discrete entropy level and 22.8% (1.228 times) increment in the sharpness content. Also, for dark color images, higher values of brightness and colorfulness are also desired, those are reported with 130.4% (2.3 times) and 196.4% (2.96 times) increased w.r.t. the input indices, respectively. In addition, the textural improvement is advocated in terms of desired comparative reduction of GLCM based metrics, namely correlation, energy and homogeneity are suppressed by 25.6% (0.744 times), 72.5% (0.275 times), and 21.8% (0.782 times), respectively. Hence, the desired objective is achieved efficiently.

V. CONCLUSION

As a concluding note, it can be said that the prime objective is to extract more and more information by imparting adaptive/content-wise boosting/improvement for visual features of the poorly acquired, remotely sensed textured satel-

lite images. Along with proper illumination improvement for dark satellite images, content dependent texture enhancement is also required for efficient post-processing usage. With this above-mentioned objective, an optimal fusion based framework is proposed for a constructive association of gamma-compressed and gamma-expanded intensity channels, with a third channel which is a novel inclusion in this paper. This third channel is a texture enhanced version of the image obtained by fractional-order RL integration based adaptive 2-D filtering. Effectively proposed, 2-D RL fractional-order integral mask is a highly efficient version of blurring/smoothing filter and when, thus processed filtered output is negatively augmented with input intensity channel, with proper intensity dependent scaling factor, leads to texture enhanced version of the input channel. The proposed PGC-RLFOIM approach can also be identified as a weighted association of all three interim channels, namely gamma compressed, gamma expanded, and the fractional-order integration based texture enhanced version. Optimal association of these interim images is planned in an intelligent manner by adaptive exploration along with efficient garnering of missing intensity levels throughout the span of permissible intensity levels. Along with this newly proposed fusion framework, a novel fitness function has been also suggested in this paper which seems very robust for all kinds of textural and non-textural image details. According to the demand, OBL-SCA is associated with the proposed framework after rigorous experimentation, so that a four-dimensional (4-D) search space can be fruitfully explored and exploited for achieving the suitable values of $(\alpha, \beta, \gamma, \nu)$ so that overall enhancement can be imparted. The approach has been found very efficient for remotely sensed satellite images and on some general images as well.

REFERENCES

- [1] J. Wang, Y. Ye, X. Pan, X. Gao, and C. Zhuang, "Fractional zero-phase filtering based on the Riemann–Liouville integral," *Signal Process.*, vol. 98, pp. 150–157, May 2014. doi: 10.1016/j.sigpro.2013.11.024.
- [2] Y.-F. Pu, J.-L. Zhou, and X. Yuan, "Fractional differential mask: A fractional differential-based approach for multiscale texture enhancement," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 491–511, Feb. 2010. doi: 10.1109/TIP.2009.2035980.
- [3] H. Singh, A. Kumar, L. K. Balyan, and G. K. Singh, "A novel optimally weighted framework of piecewise gamma corrected fractional order masking for satellite image enhancement," *Comput. Elect. Eng.*, to be published. doi: 10.1016/j.compeleceng.2017.11.014.
- [4] H. Singh, A. Kumar, L. K. Balyan, and G. K. Singh, "Swarm intelligence optimized piecewise gamma corrected histogram equalization for dark image enhancement," *Comput. Elect. Eng.*, vol. 70, pp. 462–475, Aug. 2018. doi: 10.1016/j.compeleceng.2017.06.029.
- [5] K. Singh and R. Kapoor, "Image enhancement via median-mean based sub-image-clipped histogram equalization," *Opt.-Int. J. Light Electron Opt.*, vol. 125, no. 17, pp. 4646–4651, Sep. 2014. doi: 10.1016/j.ijleo.2014.04.093.
- [6] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. Heckbert, Ed. New York, NY, USA: Academic, 1994. doi: 10.1016/B978-0-12-336156-1.50061-6.
- [7] S. C.-F. Lin et al., "Image enhancement using the averaging histogram equalization (AVHEQ) approach for contrast improvement and brightness preservation," *Comput. Elect. Eng.*, vol. 46, pp. 356–370, Aug. 2015. doi: 10.1016/j.compeleceng.2015.06.001.
- [8] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, Mar. 2013. doi: 10.1109/TIP.2012.2226047.
- [9] H. Singh, A. Kumar, L. K. Balyan, and G. K. Singh, "Slantlet filter-bank-based satellite image enhancement using gamma-corrected knee transformation," *Int. J. Electron.*, vol. 105, no. 10, pp. 1695–1715, 2018. doi: 10.1080/00207217.2018.1477199.
- [10] C. Y. Wong et al., "Histogram equalization and optimal profile compression based approach for colour image enhancement," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 802–813, Jul. 2016. doi: 10.1016/j.jvcir.2016.04.019.
- [11] C. Y. Wong et al., "Image contrast enhancement using histogram equalization with maximum intensity coverage," *J. Mod. Opt.*, vol. 63, no. 16, pp. 1618–1629, 2016. doi: 10.1080/09500340.2016.1163428.
- [12] S. C.-F. Lin et al., "Intensity and edge based adaptive unsharp masking filter for color image enhancement," *Optik*, vol. 127, no. 1, pp. 407–414, 2016. doi: 10.1016/j.ijleo.2015.08.046.
- [13] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 4, Nov. 1995, pp. 1942–1948. doi: 10.1109/ICNN.1995.488968.
- [14] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Appl. Soft Comput.*, vol. 8, pp. 687–697, Jan. 2008. doi: 10.1016/j.asoc.2007.05.007.
- [15] X. S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proc. World Congr. Nature Biol. Inspired Comput. (NABIC)*, Dec. 2009, pp. 210–214. doi: 10.1109/NABIC.2009.5393690.
- [16] S. Mirjalili, "SCA: A sine cosine algorithm for solving optimization problems," *Knowl.-Based Syst.*, vol. 96, pp. 120–133, Mar. 2016. doi: 10.1016/j.knsys.2015.12.022.
- [17] M. A. Elaziz, D. Oliva, and S. Xiong, "An improved opposition-based sine cosine algorithm for global optimization," *Expert Syst. Appl.*, vol. 90, pp. 484–500, Dec. 2017. doi: 10.1016/j.eswa.2017.07.043.
- [18] [Online]. Available: <https://www.satimagingcorp.com/gallery/quickbird/>
- [19] [Online]. Available: <https://www.satimagingcorp.com/gallery/pleiades-1/>
- [20] [Online]. Available: <https://www.satimagingcorp.com/gallery/pleiades-2/>



HIMANSHU SINGH received the B.E. degree (Hons.) in electronics and communication engineering from the Madhav Institute of Technology and Science, Gwalior, India, in 2010. He is currently pursuing the Ph.D. degree in electronics and communication engineering with the Indian Institute of Information Technology, Design and Manufacturing Jabalpur, India. His research interests include signal and image processing applications, image enhancement, and soft-computing techniques.



ANIL KUMAR received the B.E. degree in electronic and telecommunication engineering from the Army Institute of Technology, Pune University, India, in 2002, and the M.Tech. and Ph.D. degrees in electronic and telecommunication engineering from IIT Roorkee, India, in 2006 and 2010, respectively. He is currently an Assistant Professor with the Electronic and Communication Engineering Department, Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, India. He is also a Visiting Researcher with the Gwangju Institute of Science and Technology, South Korea. His research interests include design of digital filters and filter bank, biomedical signal processing, image processing, and speech processing.



L. K. BALYAN received the M.Sc. degree in applied mathematics from IIT Roorkee and the Ph.D. degree in applied mathematics from IIT Kanpur, in 2009. He is currently an Assistant Professor of mathematics with the Discipline of Natural Science, IIITDM, Jabalpur. His research interests include numerical solution of partial differential equations, and spectral methods and their applications.



HEUNG-NO LEE received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively. From 1999 to 2002, he was a Research Staff Member with HRL Laboratories, LLC, Malibu, CA, USA. From 2002 to 2008, he was an Assistant Professor with the University of Pittsburgh, PA, USA. In 2009, he joined the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea, where he currently holds a position. His areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, the future Internet, and brain-computer interface.

• • •



A new design method for FIR notch filter using Fractional Derivative and swarm intelligence

A KUMAR^{1,*}, K N MUSTIKOVILA¹, G K SINGH², S LEE³ and H-N LEE³

¹PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur 482005, India

²Indian Institute of Technology Roorkee, Roorkee 247667, India

³School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Korea

e-mail: anilkdee@gmail.com; mustikovila@iiitdmj.ac.in; gksngfee@gmail.com; seungchan@gist.ac.kr; heungno@gist.ac.kr

MS received 6 March 2018; revised 29 June 2018; accepted 22 October 2018

Abstract. In this paper, a new design method for the finite impulse response (FIR) notch filters using fractional derivative (FD) and swarm intelligence technique is presented. The design problem is constructed as a minimization of the magnitude response error w.r.t. filter coefficients. To acquire high accuracy of notch filter, fractional derivative (FD) is evaluated, and the required solution is computed using the Lagrange multiplier method. The fidelity parameters like passband error, notch bandwidth, and maximum passband ripple vary non-linearly with respect to FD values. Moreover, the tuning of appropriate FD value is computationally expensive. Therefore, modern heuristic methods, known as the constraint factor particle swarm optimization (CFI-PSO), which is inspired by swarm intelligence, is exploited to search the best values of FDs and number of FD required for the optimal solution. After an exhaustive analysis, it is affirmed that the use of two FDs results in 21% reduction in passband error, while notch bandwidth is slightly increased by 2% only. Also, it has been observed that, in the proposed methodology, at the most 66 iterations are required for convergence to optimum solution. To examine the performance of designed notch filter using the proposed method, it has been applied for removal of power line interference from an electrocardiography (ECG) signal, and the improvement in performance is affirmed.

Keywords. Notch filter; fractional derivative (FD); swarm intelligence.

1. Introduction

Filtering of any contaminated signal is the primary requirement in numerous signal processing applications. Thus, digital filters are the vital elements in digital signal processing, which have been classified as the finite impulse response (FIR) and infinite impulse response (IIR) filters. FIR filter having a transfer function with all zero's, results in always stable system functions, and are used extensively in noise filtering and filter banks [1–3]. Generally, the FIR notch filters are prominently used in elimination of interference, caused due to an individual frequency component. In early stage of research in the notch filter design [4], three methods were adopted such as: (i) windowed Fourier series approach; (ii) frequency sampling approach, and (iii) optimized FIR filter design approach [4, 5].

In the optimized FIR filter design approach [5], a reasonable amount of the passband ripples are introduced, and the frequency sampling method leads to large interpolation

error as frequency response drastically changes across the notch point. Other familiar methods to lessen the minimum and maximum error in frequency response are McClellan-Parks-Rabiner (MPR) computer program and standard linear programming technique. MPR algorithm is generally used to design the Equiripple FIR filters, whereas standard linear programming is used for the design of Equiripple FIR notch filter, but this method fails due to huge memory requirement, and also takes more computational time for convergence. Another method for designing a FIR notch filter is the multiple exchange algorithm, also known as Equiripple FIR notch filter design method [6]. Recently, a new method has been proposed in which the 'Null width' is controlled by proper selection of zero odd order derivative constraints to obtain maximally flat linear phase FIR notch filter [7].

Fractional derivative (FD) has been employed for refining the performance in various signal processing applications like: image sharpening [8, 9], event detection in biomedical signals [10], filter design accuracy [11]. FD possesses the real time phenomena of memory effect of

*For correspondence

electrical circuits and chemical reaction, which helps in smooth tracking. Therefore, fractional derivative is extensively used by several researchers [12–20]. In [12–15], authors have proposed new methods for designing simple digital FIR filters, wideband fractional delay filters using fractional derivatives. However, in these techniques, the optimal value of order of fractional derivative is determined by trial and error method. In order to overcome this problem, authors have used different swarm based techniques such as particle swarm optimization (PSO), artificial bee colony (ABC) algorithm, cuckoo search (CS) optimization, etc. to determine the optimal value of order of FD for designing FIR filters and filter banks [16–20]. A new technique using fractional derivative and swarm based optimization has been proposed for designing IIR filters [21]. The expression of a fractional derivative consists of an integration, which is a non-local operator and that is why fractional derivative is also a non-local operator. Hence, the fractional derivative has a unique property of capturing the history of a variable. This is not easily conquerable by using only integer order derivative [22]. From the reviewed literature, it is evident that several methods have been proposed for designing FIR Notch filters. However, in these techniques, there is no provision for controlling the notch bandwidth and more accurate passband response. Therefore, there is strong motivation to develop a new design technique for FIR Notch filter that has improved passband response, and required for noise filtering for numerous signal processing applications [1–3].

Therefore, in the above context, this paper describes a new technique for designing a digital FIR Notch filter using fractional derivative and swarm intelligence with the improved passband response along with suitable notch bandwidth. For this purpose, the design problem of a digital FIR Notch filter is formulated as minimization of integral square error between the ideal response and actual response subjected to the fractional derivatives are evaluated at the prescribed frequency. For determining suitable value of order of FD, which controls the notch bandwidth and precise attenuation at the individual frequency, the constraint factor particle swarm optimization (CF-PSO) is used due to its simplicity and efficient implementation. The detailed experimental analysis has been carried out to produce an optimal choice iteration count. Statistical analysis is done, which confirms the robustness of the proposed method. To examine the efficacy of state-of-the-art with the proposed method, these algorithms have been tested for noise filtering of an ECG signal. Rest of the paper is organized as follows. Section 2 briefs the swarm intelligence based optimization methods, while section 3 contains an overview of fractional derivatives. Section 4 explains the design procedure of FIR notch filter using FDC. In section 5, the proposed problem formulation is stated and a detailed explanation of the experimental set-up and the results are given in section 6. Finally, the conclusions are provided in section 7.

2. Swarm intelligence based optimization

The modern heuristic search methods are proven as the robust in problem solving of non-differentiable, multi-modal, and non-convex problems. Particle swarm optimization (PSO) [23, 24], artificial bee colony (ABC) [25], Hybrid algorithm [26], cuckoo search optimization (CSO) [27], and similar other methods are most prominent swarm intelligence based techniques. In all these techniques, solution of non-differentiable problem is searched from a search space matrix (U), which is continuously updated. PSO is inspired by the communication of biological organism, and extensively used in numerous optimization problems due to its simple structure, efficient exploration and exploitation ability [24]. The principle equations in PSO are [23]:

$$\mathbf{V}^{k+1} = \chi[W \cdot \mathbf{V}^k + \psi_1 \cdot (\mathbf{PB}^k - \mathbf{U}^k) + \psi_2 \cdot (\mathbf{GB}^k - \mathbf{U}^k)]. \quad (1)$$

In the above Eq., k is the iteration cycle count, \mathbf{V}^k represents the current velocity matrix, associated with search space matrix (U), W is the inertia weight, C_1 and C_2 are the learning coefficients rates, which evaluate following as; $\psi_1 = C_1 \cdot rand(\cdot)$ and $\psi_2 = C_2 \cdot rand(\cdot)$, while χ is the constrained factor. \mathbf{PB} represents archive of personal best solutions discovered till k^{th} iteration, whereas \mathbf{GB} is the global best solution at k^{th} iteration. New velocity is computed by using Eq. (1), which is used for updating of U as [23]:

$$\mathbf{U}^{k+1} = \mathbf{V}^{k+1} + \mathbf{U}^k. \quad (2)$$

During the course of modification, if either value of U or V gets beyond the limit, then the respective values are restored. For restoration, either new suitable value, which is either in the predefined range or ultimate value of range is assigned to out of range elements of U or V .

3. Fractional derivative (FD)

The exhaustive research in numerous signal processing applications using fractional derivatives (FD) has been fascinated [11–22]. Riemann–Liouville, Grünwald–Letnikov and Caputo are the three most prominent definitions of FD, and Grünwald–Letnikov fractional derivative is mostly used [11, 16–19].

$$D^u y(m) = \frac{d^u y(m)}{dm^u} = \lim_{\Delta \rightarrow 0} \sum_{l=0}^{\infty} \frac{(-1)^l T_l^u}{\Delta^u} y(m - l\Delta), \quad (3)$$

and the coefficient T_l^u is computed as:

$$T_l^u = \frac{\Gamma(u+1)}{\Gamma(l+1)\Gamma(u-l+1)} = \begin{cases} 1, & l=0 \\ \frac{[u(u-1)(u-2)\dots(u-l+1)]}{1, 2, 3 \dots l}, & l \geq 1 \end{cases}. \quad (4)$$

In the above Eq., $\Gamma(\cdot)$ represents a gamma function. Based on this, FDs of trigonometric function may be computed as:

$$D^u \{A \cdot \sin(\omega x + \varphi)\} = A \cdot \omega^u \cdot \sin\left(\omega x + \varphi + \frac{\pi}{2}u\right), \quad (5)$$

and

$$D^u \{A \cdot \cos(\omega x + \varphi)\} = A \cdot \omega^u \cdot \cos\left(\omega x + \varphi + \frac{\pi}{2}u\right). \quad (6)$$

4. Design of FIR Notch Filter using FD

The design problem of a digital filter is to evaluate the coefficients of a transfer function, which reasonably satisfy the approximation to the desired response. The notch filter function is to attenuate an individual frequency component decidedly, while other frequency components are kept intact. Therefore, the ideal response of a notch filter is given by:

$$H_d(e^{j\omega}) = \begin{cases} 0, & \omega = \omega_{notch} \\ 1, & \omega \neq \omega_{notch} \end{cases}. \quad (7)$$

4.1 Design procedure of FIR notch filter

The transfer function of a causal FIR filter with order of N is defined as [11]:

$$H(e^{j\omega}) = \sum_{n=0}^N h(n) \cdot e^{-j\omega n} \quad (8)$$

The filter transfer function, defined by the above equation has a linear-phase response, if the impulse response $\{h(n)\}$ is symmetric. On this basis, whether $h(n)$ is the symmetric or anti-symmetric, FIR filters are categorized into four types as Type-1 to Type-4 [11]. In this paper, Type-1 filter, whose impulse response is symmetric with even order (N) is considered. Due to symmetric response, Eq. (8) may be reframed as:

$$\begin{aligned} H(e^{j\omega}) &= e^{-j\omega L} \left\{ h(L) + 2 \cdot \sum_{n=0}^{L-1} h(n) \cdot \cos(\omega(L-n)) \right\}, \\ &= H_o(\omega) \cdot e^{-j\omega L}. \end{aligned} \quad (9)$$

Here, $L = N/2$, $H_o(\omega)$ is the magnitude response of a FIR filter, which can be rewritten as:

$$H_o(\omega) = \sum_{l=0}^L b(l) \cos(\omega l), \quad (10)$$

where

$$b(l) = \begin{cases} h(L) & l = 0 \\ 2 \cdot h(L-l) & 1 \leq l \leq L \end{cases} \quad (11)$$

Eq. (10) may also be represented in matrix form as:

$$H_o(\omega) = \mathbf{b}^T \cdot \mathbf{C}(\omega), \quad (12)$$

where

$$\mathbf{b} = [b(0) \ b(1) \ \dots \ b(L)], \quad (13)$$

and

$$\mathbf{C}(\omega) = [1 \ \cos(\omega) \ \dots \ \cos(L\omega)]. \quad (14)$$

In Eq. (12), T denotes the transpose of a vector. Now, in case of notch filter, the design problem is reduced to evaluate the coefficients of filter (\mathbf{b}) such that it should eliminate the desired individual frequency component and has unity magnitude for the rest of other frequencies. Now, the filter coefficients are determined by minimizing an error function, defined as:

$$\begin{aligned} J(\mathbf{b}) &= \int_{\omega \in ROI} (H_d(\omega) - H_o(\omega))^2 d\omega, \\ &= \mathbf{b}^T \mathbf{Q} \mathbf{b} - 2\mathbf{p}^T \mathbf{b} + \alpha, \end{aligned} \quad (15)$$

where, ROI is the region of interest, matrix \mathbf{Q} , vector \mathbf{p} , and scalar α are given by [11]:

$$\mathbf{Q} = \int_{\omega \in ROI} \mathbf{C}(\omega) \cdot \mathbf{C}(\omega)^T d\omega, \quad (16)$$

$$\mathbf{p} = \int_{\omega \in ROI} (H_d(\omega) \cdot \mathbf{C}(\omega)) d\omega, \quad (17)$$

and

$$\alpha = \int_{\omega \in ROI} \{H_d(\omega)\}^2 d\omega. \quad (18)$$

Now on the differentiation of Eq. (15) w.r.t. \mathbf{b} , and equating to zero, results in the conventional least squares design solution as $\mathbf{b}_{LS} = \mathbf{Q}^{-1} \cdot \mathbf{p}$. To yield more accuracy at notch frequency, the following constraints are employed on the response $H_o(\omega)$ at the given frequency as [11]:

$$H_o(\omega_0) = H_d(\omega_0) = 0, \quad (19)$$

and

$$DH_o(\omega)|_{\omega=\omega_0} = 0. \quad (20)$$

In case of a notch filter, the fractional derivative evaluated at ω_0 must satisfy the constraint defined as [11]:

$$D^u H_o(\omega)|_{\omega=\omega_0} = \beta(u-1) \quad (21)$$

In Eq. (21), u is the order of FD and β is the recommended constant, and for this work, it is taken as 30 [11].

By using Eqs. (6) and (10), the fractional derivative of $D^u H_o(\omega)$ can be computed as:

$$\begin{aligned} D^u H_o(\omega) &= \frac{d^u \left(\sum_{l=0}^L b(l) \cos(\omega l) \right)}{d\omega^u} = \sum_{l=0}^L b(l) \frac{d^u \cos(\omega l)}{d\omega^u}, \\ &= \sum_{l=0}^L b(l) \cdot l^u \cdot \cos\left(\omega l + \frac{\pi u}{2}\right) = \mathbf{b}^T \cdot \mathbf{c}(\omega, u), \end{aligned} \quad (22)$$

where, the vector $\mathbf{c}(\omega, u)$ is defined as:

$$\mathbf{c}(\omega, u) = \begin{bmatrix} 1^u \cdot \cos\left(\omega + \frac{\pi u}{2}\right) \\ 2^u \cdot \cos\left(2\omega + \frac{\pi u}{2}\right) \\ \vdots \\ L^u \cdot \cos\left(L\omega + \frac{\pi u}{2}\right) \end{bmatrix} \quad (23)$$

By using the equations (12), (22) and (23), the constraint equations (19), (20), and (21) are rewritten in matrix form as:

$$\mathbf{CB}_x \cdot \mathbf{b} = \mathbf{F}. \quad (24)$$

where

$$\mathbf{CB}_x = [\mathbf{C}^T(\omega_0) \quad \mathbf{c}^T(\omega_0, 1) \quad \mathbf{c}^T(\omega_0, u)]^T, \quad (25)$$

and

$$\mathbf{F} = [0 \quad 0 \quad \beta(u-1)]^T. \quad (26)$$

The constraints defined in Eq. (19) is used for achieving exact zero magnitude response at the reference notch frequency (ω_0), while Eq. (20) is used to make first order

derivative equal to be zero [11]. And the constraint defined by Eq. (21) aids in controlling 3-dB notch bandwidth [11]. Therefore, it is possible to adjust the notch bandwidth by tuning the value of u .

On merging the objective function given by Eq. (15), with constraints in Eq. (24), the definition of design problem of notch filter is expressed as:

$$\begin{aligned} \text{Minimize } J(\mathbf{b}) &= \mathbf{b}^T \mathbf{Q} \mathbf{b} - 2\mathbf{p}^T \mathbf{b} + \alpha, \\ \text{subjected to } \mathbf{CB}_x \cdot \mathbf{b} &= \mathbf{F}. \end{aligned} \quad (27)$$

The Lagrange multiplier method [11, 16] gives the optimal solution of such constrained optimization problem, and is computed as:

$$\begin{aligned} \mathbf{b}_{opt} &= \mathbf{Q}^{-1} \cdot \mathbf{p} - \mathbf{Q}^{-1} \cdot \mathbf{CB}_x^T \\ &\cdot (\mathbf{CB}_x \cdot \mathbf{Q}^{-1} \cdot \mathbf{CB}_x^T)^{-1} [\mathbf{CB}_x \cdot \mathbf{Q}^{-1} \cdot \mathbf{p} - \mathbf{F}]. \end{aligned} \quad (28)$$

This is a closed-form solution and effortlessly computable. The computational complexity of this method includes two terms, one is the computation of conventional least squares solution, which is $\mathbf{Q}^{-1} \cdot \mathbf{p}$. Second term is the product of $\mathbf{Q}^{-1} \cdot \mathbf{CB}_x^T (\mathbf{CB}_x \cdot \mathbf{Q}^{-1} \cdot \mathbf{CB}_x^T)^{-1}$ and $[\mathbf{CB}_x \cdot \mathbf{Q}^{-1} \cdot \mathbf{p} - \mathbf{F}]$, in which the computation of inverse of a matrix $(\mathbf{CB}_x \cdot \mathbf{Q}^{-1} \cdot \mathbf{CB}_x^T)^{-1}$ is expensive task. However, size of $\mathbf{CB}_x \cdot \mathbf{Q}^{-1} \cdot \mathbf{CB}_x^T$ is small of $i \times i$, where $i =$ (integral order) + (order of FD terms), which are user defined and smaller. Therefore, the computational complexity of second term is also small. Authors in [11], have designed a notch filter using single FD term and have shown its effect on notch bandwidth (W_{notch}) as depicted in figure 1(a), and the corresponding frequency response in figure 1(b). The fidelity parameter, defined as passband error (er_p) is computed as:

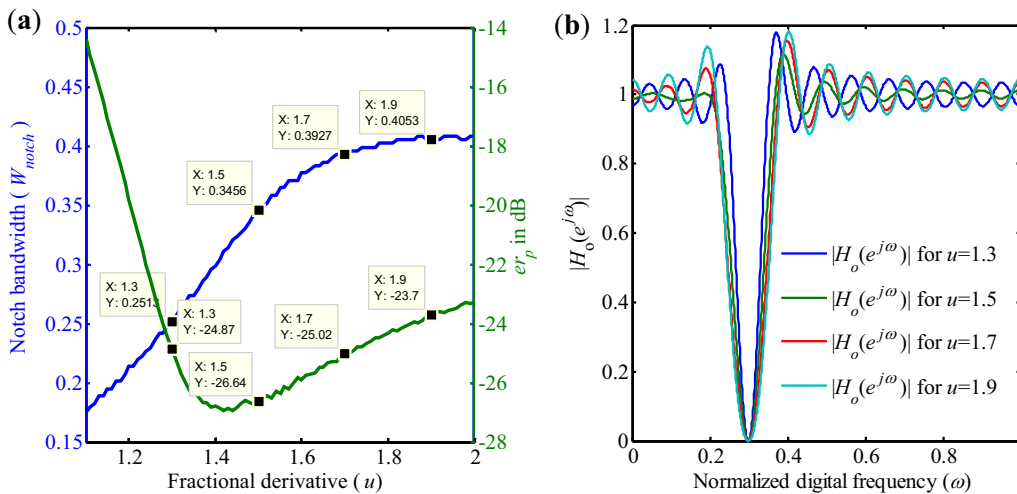


Figure 1. (a) Variation of notch bandwidth and er_p using single FD (u), (b) FIR notch filter frequency response for different FD value (u) = 1.3, 1.5, 1.7, and 1.9.

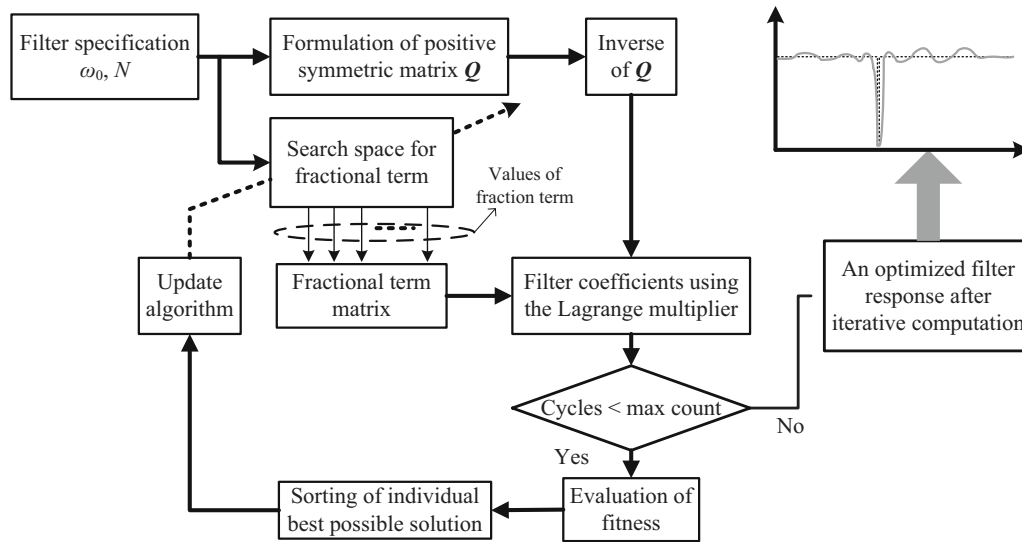


Figure 2. Block diagram of the proposed method.

$$er_p = \frac{1}{\pi} \left(\int_0^{\omega_c^1} (H_d(e^{j\omega}) - H_o(e^{j\omega}))^2 d\omega + \int_{\omega_c^2}^{\pi} (H_d(e^{j\omega}) - H_o(e^{j\omega}))^2 d\omega \right), \tag{29}$$

where, ω_c^1 and ω_c^2 are the lower and upper cut-off frequencies, given by:

$$H_o(\omega_c^l) = \sqrt{2}, \text{ where } l = 1 \text{ or } 2. \tag{30}$$

It is found that er_p varies with W_{notch} such that at $u=1.01$, W_{notch} is the minimum, however er_p is having maximum value. When u is increased, W_{notch} starts increasing with the reduction in er_p . Authors in [11], have used the step size of 0.1 for FD values, however it is observed that the step size with 0.01 attains more accurate results as shown in figure 1. When FD value is incremented with 0.01 and u is equal to 1.42, er_p attains it's the best value of -26.96 dB, which is the minimum and starts increasing, when u is greater than 1.45. Therefore, more accuracy with less W_{notch} may be achieved with high precision value of u , however it would be computational expensive in such approach. Therefore, swarm intelligence based modern heuristic approach is a suitable choice for obtaining the optimal solution, which simultaneously reduce the W_{notch} and er_p .

5. Problem formulation

In this work, the design problem of a notch filter response with less er_p and narrow W_{notch} is constructed as a minimization of Eq. (29). Here, W_{notch} is controlled by the value

and number of u . Therefore, particle swarm optimization (PSO) is used for finding the suitable FD value and number of FD used. The proposed method using FD and swarm intelligence technique is completely described in figure 2.

5.1 Particle swarm optimization

In PSO, the optimal solution is found by controlling the local and global search using search space, which is formed in the initial stage. In the proposed methodology, for acquiring more accurate solution, FD values are searched using CI-PSO. Therefore, search space (U) is formulated by a matrix containing elements uniformly distributed in the range of lower (U_l) and upper (U_u) bound, defined as:

$$U = U_l - (U_u) \oplus rand(0, 1). \tag{31}$$

Each set of a row vector of U is the possible combination of FD values for the evaluations (25) and (26). This approach ensures the independency on step size, and the self-learning mechanism of PSO helps in finding the appropriate best value. Also, with this approach, more fractional order based design can be tested with less computational cost.

5.2 The algorithmic steps to be followed for the proposed method based on FD using PSO

The complete design scheme can be framed using following steps:

- Step 1: Declare the filter specifications like: filter order (K), notch frequency (ω_{notch}), and FD order.
- Step 2: Define the ideal repose ($H_d(e^{j\omega})$) on the basis of Eq. (7).

Step 3: Compute a vector $\mathbf{C}(\omega)$, matrix \mathbf{Q} , and vector \mathbf{p} using Eqs. (14), (16), and (17). Also, evaluate the inverse of \mathbf{Q} and store for further computations.

Step 4: Initialize the search space matrix (\mathbf{U}) and velocity matrix (\mathbf{V}) with uniformly distributed random matrix within the limit of U_l to U_u as defined by Eq. (31).

Step 5: Evaluate the matrix \mathbf{CB}_x and vector \mathbf{F} for each solution containing FD value in a row vector of \mathbf{U} . After then, compute an optimal filter coefficients \mathbf{b}_{opt} using Eq. (28).

Step 6: Compute the frequency response, followed by the computation of error function, defined by Eq. (29). Store these error values as **Local Best Error**, and find out the least value of error from these. Assign this least value to **Global Best Error**. Assign the content of matrix \mathbf{U} into matrix \mathbf{PB} (personal best solutions). In the end, the solution corresponding to **Global Best Error** is kept in \mathbf{GB} (global best solutions).

Step 7: Update \mathbf{V} using Eq. (1), followed by the update of \mathbf{U} using Eq. (2).

Step 8: Restore these values, if they move beyond U_l and U_u .

Step 9: Using the updated \mathbf{U}^{n+1} , evaluate \mathbf{CB}_x , \mathbf{F} , then compute the filter coefficients \mathbf{b}_{opt} using Eq. (28) similarly as performed in step 5.

Step 10: Consider those solutions of updated \mathbf{U} , whose er_p is less than that of **Local Best Error**. After this, sort out the minimum value of **Local Best Error**, and if it is smaller than the current **Global Best Error**, then update the **Global Best Error** and \mathbf{GB} with respective value and solution.

Step 11: Repeat steps 4 to 9 till maximum number iteration are completed or er_p is dropped beyond tolerable limit.

6. Experimental set-up, results and discussion

This section elaborates the experimental set made for the design of a FIR notch filter using FD with PSO. For this purpose, MATLAB® 2014 is used on Genuine Intel (R) CPU i7 3770 @ 3.40 GHz, 4GB RAM. The grid size of 500 equally spaced sample for normalized digital frequency is taken during the experiments.

6.1 Statistical analysis of the proposed method

In PSO, size of search space is the key factor and depends on computation time. If \mathbf{U} is smaller, it results in less computation time and grows almost abruptly as the size increases. The size of \mathbf{U} is defined by the dimensionality (D) and number of solution (NS). Therefore, for obtaining the solution in reasonable computational time (t), it is required to set NS reasonably best by experimental evaluation. Therefore, in this section, various experiments have been performed to demonstrate the effect of D and NS on er_p , W_{notch} , t and convergence. Number of FDs is considered as D , and 30 trials of experiments are performed for

possible combinations of D and NS . The mean of different fidelity parameters such as er_p , W_{notch} , is computed for analysis, and it is observed that the er_p and W_{notch} are increased for D greater than 2 as illustrated in figures 3(a) and 3(b). The mean value of t is increased with the increase in number of search space solution (NS) and D as shown in figure 3(c). However, the performance measured in terms of er_p and W_{notch} are being intact irrespective of the value of NS . Therefore, $NS = 10$ is an optimal choice for acquiring the best results in reasonable computation time. The mean value of rate of decay for er_p w.r.t. iteration for different number of FD, denoted by D is shown in figure 3(d).

The computation time taken during optimization process depends on order of filter, number of fractional derivatives and search space size. From the above discussion, it is clear that two fractional derivatives are the best suited for minimum er_p . Also, swarm size equal to 10 achieves the same performance as achieved with other higher value of it, which is also observed in figure 4 that shows the variation in the best and worst performance for different swarm size values. It can be perceived that er_p quickly converges into steady state as shown in figure 3(d). To find the practical value of iteration cycles, the convergence profile is differentiated w.r.t. number of iteration (n). The value of n after, which is:

$$\frac{\partial(10 \cdot \log_{10}[\text{mean}(er_p)])}{\partial k} = \text{constant}, \quad (32)$$

and this is the best appropriate choice. On performing the above operation, the suitable value of n is found to be 13 and 66 for $D = 1$ and 2, respectively as shown in figure 5(a). It can also be observed that $D = 2$ archives 21% reduction in er_p , when compared with $D = 1$; however, slight increment in W_{notch} of 2% has occurred as depicted in figure 5(b). The frequency response of FIR notch filter designed by the proposed method is shown in figure 5(c) with notch at $\omega = 0.3$.

6.2 Comparative analysis

Based on the results obtained in the above analysis, robustness of the proposed methodology has been tested by designing different order notch filter with different notch frequencies. The maximum iteration is kept to be 70, and order is varied from 10 to 80 with increment of 10. The proposed method has been tested using single FD and two FDs, the results obtained in single FD is summarized in table 1, while table 2 summarizes the results obtained in case of two FDs.

6.3 Application in an electrocardiogram signal filtering

ECG signal processing is the most eminent and consistently evolving stream in bio-medical signal processing [28, 29].

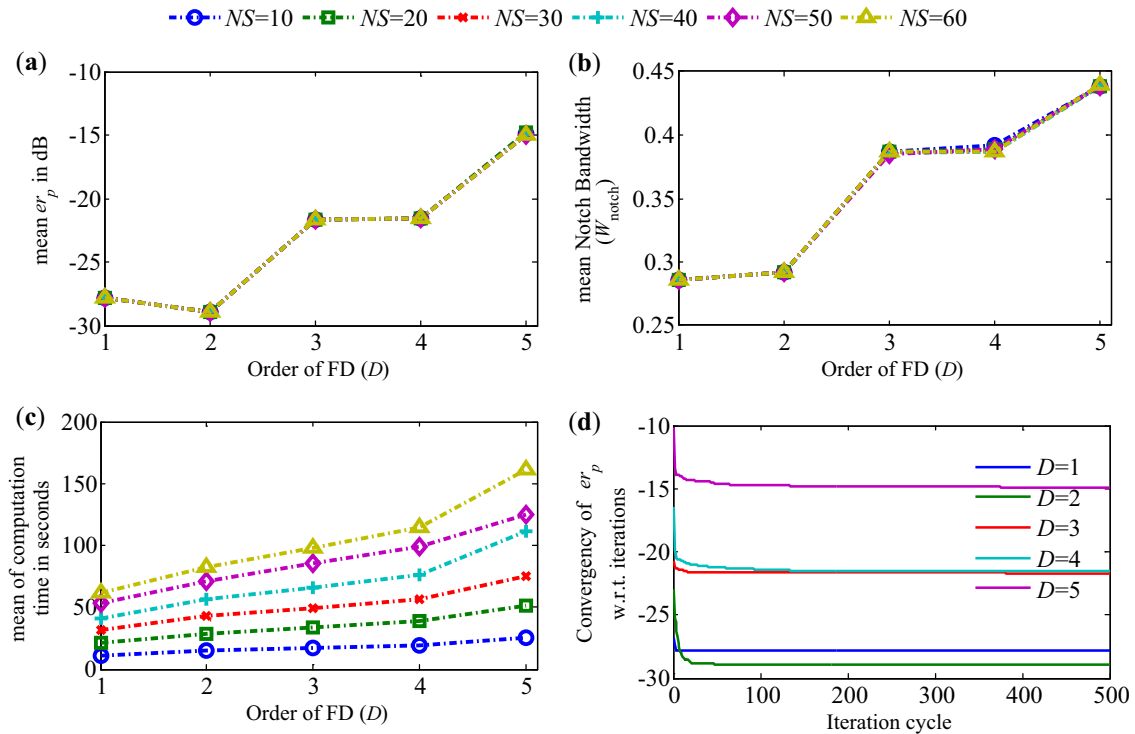


Figure 3. (a) Variation on mean value of er_p w.r.t. number of FDs (D) with different search space size (NS), (b) Variation on mean value of W_{notch} w.r.t. D with N , (c) Variation on mean value of computation time w.r.t. D with NS , (d) Convergence of er_p w.r.t. to iteration cycle for $NS = 10$ for different values of D .

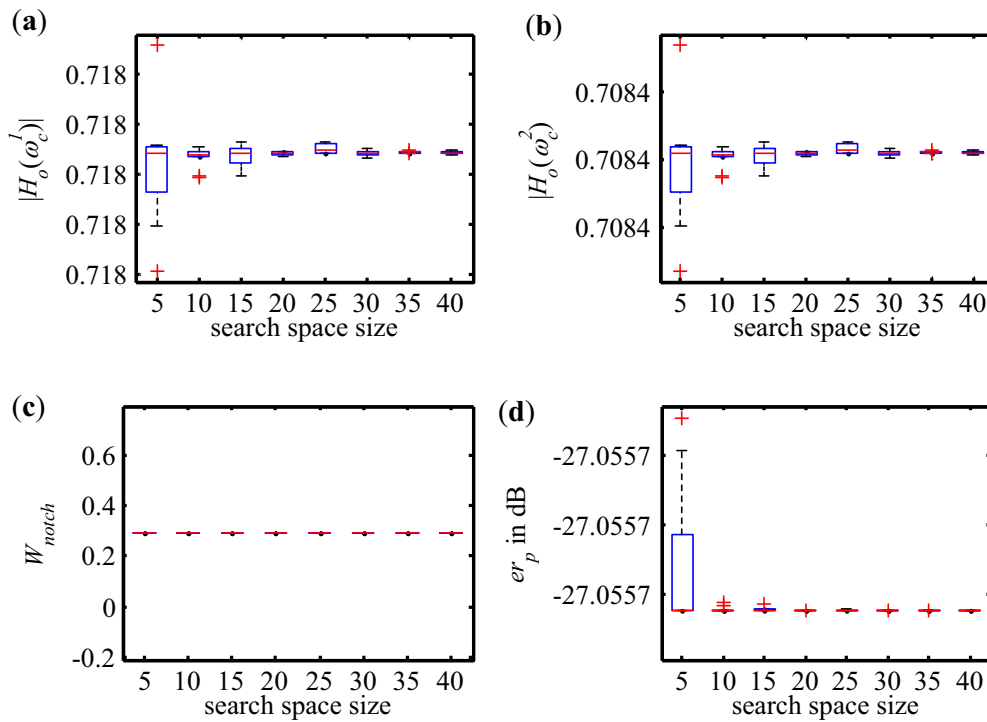


Figure 4. Variations in values of worst, median and best values obtained in 30 independent experiments performed. (a) Magnitude at lower cutoff frequency $|H_o(\omega_c^1)|$. (b) Magnitude at upper cutoff frequency $|H_o(\omega_c^2)|$. (c) W_{notch} . (d) er_p in dB.

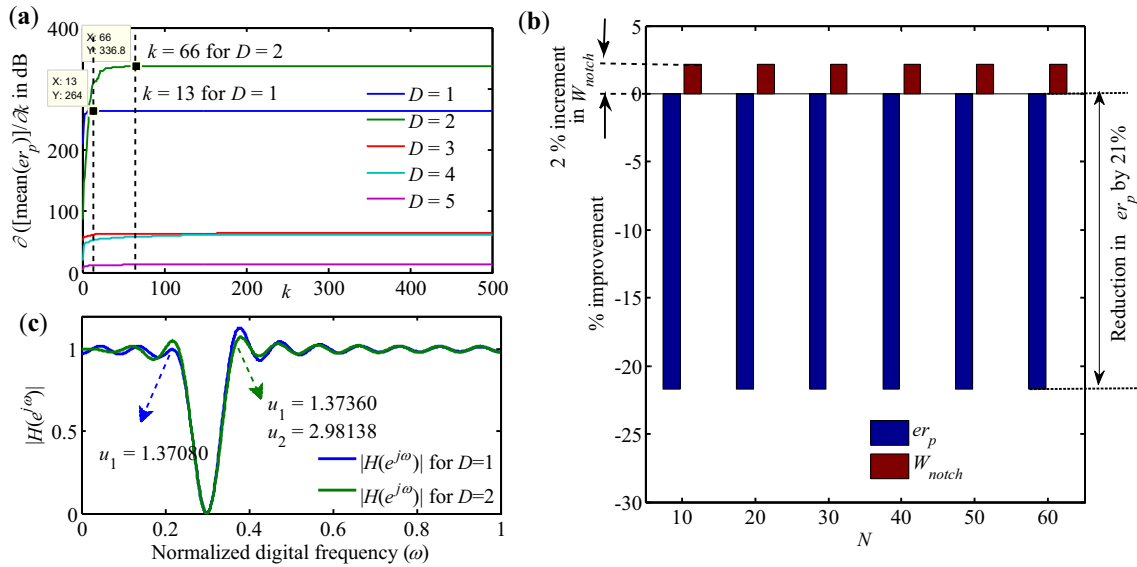


Figure 5. (a) Computation of optimal iteration count, (b) comparison of improvement in er_p and W_{notch} between design methodology of proposed method for $D = 1$ and $D = 2$, respectively, and (c) frequency response of notch filter designed using proposed methodology for $D = 1$ and $D = 2$.

Table 1. Performance of the proposed method with different order and notch frequencies for single FD.

Filter Order	$\omega_{notch} = 0.25$		$\omega_{notch} = 0.60$		$\omega_{notch} = 0.80$	
	er_p	W_{notch}	er_p	W_{notch}	er_p	W_{notch}
10	0.4672	0.4618	4.4578	0.4492	-1.8395	0.4650
20	-16.7001	0.5027	-17.8581	0.4681	-17.9987	0.4650
30	-22.7098	0.3644	-22.9152	0.3613	-24.0427	0.3424
40	-27.7795	0.2890	-27.8404	0.2859	-27.9608	0.2859
50	-29.2317	0.2419	-28.7865	0.2482	-29.2498	0.2450
60	-29.6564	0.2168	-30.0488	0.2105	-29.9702	0.2073
70	-30.7874	0.1948	-30.6193	0.1916	-30.7997	0.1854
80	-31.5196	0.1696	-31.4998	0.1696	-31.4098	0.1696

Table 2. Performance of proposed method with different order and notch frequencies for two FDs.

Filter Order	$\omega_{notch} = 0.25$		$\omega_{notch} = 0.60$		$\omega_{notch} = 0.80$	
	er_p	W_{notch}	er_p	W_{notch}	er_p	W_{notch}
10	0.4703	0.4609	4.3841	0.4496	-1.9289	0.5759
20	-16.6999	0.5023	-17.8597	0.4681	-18.0006	0.4618
30	-22.8581	0.3644	-23.0752	0.3641	-24.2116	0.3456
40	-28.8163	0.2922	-28.9032	0.2922	-29.0729	0.2890
50	-31.3496	0.2532	-30.9587	0.2576	-31.2656	0.2513
60	-31.5176	0.2262	-31.5415	0.2205	-31.6355	0.2212
70	-31.4722	0.1970	-31.5991	0.1963	-31.6769	0.1910
80	-31.6336	0.1715	-31.6621	0.1696	-31.7297	0.1715

One of the important parts is extraction of QRS complex and analyzing its characteristics to diagnose the irregularities in the heart rhythm. The notch filters are widely used

in application, where an individual harmonic elimination is required such as interference of power line in an electrocardiogram (ECG) recording, open-loop voltage across the

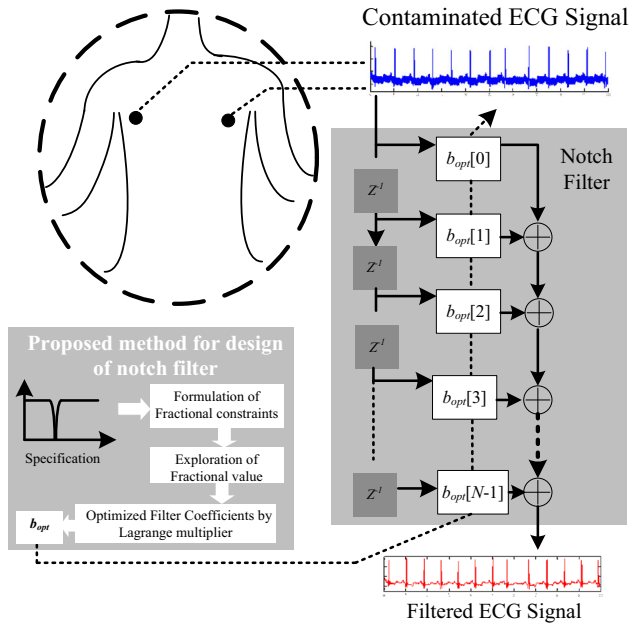


Figure 6. Proposed methodology for design of notch filter for power line interference removal from ECG signal.

input of an analog instrument and many such [7]. In this section, power line interference in an ECG is filtered by

using the designed notch filter as shown in figure 6. First, an artificial ECG is generated, and then contaminated with power line interference, and finally filtered using the designed notch filter. This experiment is also performed on ECG recorded signals from MIT-BIH [30]. The quality of filtering is judged by finding the value of following [31, 32]: mean squared error (*MSE*):

$$MSE = \frac{1}{N_s} \sum |x(n) - \hat{x}(n)|^2, \text{ where } N_s \text{ number of samples,} \tag{33}$$

where N_s number of samples. Percent root mean square difference (*PRD*):

$$PRD = \left(\frac{\sum |x(n) - \hat{x}(n)|^2}{\sum [x(n)]^2} \right)^{1/2} \cdot 100, \tag{34}$$

and signal to noise ratio (*SNR*):

$$SNR \text{ dB} = 10 \cdot \log_{10} \left(\frac{\sum [x(n)]^2}{\sum |x(n) - \hat{x}(n)|^2} \right) \tag{35}$$

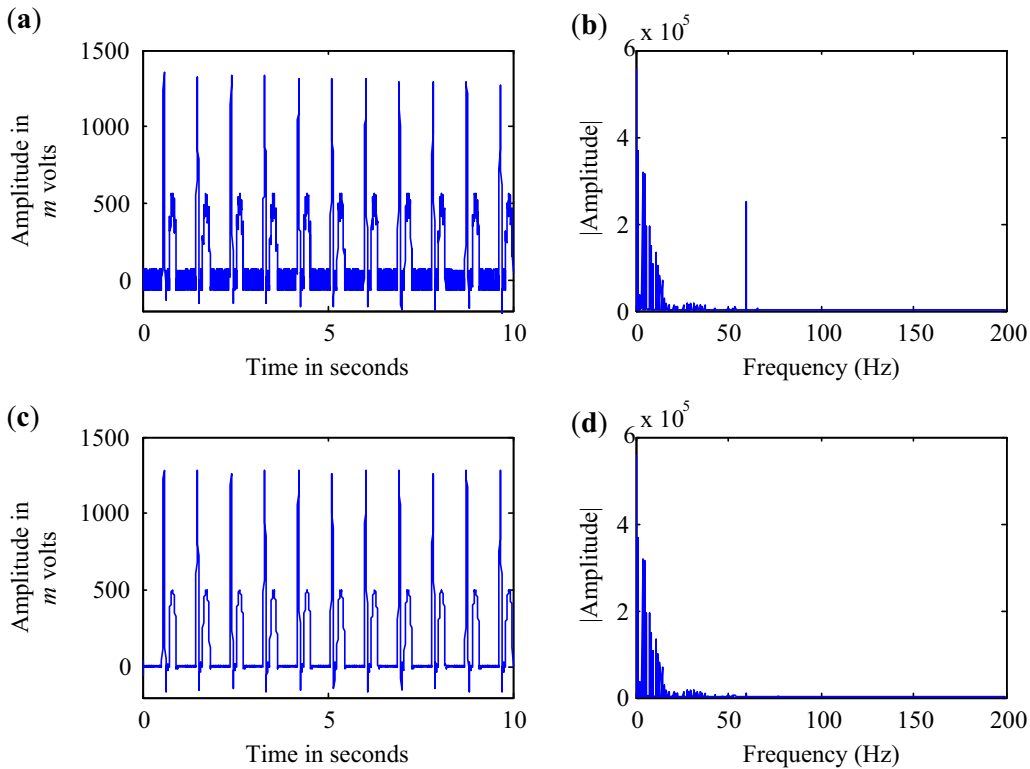


Figure 7. (a) Contaminated synthesized ECG signal, using 400 Hz sampling frequency, with 60 Hz interference, (b) Contaminated synthesized ECG signal spectrum, (c) filtered synthesized ECG signal by filter designed using proposed methodology, and (d) spectrum of filtered ECG signal.

An ECG signal with the sampling rate of 400 Hz is synthesized, and 60 Hz interference has been introduced as shown in figures 7(a) and 7(b). This contaminated signal has been filtered using the notch filters designed using FD approach as suggested in [11], and by the proposed design approach. It can be perceived from table 3 that filtering using notch filter designed by the proposed methodology using second order FD obtains better value of fidelity parameters. This is possible due to the filter designed with second order FD, which has better passband accuracy with optimal notch bandwidth. The obtained performance has been compared and summarized in table 3. It is evident that with the second order fractional derivative approach, filter achieves more accurate filtering results when compared to single order fractional derivative approach as given in [11].

The designed filters are also tested for real time ECG signal taken from [30]. These signals are mixed with 50 and 60 Hz power line signal. The sampling frequency of ECG signal is 360 Hz, and if these signals are interfered by 50 Hz power line signal, then notch filter with $(\frac{50}{360}) \times 2 = 0.2778$, normalized digital frequency is required. Whereas in case of 60 Hz power line signal interference, it is required that notch frequency should be 0.3333 normalized. In trial based approach [11], it took 3.7813 seconds for completion, and then additional time in sorting of best solution from entire listed output solutions. If same approach is adopted for two derivatives, then it would take more computation time. Whereas the proposed technique takes maximum of 7.39 seconds for obtaining the optimal values of u for $D = 2$. ECG signals are

Table 3. Performance evaluation of notch filter designed by proposed technique in filtering of synthesise ECG signal.

Design Technique	FD Order	Fractional value (u)	MSE	PRD	SNR
FD [11]	1	1.300000	164.2868	0.1704	27.6859
FD [11]	1	1.500000	88.7068	0.0920	30.3623
FD [11]	1	1.700000	165.6624	0.1718	27.6497
FD [11]	1	1.900000	386.6205	0.4010	23.9690
Proposed	1	1.370809	116.2129	0.1205	29.1894
Proposed	2	1.373603, 2.981385	81.9271	0.0850	30.7076

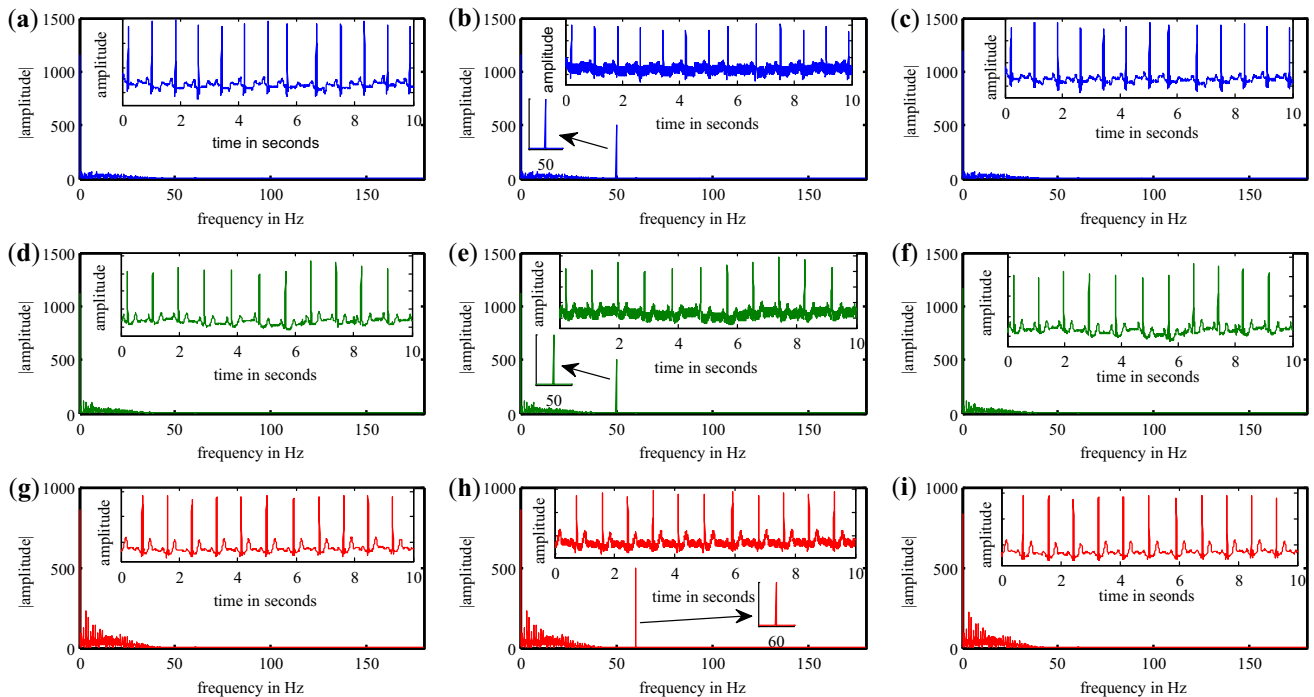


Figure 8. (a) ECG signal record MIT BIH 100, (b) contaminated ECG signal with 50 Hz interference, (c) filtered ECG signal, (d) ECG signal record MIT BIH 101, (e) contaminated ECG signal with 50 Hz interference, and (f) filtered ECG signal. (d) ECG signal record MIT BIH 103, (e) contaminated ECG signal with 60 Hz interference, and (g) filtered ECG signal. Filtering has been performed using filter designed using proposed methodology with $D = 2$.

Table 4. Performance evaluation of designed notch filter using real time recorded ECG signal take from ECG database from [30].

Signal	Sinusoid frequency	Technique	MSE	PRD	SNR
MIT-BIH 100	50 Hz	1-D using approach in [11]	0.00099	0.75484	21.22147
MIT-BIH 101	50 Hz		0.00086	0.62526	22.03940
MIT-BIH 106	50 Hz		0.00189	1.02824	19.87906
MIT-BIH 103	60 Hz		0.00049	0.32498	24.88146
MIT-BIH 104	60 Hz		0.00089	0.67221	21.72498
MIT-BIH 105	60 Hz	2-D using proposed approach	0.00032	0.21904	24.72498
MIT-BIH 100	50 Hz		0.00076	0.57848	22.37713
MIT-BIH 101	50 Hz		0.00066	0.47975	23.18982
MIT-BIH 106	50 Hz		0.00169	0.92277	20.34904
MIT-BIH 103	60 Hz		0.00045	0.29544	25.29529
MIT-BIH 104	60 Hz		0.00085	0.63760	21.95449
MIT-BIH 105	60 Hz		0.00085	0.63760	26.59480

contaminated by both 50 and 60 Hz interference and filtered as depicted in figure 8, and performance is summarized in table 4. The optimal value of u is found to be 1.389992 and 2.956899 in case of notch frequency equals to 0.2778, while it should be 1.390799 and 2.973813 for normalized notch frequency equals to 0.3333.

7. Conclusions

In this paper, a new design approach using fractional derivatives, which are explored using CFI-PSO, is presented. The exhaustive experimentation results have revealed that two fractional derivatives with second order derivative is sufficient for the design of optimal notch filter. There is reduction in passband error by 21%, however there is nominal increment of notch bandwidth by 2.1%, when compared with the double FD with single FD design approach. The thorough analysis made for analyzing the effect of swarm size reveals that swarm size consisting of ten solutions is the best, which also results in less computation time. On differentiating the mean of convergence w.r.t. iteration, it gives the reasonable iteration count for the convergence and found to be 13 for single and 66 for two FD based design. The designed filter is tested for power line interference removal, and was found to be very efficient.

Acknowledgement

This work was supported by the National Research Foundation (NRF) of Korea grant funded by the Korean government (MSIP) (NRF-2018R1A2A1A19018665).

References

- [1] Roy S C D, Jain S B and Kumar B 1997 Design of Digital FIR Notch Filters from Second Order IIR Prototype. *IETE J. Res.* 43(4): pp. 275–279
- [2] Sharma I, Kuldeep B, Kumar A and Singh V K 2016 Performance of swarm based optimization techniques for designing digital FIR filter: A comparative study. *Eng. Sci. Technol. Int. J.* 19(3): 1564–1572
- [3] Kumar A and Kuldeep B 2012 Design of M-channel cosine modulated filter bank using modified Exponential window. *J. Franklin Inst.* 349(3): 1304–1315
- [4] Yu T H, Mitra, S K and Babic H 1990 Design of linear phase FIR notch filters. *Sadhana* 15(3): 133–155
- [5] Hirano K, Nishimura S and Mitra S 1974 Design of digital notch filters. *IEEE Trans. Circuits Syst.* 21(4): 540–546
- [6] Tseng, C C and Pei S C 1990 Design of an equiripple FIR notch filter using a multiple exchange algorithm. *Signal Processing* 75(3): 225–237
- [7] Deshpande R, Jain S B and Kumar B 2008 Design of maximally flat linear phase FIR notch filter with controlled null width. *Signal Processing* 88(10): 2584–2592
- [8] Tseng C C and Lee S L 2012 Digital image sharpening using fractional derivative and mach band effect. In: *Proceedings International Symposium on Circuits and Systems*, IEEE, Seoul, South Korea, pp. 2765–2768
- [9] Mathieu B, Melchior P, Oustaloup A and Ceyral C 2003 Ceyral, Fractional differentiation for edge detection. *Signal Processing* 83(11): 2421–2432
- [10] Ferdi Y, Herbeuval J P, Charef A and Boucheham B. 2003 R wave detection using fractional digital differentiation. *ITBM-RBM.* 24(5): 273–280
- [11] Tseng C C and Lee S L 2012 Design of linear phase FIR filters using fractional derivative constraints. *Signal Processing* 92(5): 1317–1327
- [12] Tseng C C and Lee S L 2013 Fractional Derivative Constrained Design of FIR Filter with Prescribed Magnitude and Phase Responses. In: *Proceedings of European Conference on Circuit Theory and Design*, IEEE, Dresden, Germany, pp. 1–4
- [13] Tseng C C and Lee S.L 2010 Design of wideband fractional delay filters using derivative sampling method. *IEEE Trans. Circuits Syst. I Regul. Pap.* 57(8): 2087–2098
- [14] Tseng C C 2001 Design of fractional order digital FIR differentiators. *IEEE Signal Process. Lett.* 8(3): 77–79
- [15] Tseng C C and Lee S L 2012 Designs of Fixed-Fractional-Delay Filters Using Fractional-Derivative Constraints. *IEEE Trans. Circuits Syst. II Express Briefs.* 59(10): 683–687

- [16] Baderia K, Kumar A and Singh G K 2015 Design of multi-channel filter bank using ABC optimized fractional derivative constraints. In: *Proceedings of International Conference on Communications and Signal Processing*, Melmaruvathur, India, pp. 0490–0494
- [17] Baderia K, Kumar A and Singh G K 2015 Hybrid method for designing digital FIR filters based on fractional derivative constraints. *ISA Trans.* 58: 493–508
- [18] Kuldeep B, Singh V K, Kumar A and Singh G K 2015 Design of two-channel filter bank using nature inspired optimization based fractional derivative constraints. *ISA Trans.* 54: 101–116
- [19] Kuldeep B, Kumar A and Singh G K 2015 Design of quadrature mirror filter bank using Lagrange multiplier method based on fractional derivative constraints. *Eng. Sci. Technol. Int. J.* 18(2): 235–243
- [20] Kuldeep B, Kumar A and Singh G K 2015 Design of Multi-channel Cosine-Modulated Filter Bank Based on Fractional Derivative Constraints Using Cuckoo Search Algorithm. *Circuits, Syst. Signal Process.* 34(10): 3325–3351
- [21] Agrawal N, Kumar A and Bajaj V 2017 Design Method for Stable IIR Filters With Nearly Linear-Phase Response Based on Fractional Derivative and Swarm Intelligence *IEEE Trans. Emerg. Top. Comput. Intell.* 1(1): 464–477
- [22] Charef A, Djouambi A and Idiou D 2014 Linear fractional order system identification using adjustable fractional order differentiator. *IET Signal Process* 8(4): 398–409
- [23] Poli R, Kennedy J and Blackwell T 2007 Particle swarm optimization An overview. *Swarm Intelligence* 1(1): 33–57
- [24] Ahirwal M K, Kumar A and Singh G K 2014 Adaptive filtering of EEG/ERP through noise cancellers using an improved PSO algorithm. *Swarm Evol. Comput.* 14: 76–91
- [25] Karaboga D and Akay B 2009 A comparative study of Artificial Bee Colony algorithm. *Appl. Math. Comput.* 214(1): 108–132
- [26] Rafi S M, Kumar A and Singh G K 2013 An improved particle swarm optimization method for multirate filter bank design. *J. Franklin Inst.* 350(4): 757–769
- [27] Agrawal N, Kumar A, Bajaj V and Singh G K 2018 Design of Bandpass and Bandstop Infinite Impulse Response Filters using Fractional Derivative. *IEEE Trans. Ind. Electron.* 66(2): 1285–1295
- [28] Dai H, Yin L and Li Y 2016 QRS residual removal in atrial activity signals extracted from single lead: a new perspective based on signal extrapolation. *IET Signal Process.* 10(9): 1169–1175
- [29] Khamis H, Weiss R, Xie Y, Chang C W, Lovell N H and Redmond S J 2016 QRS Detection Algorithm for Telehealth Electrocardiogram Recordings *IEEE Trans. Biomed. Eng.* 63(7): 1377–1388
- [30] PhysioBank ATM, MIT-BIH arrhythmia ECG signal database, (n.d.).
- [31] Kumar R, Kumar A and Pandey R K 2013 Beta wavelet based ECG signal compression using lossless encoding with modified thresholding. *Comput. Electr. Eng.* 39(1): 130–140
- [32] Kumar R, Kumar A and Singh G K 2016 Hybrid method based on singular value decomposition and embedded zero tree wavelet technique for ECG signal compression. *Comput. Methods Programs Biomed.* 129: 135–148

Fast Mixed Integer Quadratic Programming for Sparse Signal Estimation

SANGJUN PARK^{ID} AND HEUNG-NO LEE^{ID}, (Senior Member, IEEE)

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Heung-No Lee (heungno@gist.ac.kr)

This work was supported by the National Research Foundation of Korea through the Korean Government (MSIP) under Grant NRF-2018R1A2A1A19018665.

ABSTRACT It has been recently shown that the l_0 -norm problem can be reformulated into a mixed integer quadratic programming (MIQP) problem. CPLEX, a commercial optimization software package that can solve integer programming problems, is used to find the global solution to this MIQP problem for sparse signal estimation. However, CPLEX uses an exhaustive approach to search a feasible space to this MIQP problem. Thus, its running time grows exponentially as the problem dimension grows. This means that CPLEX quickly becomes computationally intractable for higher dimension problems. In this paper, we aim to propose a fast first-order-type method for solving this MIQP problem based on the alternating direction method. We conduct extensive simulations to demonstrate that: 1) our method is used to estimate a sparse signal by solving this problem and 2) our method is computationally tractable for problem dimensions up to the order of 1 million.

INDEX TERMS Alternating direction method, compressed sensing, mixed integer quadratic program.

I. INTRODUCTION

Compressed sensing [1] has attracted attention because it allows for the acquisition of signal samples at a rate lower than the Nyquist rate. The theory of compressed sensing is built under a sparsity assumption that an n -dimensional signal \mathbf{x} can be sparsely represented using a few non-zero coefficients in a basis. This sparse signal is sampled to yield an m -dimensional measurement vector $\mathbf{b} = \mathbf{F}\mathbf{x} + \mathbf{n}$, where \mathbf{F} is an $m \times n$ sensing matrix and \mathbf{n} is an $m \times n$ noise vector. Since $m < n$, the problem of estimating \mathbf{x} is ill-posed. However, the theory shows that \mathbf{x} is reliably estimated by solving the l_0 -norm problem:

$$\min_{\mathbf{x}} \tau \|\mathbf{x}\|_0 + 2^{-1} \|\mathbf{b} - \mathbf{F}\mathbf{x}\|_2^2, \quad (1)$$

where τ is a positive regularization value. In (1), the l_0 -norm function is non-convex and discontinuous. Indeed, (1) is known to be NP-hard. Instead of solving (1), researchers aim to solve an l_1 -norm problem. This problem is formulated by relaxing the l_0 -norm function in (1) and is given by

$$\min_{\mathbf{x}} \tau \|\mathbf{x}\|_1 + 2^{-1} \|\mathbf{b} - \mathbf{F}\mathbf{x}\|_2^2. \quad (2)$$

Candes and Tao [1] have proved that a solution to (2) is equivalent to a solution to (1) if \mathbf{F} satisfies a *restricted isometry constant* (RIC) condition. Many l_1 -norm-based methods have

been proposed to solve (2). The earliest method is l_1l_s [2]. This method is based on an interior point technique and can estimate \mathbf{x} from a small number of iterations. In each of iteration, l_1l_s solves a linear equation system expressed in a matrix-vector product form. The matrix in each system changes as the iteration passes. Thus, factorization methods such as the LU decomposition and the QR decomposition can be used to reduce the computations for solving this system. However, solving multiple linear equation systems can be still burdensome. This makes its computational cost too high for high-dimensional \mathbf{x} . Then, *gradient projection sparse recovery* [3], homotopy [4], split-Bregman [5] and *your algorithms for l_1* (YALL1) [6] have been proposed to solve (2). These are first-order-type methods that do not require matrix-inversions in all iterations. This implies that they are computationally tractable to estimate high-dimensional \mathbf{x} . But, there are known problems on (2). First, the l_1 -norm function yields a biased estimation for large non-zero magnitudes, while the l_0 -norm function considers all non-zero magnitudes equally [7]. Second, if \mathbf{F} does not satisfy the RIC condition, – either m is small or the elements of \mathbf{F} are correlated – then a solution to (2) is sub-optimal [8].

In the literature, l_0 -norm-based methods such as *iterative hard thresholding* (IHT) [9], variants of IHT [10]–[12],

and *mean doubly augmented Lagrangian* (MDAL) [13] have been proposed to solve (1). Dong and Zhu [12] have shown that their method is superior to homotopy [4]. Dong and Zhang [13] have shown that MDAL restores images with higher quality than those recovered by split-Bregman [5]. These results in [12] and [13] suggest that more accurate sparse signal estimation is conducted using the l_0 -norm function rather than the l_1 -norm function.

Recently, Bourguignon *et al.* [14] have proposed an novel approach to solve (1). This approach aims to find an estimate for \mathbf{x} and the positions of the non-zero elements of \mathbf{x} , i.e., the support set. From (1), they have made a mixed integer quadratic programming (MIQP) problem:

$$\begin{aligned} \min_{\mathbf{u} \in \{0,1\}^n, \mathbf{x} \in \mathbb{R}^n} \quad & \tau \mathbf{1}_n^T \mathbf{u} + 2^{-1} \|\mathbf{b} - \mathbf{F}\mathbf{x}\|_2^2 \\ \text{subject to} \quad & |\mathbf{x}| \leq M\mathbf{u} \end{aligned} \quad (3)$$

where the binary vector \mathbf{u} indicates the support set and M is a positive value. For (3), M can be known in practical contexts. For example, if \mathbf{x} is an 8-bit greyscale image, M is set to be 255. Bertsimas *et al.* [15] have proposed methods to estimate upper bounds on M if both \mathbf{F} and \mathbf{b} are known. Bourguignon *et al.* [14] used CPLEX [16] to solve (3) and demonstrated that CPLEX is superior to IHT [9] for sparse signal estimation. According to explanations in [14], this result is because CPLEX exhaustively searches for a whole feasible space to find the global solution to (3) while IHT finds a local solution to (1).

CPLEX [16] is a commercial solver which can be used to solve MIQP problems. Then, CPLEX is implemented based on a branch-and-cut method [30] that is a combination of a cutting plane method [31] with a branch-and-bound method [17]. As noted in [22], the branch-and-cut method has non-polynomial computational costs in the worst-case and can be troublesome to solve MIQP problems with large variables. This implies the computational intractability of using CPLEX in solving integer programming problems with large variables. In Section V, we empirically confirm this computational intractability.

In this paper, we aim to propose a fast method based on the *alternating direction method* (ADM) for solving (3). We analyze the computational cost per iteration of the proposed method, referred to as ADM-MIQP. According to this result, we can show that ADM-MIQP is a first-order-type method. We evaluate the quality of its solution using metrics defined as follows.

First, we define *support set error* (SSE) as

$$d_1(\mathbf{u}, \hat{\mathbf{u}}) := k^{-1} \sum_{i=1}^n \|\mathbf{u}(i) - \hat{\mathbf{u}}(i)\|_1, \quad (4)$$

where $\hat{\mathbf{u}}$ is a solution to (3) and \mathbf{u} is constructed from

$$\begin{aligned} \mathbf{u}(i) &= 0 & \text{if } i \notin \mathcal{I} \\ \mathbf{u}(i) &= 1 & \text{if } i \in \mathcal{I}, \end{aligned}$$

where \mathcal{I} is the support set to be detected. Second, we define *mean square error* (MSE) as

$$d_2(\mathbf{x}, \tilde{\mathbf{x}}) := n^{-1} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2, \quad (5)$$

where \mathbf{x} is an original signal and $\tilde{\mathbf{x}}$ is an estimate of \mathbf{x} . Then, we compare ADM-MIQP with both YALL1 and MDAL in terms of SSE and MSE. We observe the following:

- ADM-MIQP significantly surpasses both MDAL and YALL1 in terms of both MSE and SSE.
- ADM-MIQP exhibits good estimation performance close to the performance of ORACLE that knows support set *a priori*.
- ADM-MIQP is computationally tractable for solving (3) with the problem dimension up to the order of one million.
- ADM-MIQP exhibits a computational cost given by $O(n^{1.3})$ in our simulations.

The rest of this paper is organized as follows. Section II gives notations used in this paper and a summary about ADM. Section III elucidates the derivation and computational costs associated with ADM-MIQP. Also, Section III gives results of comparison between our proposed approach with that of [22] for solving our problem. Section IV gives simulation studies and shows the superiority of ADM-MIQP compared to other ADM-based methods [6], [13]. Section V gives conclusions of this paper.

II. PRELIMINARIES

A. NOTATIONS

We present some notations frequently used in this paper and their meanings in Table 1.

TABLE 1. Summary of the notations.

Notation	Definitions
$\mathbf{1}_n$	The $n \times 1$ vector of ones
$\mathbf{0}_n$	The $n \times 1$ vector of zeros
\mathbf{I}_n	The $n \times n$ identity matrix
\mathbf{O}_n	The $n \times n$ matrix of zeros
$\mathbf{f}(i)$	The i^{th} element of a column vector \mathbf{f}
$\mathbf{f}[i:l]$	The column vector constructed by collecting elements of a column vector \mathbf{f} from the i^{th} element to l^{th} element
$\mathbf{F}_{\mathcal{I}}$	The matrix is constructed by collecting columns of \mathbf{F} corresponding to indices of a set \mathcal{I}

B. ALTERNATION DIRECTION METHOD (ADM)

A branch and bound method [17] finds the global solution to a MIQP problem. But, since this method has non-polynomial computational costs, it is computationally intractable for solving MIQP problems with large variables. We turn instead to ADM for solving the MIQP problem (3). In this subsection, we introduce ADM and provide its recent results.

It is well-known that ADM is a powerful technique for solving a large-scale convex problem. ADM involves the following steps: *i*) ADM splits this problem into sub-problems and *ii*) solves alternatively these sub-problems until conditions are satisfied. ADM is then proven to find the global solution to this problem as the iteration continues [18], [19]. As the number of iterations approaches infinity, the solution

generated by ADM converges to an optimal solution which satisfies the *Karush-Kuhn-Tucker* conditions to a convex problem.

Recently, ADM has been empirically shown to be a powerful technique to find accurate solutions to integer programming problems [20]–[22]. Yadav *et al.* [20] have used ADM to solve an image separation problem that can be modeled as a binary quadratic programming. Souto and Dinis [21] have then solved a signal decoding problem modeled as an integer quadratic programming with an equality constraint using ADM. Last, Takapoui *et al.* [22] have solved problems modeled as MIQPs with an equality constraint, and shown that ADM could be greatly faster than a commercial integer programming method. We are motivated to derive a computationally tractable and accurate method to solve (3) using ADM, inspired by these results in [20]–[22].

III. SPARSE SIGNAL ESTIMATION VIA MIQP PROBLEM

The MIQP problem (3) has an inequality constraint and this constraint can be formulated into an equality constraint. Thus, we can use the approach of [22] to solve (3) by taking a further formulation. But, no explicit discussion on how this approach can be used to solve a MIQP problem which has an inequality constraint was given in [22]. In the sub-section III.C, we derive an algorithm based on the approach of [22] for solving (3). We call it *Takapoui's Algorithm with Inequality Constraint* (TAIC). We then compare ADM-MIQP with TAIC with respect to the computational cost per iteration. We show that ADM-MIQP requires much less computation per iteration than TAIC does.

A. DERIVATION OF ADM-MIQP

It is convenient to solve a single minimization problem rather than a joint minimization problem. To this end, we define

$$\mathbf{d} = [\mathbf{x}^T \quad \mathbf{u}^T]^T \in \mathbb{R}^n \times \{0, 1\}^n,$$

which is nonconvex. Then, (3) is reformulated into

$$\min_{\mathbf{d}} 2^{-1} \mathbf{d}^T \mathbf{Q} \mathbf{d} + \mathbf{q}^T \mathbf{d} \quad \text{subject to } \mathbf{A} \mathbf{d} \leq \mathbf{0}_{2n},$$

where

$$\mathbf{Q} = \begin{bmatrix} \mathbf{F}^T \mathbf{F} & \mathbf{O}_n \\ \mathbf{O}_n & \mathbf{O}_n \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} -\Phi^T \mathbf{b} \\ \tau \mathbf{1}_n \end{bmatrix},$$

and

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_n & -M \mathbf{I}_n \\ -\mathbf{I}_n & -M \mathbf{I}_n \end{bmatrix}.$$

We then define a nonnegative vector \mathbf{z} . Then, we obtain

$$\begin{aligned} & \min_{\mathbf{d}, \mathbf{z}} 2^{-1} \mathbf{d}^T \mathbf{Q} \mathbf{d} + \mathbf{q}^T \mathbf{d} + I_{\mathcal{X}}(\mathbf{z}) \\ & \text{subject to } \mathbf{A} \mathbf{d} + \mathbf{z} = \mathbf{0}_{2n} \end{aligned} \quad (6)$$

where $I_{\mathcal{X}}(\mathbf{z})$ is an indicator function of $\mathcal{X} := \{\mathbf{z} | \mathbf{z} \geq \mathbf{0}_{2n}\}$, i.e., $I_{\mathcal{X}}(\mathbf{z}) = 0$ for $\mathbf{z} \in \mathcal{X}$ and $I_{\mathcal{X}}(\mathbf{z}) = \infty$ for $\mathbf{z} \notin \mathcal{X}$.

We apply ADM into (6) to obtain

$$\begin{aligned} \mathbf{d}_{t+1} &= \arg \min_{\mathbf{d}} 2^{-1} \mathbf{d}^T [\mathbf{Q} + \rho \mathbf{A}^T \mathbf{A}] \mathbf{d} + \mathbf{q}_1^T \mathbf{d}, \\ \mathbf{z}_{t+1} &= \arg \min_{\mathbf{z}} I_{\mathcal{X}}(\mathbf{z}) + \rho 2^{-1} \mathbf{z}^T \mathbf{z} + (\rho \mathbf{A} \mathbf{d}_{t+1} - \lambda_t)^T \mathbf{z}, \\ \lambda_{t+1} &= \lambda_t - \rho (\mathbf{A} \mathbf{d}_{t+1} + \mathbf{z}_{t+1}), \end{aligned} \quad (7)$$

where $\mathbf{q}_{1,t} = \mathbf{q} - \mathbf{A}^T (\lambda_t - \rho \mathbf{z}_t)$, λ is the dual variable, and $\rho > 0$ is a penalty value. The sub-problem on \mathbf{d} is an MIQP. Thus, solving this problem is difficult, but we separate it into a pair of problems in terms of \mathbf{x} and \mathbf{u} , respectively:

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg \min_{\mathbf{x}} 2^{-1} \mathbf{x}^T (\mathbf{F}^T \mathbf{F} + 2\rho \mathbf{I}_n) \mathbf{x} + \mathbf{q}_{1,t}^T [1:n] \mathbf{x}, \\ \mathbf{u}_{t+1} &= \arg \min_{\mathbf{u}} \rho M^2 \mathbf{u}^T \mathbf{u} + \mathbf{q}_{1,t}^T [n+1:2n] \mathbf{u}. \end{aligned}$$

Since the sub-problem on \mathbf{x} has a quadratic objective function, we have an analytic closed-form solution:

$$\begin{aligned} \mathbf{x}_{t+1} &= -(\mathbf{F}^T \mathbf{F} + 2\rho \mathbf{I}_n)^{-1} \mathbf{q}_{1,t} [1:n] \\ &= (\mathbf{F}^T \mathbf{D} \mathbf{F} \mathbf{q}_{1,t} [1:n] - \mathbf{q}_{1,t} [1:n]) / (2\rho) \end{aligned} \quad (8)$$

where the second equality is due to the Woodbury formula [23] and $\mathbf{D} := (\mathbf{F} \mathbf{F}^T + 2\rho \mathbf{I}_m)^{-1}$. The sub-problem on \mathbf{u} is a binary quadratic programming. Since $\mathbf{u}^T \mathbf{u} = \mathbf{1}_n^T \mathbf{u}$, we have

$$\mathbf{u}_{t+1} = \arg \min_{\mathbf{u}} (\rho M^2 \mathbf{1}_n + \mathbf{q}_{1,t} [n+1:2n])^T \mathbf{u},$$

which has an analytic closed-form solution as follows:

$$\begin{aligned} \mathbf{u}_{t+1}(i) &= 0 \quad \text{if } \eta_t \geq 0 \\ \mathbf{u}_{t+1}(i) &= 1 \quad \text{if } \eta_t < 0 \end{aligned} \quad (9)$$

where $\eta_t = \rho M^2 + \mathbf{q}_{1,t}(n+i)$. The sub-problem on \mathbf{z} is solved to yield a solution:

$$\mathbf{z}_{t+1} = \max \left(\mathbf{0}_{2n}, \lambda_t / \rho + \begin{bmatrix} M \mathbf{u}_{t+1} - \mathbf{x}_{t+1} \\ M \mathbf{u}_{t+1} + \mathbf{x}_{t+1} \end{bmatrix} \right), \quad (10)$$

where ‘‘max’’ operation is performed element-wise.

In summary, we have formulated (6) from (3) by adding the non-negative vector and the indicator function. We then have applied ADM into (6) to produce the iterations given in (7). We next provided analytic solutions to these sub-problems. Then, we summarized ADM-MIQP in Table 2.

In [18] and [19], it has been proved that for any positive penalty value ρ , ADM can find the global solution to a convex problem. The penalty value only affects the convergence speed, not the quality of the solution. Researchers have discussed how this penalty value can be chosen to improve the speed [18], [19]. However, our problem (3) is non-convex due to the non-convex variable \mathbf{d} . In the literature, there are no convergence studies for non-convex problems with non-convex variables, to the best of our knowledge. It is difficult to find convergence conditions for the penalty value in the problem (3) that is being solved using ADM-MIQP. Afonso *et al.* [32] have solved (1) using their own algorithm

TABLE 2. The pseudo code of ADM-MIQP.

Parameters: $\mathbf{x}_0 = \mathbf{0}_n, \mathbf{u}_0 = \mathbf{0}_n, \mathbf{z}_0 = \mathbf{0}_n, \lambda_0 = \mathbf{0}_{2n}$ $\mathbf{F}, \mathbf{b}, \mathbf{A}, \rho, \tau, M, \varepsilon, \text{MaxIter}$	
Step 1:	set $\mathbf{D} := (\mathbf{F}\mathbf{F}^T + 2\rho\mathbf{I}_m)^{-1}$
Step 2:	for $t = 0, 1, 2, \dots, \text{maxIter}$
Step 3:	update \mathbf{x}_{t+1} by (8).
Step 4:	update \mathbf{u}_{t+1} by (9).
Step 5:	update \mathbf{z}_{t+1} by (10).
Step 6:	update $\mathbf{d}_{t+1} = [\mathbf{x}_{t+1}^T \ \mathbf{u}_{t+1}^T]^T$.
Step 7:	update $\lambda_{t+1} = \lambda_t - \rho(\mathbf{A}\mathbf{d}_{t+1} + \mathbf{z}_{t+1})$.
Step 8:	If $\frac{\ \mathbf{d}_{t+1} - \mathbf{d}_t\ _2}{\ \mathbf{d}_{t+1}\ _2} \leq \varepsilon$, then go to Step 10
Step 9:	end for
Step 10:	set $\tilde{\mathcal{I}} = \{i \mathbf{u}_i(i) = 1\}$.
Step 11:	set $\mathbf{x}_{sol} = \begin{cases} \mathbf{x}(i) = 0 & \text{if } i \notin \tilde{\mathcal{I}} \\ \mathbf{x}_{\tilde{\mathcal{I}}} = \mathbf{F}_{\tilde{\mathcal{I}}}^\dagger \mathbf{b} & \text{o.w.} \end{cases}$.

derived based on ADM. They have set their penalty value as $\rho = \tau/10$ in their simulations. Ghadimi *et al.* [18] have made a tool for setting the penalty value for a strictly convex problem with an inequality constraint. This tool takes a matrix given in the constraint as its input. By inspired by these works, we have relied upon extensive simulations with various penalty values given by a combination of τ and M , i.e.

$$\rho \in \left\{ \tau/M, \tau/M^2, \dots, \tau M^2 \right\},$$

where M is the element of our matrix \mathbf{A} . Based on results of these simulations, we set the penalty value as $\rho = \tau/M$ and use this value in our simulations. In our simulations, we empirically observe that ADM-MIQP with this penalty value can be used to solve (3) for estimating a sparse signal with the accuracy of $\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2}{\|\mathbf{x}\|_2^2} \leq \varepsilon$ where \mathbf{x} is an original sparse signal, $\tilde{\mathbf{x}}$ is the estimate sparse signal and ε is sufficiently small.

Any warm-start techniques can be applied into ADM-MIQP for improving its performance. We run ADM-MIQP multiple times with different initial variables randomly generated. Then, we have different solutions, i.e.,

$$\{\mathbf{d}^1, \mathbf{d}^2, \dots, \mathbf{d}^L\}$$

where L is the number of runs of ADM-MIQP. We then select a solution among these multiple solutions via

$$\mathbf{d}_{sol} := \underset{\mathbf{d} \in \{\mathbf{d}^1, \mathbf{d}^2, \dots, \mathbf{d}^L\}}{\text{argmin}} \quad 2^{-1} \mathbf{d}^T \mathbf{Q} \mathbf{d} + \mathbf{q}^T \mathbf{d}.$$

This selected solution is at least guaranteed to be better than the other unselected solutions in terms of the cost function.

B. COMPUTATION COSTS PER ITERATION

We aim to show that ADM-MIQP is a first-order-type method. The costs of updating \mathbf{z} and \mathbf{u} are both $O(n)$. Then, the cost of updating \mathbf{x} is $O(mn + m^3)$, due to both the matrix inversion and the matrix-vector products. If \mathbf{D} is stored, then this cost can be reduced to $O(mn)$.

Next, in applications such as a single pixel camera [24], [25], a lensless camera [26], [27], for an image compression [28], a sensing matrix is constructed by randomly taking m rows from an orthogonal matrix. Then, \mathbf{D} becomes a constant value $\frac{1}{1+2\rho}$. As a result, the update on \mathbf{x} is given as

$$\mathbf{x}_{t+1} = \left((1 + 2\rho)^{-1} \mathbf{F}^T \mathbf{F} \mathbf{q}_{1,t} [1 : n] - \mathbf{q}_{1,t} [1 : n] \right) / (2\rho). \tag{11}$$

Indeed, if \mathbf{F} is a partial discrete cosine transform (DCT) matrix, all matrix-vector products in (11) can be performed by the fast Fourier transform operation. That is, the update cost for \mathbf{x} can be significantly reduced to $O(n \log n)$.

C. COMPUTATION COSTS PER ITERATION

We now derive the algorithm called TAIC (*Takapoui's Algorithm with Inequality Constraint*) by following the approach of [22] for solving the MIQP problem (3) which only has the inequality constraint. As shown in the subsection III.A, it is noted that (3) is equal to (7). Then, we define the symbols as follows:

$$\begin{aligned} \tilde{\mathbf{d}} &:= \begin{bmatrix} \mathbf{d} \\ \mathbf{z} \end{bmatrix} \in \mathbb{R}^n \times \{0, 1\}^n \times \mathbb{R}_{+}^{2n}, \\ \tilde{\mathbf{A}} &:= \begin{bmatrix} \mathbf{A} & \mathbf{I}_{2n} \end{bmatrix} \in \mathbb{R}^{2n \times 4n}, \\ \tilde{\mathbf{q}} &:= \begin{bmatrix} \mathbf{q} \\ \mathbf{0}_{2n} \end{bmatrix} \in \mathbb{R}^{4n \times 1} \text{ and } \tilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{Q} & \mathbf{O}_{2n} \\ \mathbf{O}_{2n} & \mathbf{O}_{2n} \end{bmatrix} \in \mathbb{R}^{4n \times 4n}, \end{aligned}$$

where \mathbf{z} is a slack variable. With these symbols, we can reformulate (7) into an MIQP problem with an equality

$$\min_{\tilde{\mathbf{d}}} 2^{-1} \tilde{\mathbf{d}}^T \tilde{\mathbf{Q}} \tilde{\mathbf{d}} + \tilde{\mathbf{q}}^T \tilde{\mathbf{d}} \quad \text{subject to } \tilde{\mathbf{A}} \tilde{\mathbf{d}} = \mathbf{0}_{2n}, \tilde{\mathbf{d}} \in \tilde{\mathcal{X}} \tag{12}$$

where $\tilde{\mathcal{X}} := \mathbb{R}^n \times \{0, 1\}^n \times \mathbb{R}_{+}^{2n}$ is a non-convex set. Similar to (6), we also reformulate (12) into a standard form of ADM as follows:

$$\begin{aligned} \min_{\tilde{\mathbf{d}}, \tilde{\mathbf{z}}} 2^{-1} \tilde{\mathbf{d}}^T \tilde{\mathbf{Q}} \tilde{\mathbf{d}} + \tilde{\mathbf{q}}^T \tilde{\mathbf{d}} + I_{\tilde{\mathcal{X}}}(\tilde{\mathbf{z}}) \\ \text{subject to } \begin{bmatrix} \tilde{\mathbf{A}} \\ \mathbf{I}_{4n} \end{bmatrix} \tilde{\mathbf{d}} - \begin{bmatrix} \mathbf{O}_{2n \times 4n} \\ \mathbf{I}_{4n} \end{bmatrix} \tilde{\mathbf{z}} = \mathbf{0}_{6n} \end{aligned} \tag{13}$$

where $I_{\tilde{\mathcal{X}}}(\tilde{\mathbf{z}})$ is an indicator function of $\tilde{\mathcal{X}}$ and $\mathbf{O}_{2n \times 4n}$ is the $2n \times 4n$ matrix of zeros. TAIC is then implemented via

$$\begin{aligned} \tilde{\mathbf{d}}_{t+1} &= \underset{\tilde{\mathbf{d}}}{\text{argmin}} \quad 2^{-1} \tilde{\mathbf{d}}^T \tilde{\mathbf{Q}} \tilde{\mathbf{d}} + \tilde{\mathbf{q}}^T \tilde{\mathbf{d}} \\ &\quad + 2^{-1} \rho \left\| g \left(\tilde{\mathbf{d}}, \tilde{\mathbf{z}}_t, \tilde{\lambda}_t \right) \right\|_2^2, \\ \tilde{\mathbf{z}}_{t+1} &= \underset{\tilde{\mathbf{z}}}{\text{argmin}} \quad I_{\tilde{\mathcal{X}}}(\tilde{\mathbf{z}}) + 2^{-1} \rho \left\| g \left(\tilde{\mathbf{d}}_{t+1}, \tilde{\mathbf{z}}, \tilde{\lambda}_t \right) \right\|_2^2, \\ \tilde{\lambda}_{t+1} &= \tilde{\lambda}_t - \rho g \left(\tilde{\mathbf{d}}_{t+1}, \tilde{\mathbf{z}}_{t+1}, \mathbf{0}_{6n} \right), \end{aligned} \tag{14}$$

where $\tilde{\lambda}$ is the dual variable, $\rho > 0$ is a penalty value, and

$$g(\tilde{\mathbf{d}}, \tilde{\mathbf{z}}, \tilde{\lambda}) := \begin{bmatrix} \tilde{\mathbf{A}} \\ \mathbf{I}_{4n} \end{bmatrix} \tilde{\mathbf{d}} - \begin{bmatrix} \mathbf{O}_{2n \times 4n} \\ \mathbf{I}_{4n} \end{bmatrix} \tilde{\mathbf{z}} + \frac{\tilde{\lambda}}{\rho}.$$

It is noted that (7) is formed by adding *one slack variable* to (3). But, (13) is formed by adding *two slack variables* into (3). Thus, there is an intuition that TAIC requires more computational costs per iteration than ADM-MIQP does.

To investigate the validation of our intuition, we restrict our attentions to the sub-problem on $\tilde{\mathbf{d}}$ in (14) that can be simplified to

$$\tilde{\mathbf{d}}_{t+1} = \arg \min_{\tilde{\mathbf{d}}} 2^{-1} \tilde{\mathbf{d}}^T \tilde{\mathbf{D}} \tilde{\mathbf{d}} + \mathbf{h}_t^T \tilde{\mathbf{d}} \quad (15)$$

where $\tilde{\mathbf{D}} := [\tilde{\mathbf{Q}} + \rho(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}} + \mathbf{I}_{4n})] \in \mathbb{R}^{4n \times 4n}$ and

$$\mathbf{h}_t := \tilde{\mathbf{q}} - \rho \begin{bmatrix} \tilde{\mathbf{A}} \\ \mathbf{I}_{4n} \end{bmatrix}^T \left(\begin{bmatrix} \mathbf{O}_{2n \times 4n} \\ \mathbf{I}_{4n} \end{bmatrix} \tilde{\mathbf{z}}_t - \tilde{\lambda}_t \right).$$

The sub-problem on $\tilde{\mathbf{d}}$ in (7) has been decomposed into a pair of problems on \mathbf{x} and \mathbf{u} , respectively. But, $\tilde{\mathbf{D}}$ is a non-diagonal matrix that implies that the sub-problem in (15) cannot be decomposed. We then consider an analytic closed form solution to (15) as follows:

$$\tilde{\mathbf{d}}_{t+1} = -\tilde{\mathbf{D}}^{-1} \mathbf{h}_t. \quad (16)$$

For saving computational costs, the inverse matrix in (16) can be stored. Even with this stored matrix, TAIC takes $O(16n^2)$ computational cost per iteration for conducting (16) due to the matrix-vector product. This cost can be negligible for a small value of n . For a large value of n , it cannot be ignored. On the other hands, ADM-MIQP takes $O(mn)$ computational cost per iteration. It can be seen that TAIC takes more computational costs per iteration for updating the other variables than ADM-MIQP does. Thus, it can be concluded that the cost of ADM-MIQP is greatly less than that of TAIC.

IV. SIMULATIONS STUIDES

We conduct simulations to show that ADM-MIQP gives a solution to (3). We compare ADM-MIQP with MDAL and YALL1. The reasons for selecting both MDAL and YALL1 as comparative approaches are *a)* these methods are also based on ADM and *b)* are known to be computationally tractable. We define a Gaussian sparse vector ensemble and a Gaussian noise vector ensemble as follows.

Definition 1: The Gaussian sparse vector ensemble is an ensemble of n -dimensional k -sparse vectors, where each vector \mathbf{x} is generated as follows: *a)* the positions of the non-zero values of \mathbf{x} are randomly selected, *b)* the non-zero values are taken from the standard normal distribution and *c)* \mathbf{x} is normalized to produce the l_2 -norm for \mathbf{x} unit.

Definition 2: The Gaussian noise vector ensemble is an ensemble of m -dimensional noise vectors whose elements are independent and identically distributed Gaussian with zero mean and variance σ^2 .

We define the *signal-to-noise ratio* (SNR) as

$$\text{SNR [dB]} := 10 \log_{10} \left(\frac{\|\mathbf{F}\mathbf{x}\|_2^2}{(m\sigma^2)} \right).$$

We set the parameters of ADM-MIQP, MDAL, and YALL1 as follows. The regularization value is set as $\tau = \sigma\sqrt{2\log n}$ if SNR [dB] is finite and $\tau = 10^{-4}$ if SNR [dB] is infinite. The M value is set as $M = \max_i |\mathbf{x}(i)|$. As we have stated in the sub-section III.A, our penalty value is set as $\rho = \tau/M$. The penalty value of YALL1 is set as $\rho = \|\mathbf{b}\|_1/m$, used in [6]. But, MDAL with the penalty value used in [13] failed to yield an accurate solution in our simulation. We conducted extensive simulations to find the penalty value for MDAL. Thus, in our simulations, it was set as $\rho = 10\tau$. We terminated these methods either when the number of iterations exceeded 2000 or when $\frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2}{\|\mathbf{x}_{t+1}\|_2} \leq 10^{-4}$, as was done in [6], for YALL1, and when $\frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2}{\|\mathbf{b}\|_2} \leq 10^{-4}$, as in [13] for MDAL and when $\frac{\|\mathbf{d}_{t+1} - \mathbf{d}_t\|_2}{\|\mathbf{d}_{t+1}\|_2} \leq 10^{-4}$ for ADM-MIQP.

We kept in mind that a solution for \mathbf{x} in (3) must satisfy a convex constraint

$$\mathbf{x} \in \{\mathbf{x} \mid -M \leq \mathbf{x}(i) \leq M\},$$

where $i = 1, 2, \dots, n$. However, both YALL1 and MDAL are not designed to use this constraint. Therefore, we extended these methods to use the constraint for a fair comparison. Since the constraint is convex, this extension was easily carried out by adding the following:

$$\mathbf{x}_t(i) = \min(\max(\mathbf{x}_t(i), M), -M),$$

where $\mathbf{x}_t(i)$ is the i^{th} element of an intermediate solution at the t^{th} iteration. All simulations are conducted on a computer with Intel (R) Core (TM) i7-3820 processor clocked at 3.6 GHz. The MATLAB codes are in [29].

A. CONVERGENCE BEHAVIORS OF ADM-MIQP

We remind that both SSE defined in (4) and MSE defined in (5) can be used to evaluate the quality of a solution given by ADM-MIQP. We use both of the metrics to study how this solution behaves. Since the elements of \mathbf{u} are either 0 or 1, we have

$$\begin{aligned} kd_1(\mathbf{u}, \mathbf{u}_t) + nd_2(\mathbf{x}, \mathbf{x}_t) &= \|\mathbf{u} - \mathbf{u}_t\|_2^2 + \|\mathbf{x} - \mathbf{x}_t\|_2^2 \\ &= \left\| \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} - \begin{bmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{bmatrix} \right\|_2^2 \\ &= \|\mathbf{d} - \mathbf{d}_t\|_2^2 \end{aligned} \quad (17)$$

where \mathbf{d}_t is the t^{th} solution of the ADM-MIQP and \mathbf{d} is a feasible solution to (3). Thus, if both the metrics are small, the l_2 -norm between the t^{th} solution and the \mathbf{d} is also small. Based on this relation, we define the convergence of ADM-MIQP.

Definition 3: A solution $\mathbf{d}_t = [\mathbf{x}_t^T \mathbf{u}_t^T]^T$ of by ADM-MIQP is convergent to a point $\mathbf{d} = [\mathbf{x}^T \mathbf{u}^T]^T$ to (3) if there exists a positive integer T such that for every positive

ε_1 and ε_2 , we then have $d_1(\mathbf{u}, \mathbf{u}_t) < \varepsilon_1$ and $d_2(\mathbf{x}, \mathbf{x}_t) < \varepsilon_2$ for all $T \leq t \leq \text{maxIter}$, where maxIter is the maximum number of iteration.

To show that ADM-MIQP can find a converged solution to the MIQP problem (3), the problem dimension n , the number of measurements m and the sparsity level k were set as 1024, 307 and 30, respectively. Two values for SNR [dB] were considered: 35 and 45, respectively. We generated 1000 independent realizations of the set $(\mathbf{F}, \mathbf{x}, \mathbf{n})$ where \mathbf{F} was made by randomly taking 307 rows of the 1024×1024 DCT matrix, \mathbf{x} was taken from the Gaussian sparse vector ensemble, and \mathbf{n} was taken from the Gaussian noise vector ensemble. We determined average values for both MSE defined in (5) and SSE defined in (4). We then plotted the results in Figs. 1 and 2, respectively.

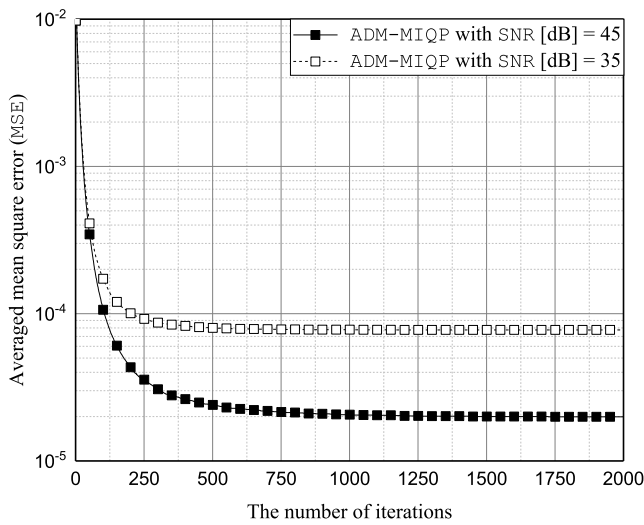


FIGURE 1. It plots the average MSE of ADM-MIQP depending on the number of iterations. The problem dimension n , the number of measurements m and the sparsity level k are set to be 1024, 307 and 30, respectively.

For all the SNRs investigated, both MSE and SSE gradually decreased and were eventually saturated. For SNR [dB] = 45, at the 250th and 500th iterations, MSEs were 3.5×10^{-5} and 2.4×10^{-5} , respectively. Finally, MSE converged to 2×10^{-5} after $O(10^3)$ iterations. This means that an estimate of \mathbf{x} can converge to an original sparse signal. Next, we considered SSE at SNR [dB] = 45. At the 250th and 500th iterations, SSEs were 0.041 and 0.031, respectively. Eventually, SSE converged to 0.029 after $O(10^3)$ iterations. This suggests that the detected support set converges to an original support set. Due to (17), after $O(10^3)$ iterations, we observed

$$\|\mathbf{d} - \mathbf{d}_t\|_2^2 < O(10^{-c})$$

where $c \approx 1$. This observation shows the convergence of ADM-MIQP under the definition 3.

B. COMPARISON STUDIES AND DISCUSSION

Let $\alpha := m/n$ be an *under-sampling* ratio and $\beta := k/m$ be an *over-sampling* ratio. The phase transition for a given method

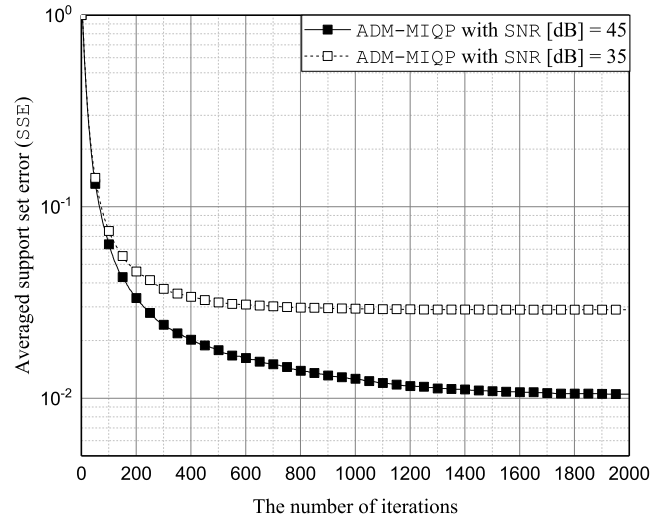


FIGURE 2. It plots the average SSE of ADM-MIQP depending on the number of iterations. The problem dimension n , the number of measurements m and the sparsity level k are set to be 1024, 307 and 30, respectively.

shows how accurately this method can estimate sparse signals in the (α, β) plane with n . We conducted simulations to study the phase transitions in computations obtained by ADM-MIQP, MDAL and YALL1. Then, the aims of this phase transition study include being aware of the overall performance of ADM-MIQP and understanding which of these ADM-based methods, each of which solves different problems to estimate sparse signals, achieves the best performance for this sparse signal estimation.

The problem dimension n was set as 1024. Then, a 15×15 uniformly spaced grid on the (α, β) plane was constructed for $\alpha, \beta \in \{0.15, 0.175, \dots, 0.5\}$. We generated 1000 independent realizations of the set (\mathbf{F}, \mathbf{x}) , where \mathbf{F} was derived by randomly taking m rows of the 1024×1024 DCT matrix and \mathbf{x} was taken from the Gaussian sparse vector ensemble. The estimate $\tilde{\mathbf{x}}$ was considered to be successful if $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 / \|\mathbf{x}\|_2^2 \leq 10^{-4}$. In Fig. 3, we illustrated the phase transitions for all these methods. The solid line represents a 99% probability of success. That is, for points lying in the graphical area below this line, there was at least 99% probability of success in problem solving. The area beneath the dashed-line then represents a 50% probability of success.

First, we fixed the over-sampling ratio. We then considered the under-sampling ratio to attain a 99% probability of success. The under-sampling ratio for ADM-MIQP was found to be the smallest. As an example, for a fixed $\beta = 0.25$, we observed that the under-sampling ratios of ADM-MIQP, MDAL, and YALL1 were 0.25, 0.275, and 0.325, respectively. The under-sampling ratio was proportional to m because n was fixed. This implies that ADM-MIQP requires the smallest value of m for sparse signal estimation, when compared with the other methods.

Second, we fixed the under-sampling ratio and considered the over-sampling ratio to achieve a 99% probability of success. We observed that for ADM-MIQP, the over-sampling

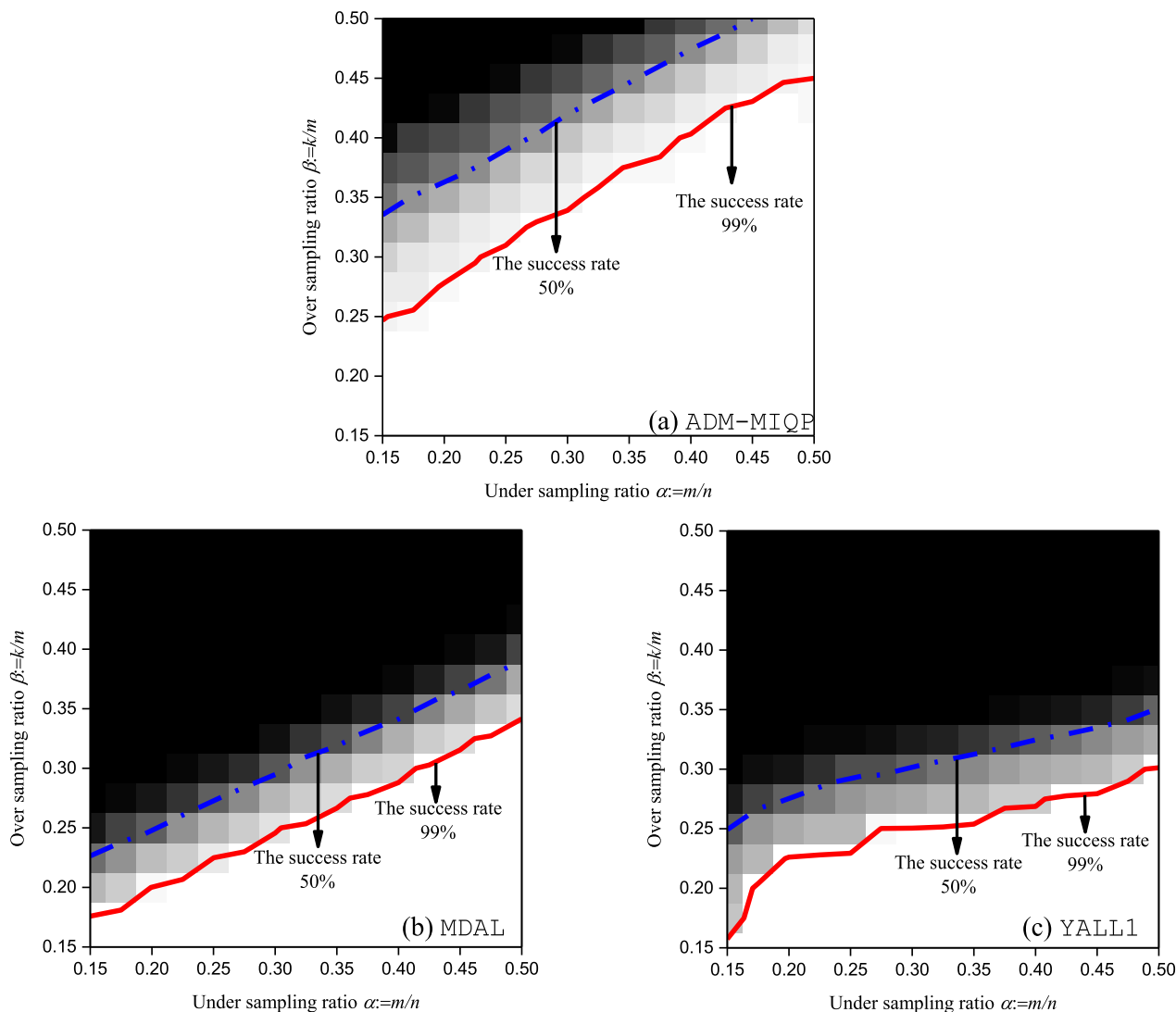


FIGURE 3. It plots the empirical phase transitions of the ADM-based methods such as ADM-MIQP, MDAL and YALL1, respectively.

ratio was the largest. For a fixed $\alpha = 0.3$, the over-sampling ratios of ADM-MIQP, MDAL and YALL1 were 0.325, 0.25, and 0.225, respectively. The over-sampling ratio was proportional to k for a fixed under-sampling ratio. This shows that ADM-MIQP can estimate \mathbf{x} with the higher value of k in which the other methods cannot.

Next, we conducted simulations to study the performance of all these methods by varying k for a fixed n and m under noisy cases. To this end, SNR [dB], n and m were set as 35, 1024, and 307, respectively and k was varied between 30 and 100. Then, we generated 1000 independent realizations of the set $(\mathbf{F}, \mathbf{x}, \mathbf{n})$ where \mathbf{F} , \mathbf{x} , and \mathbf{n} were obtained through the manner discussed in the sub-section III.A. Then, we obtained the average MSE for each method and plotted these values in Fig. 4.

For any k , ADM-MIQP can achieve the lowest MSE when compared with MDAL and YALL1. This means that ADM-MIQP can more accurately estimate \mathbf{x} than the other methods can. The MSE gap between ADM-MIQP and

ORACLE is small. At $k = 40$, as an example, MSEs of ADM-MIQP and ORACLE are 7×10^{-6} and 4×10^{-6} , respectively. This suggests that ADM-MIQP can achieve a performance close to that achieved by ORACLE.

Since both MDAL and YALL1 are originally designed to find an estimate of \mathbf{x} , not the support set, we needed to construct the support set based on the estimate $\tilde{\mathbf{x}}$ in order to measure SSEs for these methods. For this purpose, we set a threshold value

$$\zeta = 0.8 \min_i |\mathbf{x}(i)|$$

and constructed the support set $\hat{\mathbf{u}}$ by

$$\hat{\mathbf{u}}(i) = 0 \quad \text{if } |\tilde{\mathbf{x}}(i)| < \zeta,$$

$$\hat{\mathbf{u}}(i) = 1 \quad \text{if } |\tilde{\mathbf{x}}(i)| \geq \zeta.$$

where $i = 1, 2, \dots, n$. Under the same conditions used in the experiment depicted in Fig. 4, we independently made 1000 realizations of the set $(\mathbf{F}, \mathbf{x}, \mathbf{n})$. We then determined the

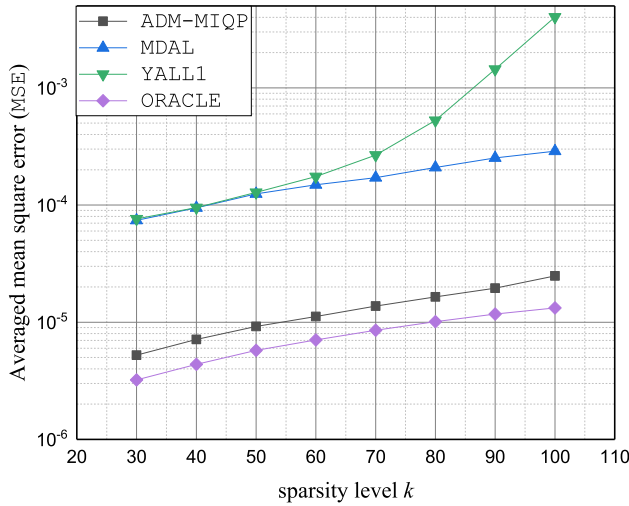


FIGURE 4. It plots the average MSEs of ADM-MIQP, MDAL, YALL1 and ORACLE depending on the sparsity level k . The problem dimension n , the number of measurements m and SNR [dB] are set to be 1024, 307 and 35, respectively.

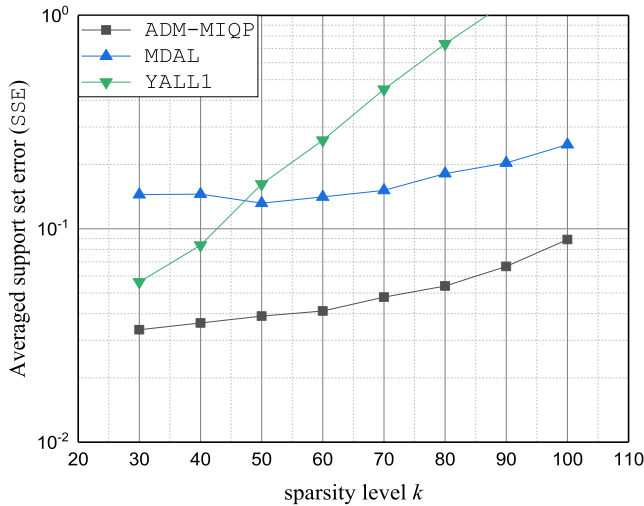


FIGURE 5. It plots the average SSEs of ADM-MIQP, MDAL and YALL1 depending on the sparsity level k . The problem dimension n , the number of measurements m and SNR [dB] are set to be 1024, 307 and 35, respectively.

average SSE for each of the methods and plotted the results in Fig. 5. As with MSE, for any k , ADM-MIQP was found to achieve the lowest SSE. As an example, at $k = 80$, SSEs of ADM-MIQP, MDAL, and YALL1 were 0.04, 0.14, and 0.26, respectively. This means that ADM-MIQP can more accurately detect the support set than the other methods can. Next, at $k = 60$, we counted the number of events for which $\sum_{i=1}^n \|\mathbf{u}(i) - \hat{\mathbf{u}}(i)\|_1 \leq 6$, i.e., for which the support set error could occur within 10%. The results for ADM-MIQP, MDAL, and YALL1 were 962, 227, and 349 events respectively. This suggests that ADM-MIQP surpasses the other methods.

Thus far, we have shown that ADM-MIQP is superior to other ADM-based methods in terms of MSE and SSE. There are multiple reasons for why this is the case.

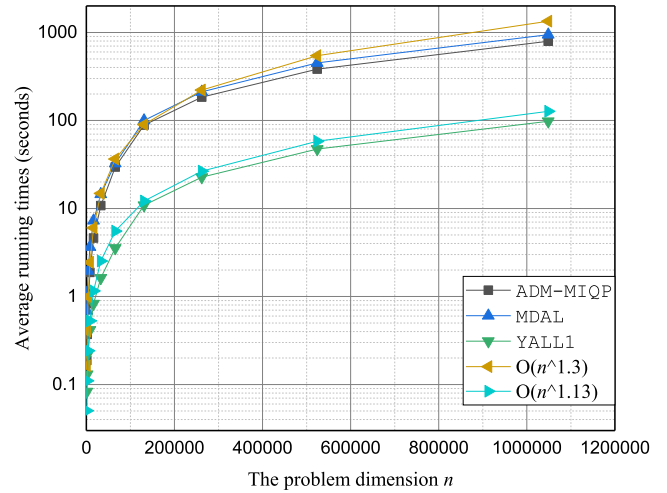


FIGURE 6. It plots the average running times of ADM-MIQP, MDAL, YALL1 and CPLEX depending on the problem dimension n with $m = \lfloor 0.3n \rfloor$, $k = \lfloor 0.3m \rfloor$ and SNR [dB] = 45. ADM-MIQP, MDAL and YALL1 have the polynomial computational order.

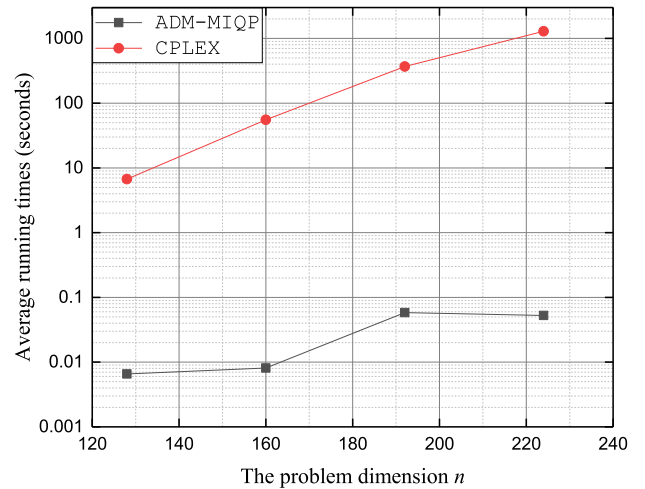


FIGURE 7. It plots the average running times of ADM-MIQP and CPLEX depending on the problem dimension n with $m = \lfloor 0.3n \rfloor$, $k = \lfloor 0.2m \rfloor$ and SNR [dB] = 45. This figure shows that ADM-MIQP is significantly faster than CPLEX.

First, ADM-MIQP is designed to solve (3). The binary vector \mathbf{u} in (3) indicates the support set and $\mathbf{1}_n^T \mathbf{u}$ counts the number of ones in \mathbf{u} . This means that ADM-MIQP aims to find a solution that both the cardinality of the support set and the data-fidelity are jointly minimized. Minimizing the cardinality of the support set is a characteristic of l_0 -norm based methods. This is the reason for the superiority of our method over YALL1.

Second, Dong and Zhang [13] have empirically reported that MDAL finds a local solution to the l_0 -norm problem. By contrast, methods based on ADM tend to find the global solution to a MIQP problem, as reported in [20]–[22]. Then, as reported in [14], CPLEX is capable of finding the global solution to (3). To understand whether ADM-MIQP finds the global solution or not, we compared the solution of

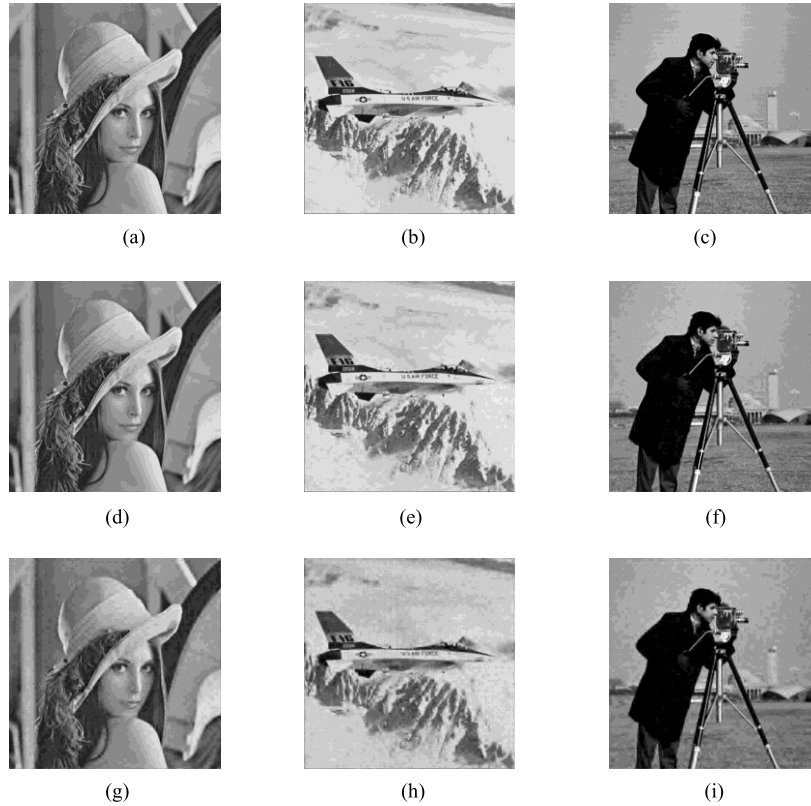


FIGURE 8. The original grayscale images of size 512×512 are shown in the first row. The images recovered by ADM-MIQP are shown in the second row. The images recovered by MDAL are shown in the third row. The PSNR value of each recovered image is averaged 10 trials at $m = \lfloor 0.15n \rfloor$ and $k = \lfloor 0.05n \rfloor$. (a) Lena. (b) Airplane. (c) Cameraman. (d) 33.83 dB. (e) 31.19 dB. (f) 33.71 dB. (g) 26.91 dB. (h) 22.27 dB. (i) 27.31 dB.

ADM-MIQP and that of CPLEX. We independently generated 100 realizations of the set (\mathbf{F}, \mathbf{x}) by assuming that n , m , and k were 200, 80, and 10 respectively, where \mathbf{F} was a partial orthogonal sensing matrix and \mathbf{x} was taken from the Gaussian sparse vector ensemble. We determined the average of the objective function

$$\tau \mathbf{1}_n^T \mathbf{u} + 2^{-1} \|\mathbf{b} - \mathbf{F}\mathbf{x}\|_2^2$$

for each method, as well as the average for *normalized* MSE,

$$\|\mathbf{x}_C - \mathbf{x}_A\|_2^2 / \|\mathbf{x}_C\|_2^2$$

where \mathbf{x}_A is an estimate of \mathbf{x} obtained by ADM-MIQP and \mathbf{x}_C is an estimate of \mathbf{x} obtained by CPLEX. The value of the objective function of CPLEX, and that of ADM-MIQP, were 0.0099 and 0.0094, respectively, and the *normalized* MSE was 0.0033. The gap between these values and the *normalized* MSE were both small. This indicates that ADM-MIQP indeed finds the global solution to (3). This makes ADM-MIQP a superior approach to MDAL.

We observed that ADM-MIQP is computationally tractable for solving (3) up to the problem dimension n of the order of one million. To this end, SNR [dB] was set as 45 and n was varied from 1024 to 1048576. For a fixed n , we altered m and k to $m = \lfloor 0.3n \rfloor$ and $k = \lfloor 0.3m \rfloor$. The number of iterations was set as 1000. At each point $(n, m, k, \text{SNR [dB]})$,

we generated 500 independent realizations of the set $(\mathbf{F}, \mathbf{x}, \mathbf{n})$, where \mathbf{F} , \mathbf{x} , and \mathbf{n} are obtained by the approach given in the sub-section IV.A. We determined the average running time for each method and plotted the results in Fig. 6.

In Fig. 6, the average running times for each method grow linearly with n . We calculated the order of the average running times for ADM-MIQP, MDAL and YALL1 with respect to n . The orders are roughly $O(n^{1.3})$, $O(n^{1.3})$, and $O(n^{1.13})$ respectively. These orders show that ADM-MIQP has polynomial computation costs, leading to that ADM-MIQP is still computationally tractable for solving (3) with the large problem dimension. Finally, YALL1 was found to be a faster method than ADM-MIQP. This is because the l_1 -norm problem (2), solved by YALL1, is easier to solve than (3). Despite this, if the running time for ADM-MIQP is acceptable, ADM-MIQP gains significant improvements on sparse signal estimation.

We conducted simulations to compare ADM-MIQP with CPLEX in terms of the running time. MaxIter was set as 1000. SNR [dB] was set as 45 and n was varied from 128 to 224. Then, both m and k were altered to $m = \lfloor 0.3n \rfloor$ and $k = \lfloor 0.2m \rfloor$. At each point $(n, m, k, \text{SNR [dB]})$, we made 50 independent realizations of the set $(\mathbf{F}, \mathbf{x}, \mathbf{n})$, where \mathbf{x} and \mathbf{n} are obtained by the approach given in the sub-section IV.A and \mathbf{F} is a partial orthogonal matrix.

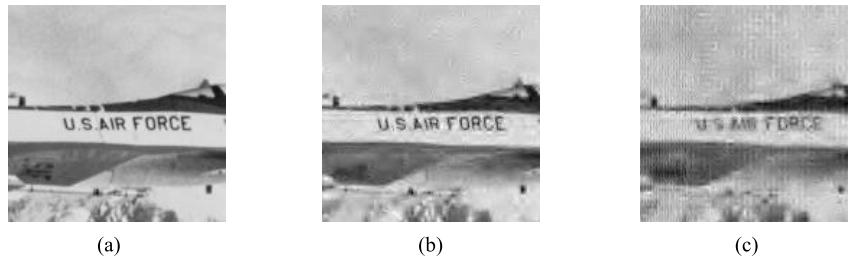


FIGURE 9. The images are corresponding to the part of the original and each recovered airplane images. (a) The original image. (b) The image recovered by ADM-MIQP. (c) The image recovered by MDAL.

In Fig. 7, the average running time of CPLEX rapidly grows with n . Even n was roughly doubled, the time rapidly increased. At $n = 128$ and $n = 224$, the times are 6.7 secs and 2113 secs, respectively. This observation can be in accordance with the statement in Section I that CPLEX has the computational intractability in solving (3) with large variables. On the other hands, the average running time of ADM-MIQP does not rapidly grow with n . This observation shows that ADM-MIQP is faster than CPLEX.

C. AN IMAGE RECOVERY EXAMPLE

We conducted an image recovery experiment to demonstrate the successful application of ADM-MIQP. For this study, the discrete wavelet transform was applied onto each image. The k largest magnitude values of the transformed image were retained. For each image, k non-zero values were stacked to form a sparse vector, to be compressed to get an m -dimensional measurement vector using a partial DCT matrix. Both MDAL and ADM-MIQP were used to recover the image. To evaluate the qualities of the recovered images, we used the following *peak-signal-to-noise ratio* (PSNR):

$$\text{PSNR [dB]} := 10 \log_{10} \left(n \times 255^2 / \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 \right), \quad (18)$$

where $\tilde{\mathbf{x}}$ is an original image and is the recovered image.

In Fig. 8, we illustrate the original greyscale images of size 512×512 with a problem dimension $n = 262144$. We have also showed the images recovered by each method and their PSNRs. These PSNR values were the averages of results from 10 trials where $m = \lfloor 0.15n \rfloor$ and $k = \lfloor 0.05n \rfloor$.

It is immediately observed that ADM-MIQP recovers images with higher quality than MDAL in terms of PSNR. ADM-MIQP then preserves the detailed information in the original images. For example, let us consider the text part “US AIR Force” of the recovered airplane image. As shown in Fig. 9, we clearly see this text in (b), recovered by ADM-MIQP, we cannot make out it in (c), recovered by MDAL. This result shows that ADM-MIQP surpasses MDAL in this image recovery example.

V. CONCLUSION

We proposed a fast method referred to as ADM-MIQP to solve the mixed integer quadratic programming problem (3)

formulated in [14] from the l_0 -norm problem (1). We derived ADM-MIQP using the alternating direction method, which has been recently used to solve integer programming problems in [20]–[22]. We then showed that ADM-MIQP is a first-order-type method. That is, matrix-vector products are only used to implement ADM-MIQP. We selected MDAL [13] and YALL1 [6] as competitors to ADM-MIQP because these methods are based on ADM to solve the l_0 -norm and the l_1 -norm problems, respectively. We also compared ADM-MIQP with ORACLE, an approach which involved *a priori* knowledge of the support set. We used both *support set error* (SSE) (4) and *mean square error* (MSE) (5) to assess the quality of a solution obtained by each method.

We empirically demonstrated that ADM-MIQP could achieve a significantly better performance than MDAL and YALL1 in terms of both SSE and MSE. We also showed that ADM-MIQP eventually achieved a performance close to that of ORACLE in terms of MSE. We showed that ADM-MIQP is computationally tractable for solving (3) up to the order of one million in the problem dimension. We confirmed that the computational cost of ADM-MIQP is $O(n^{1.3})$ in our simulations. We concluded that ADM-MIQP is efficient in finding an accurate solution to (3) when the problem dimension n is large.

The next step is to conduct convergence analysis for ADM-MIQP. Specifically, it will be interesting to prove that a solution of ADM-MIQP is convergent. Also, this work can be extended to determine the appropriate penalty value that would guarantee the convergence of ADM-MIQP.

REFERENCES

- [1] E. J. Candés, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [2] S.-J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, “An interior point method for large-scale l_1 -regularized least squares,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.
- [3] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.
- [4] L. Xiao and T. Zhang, “A proximal-gradient homotopy method for the sparse least-squares problem,” *SIAM J. Optim.*, vol. 23, no. 2, pp. 1062–1091, 2013.
- [5] T. Goldstein and S. Osher, “The split Bregman method for l_1 -regularized problems,” *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

- [6] F. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1 -problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.
- [7] F. Wen, L. Pei, Y. Yang, W. Yu, and P. Liu, "Efficient and robust recovery of sparse signal and image using generalized nonconvex regularization," *IEEE Trans. Comput. Imag.*, vol. 3, no. 3, pp. 566–579, Dec. 2017.
- [8] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 3, p. 035020, May 2008.
- [9] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, Nov. 2009.
- [10] T. Blumensath and M. E. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, Apr. 2010.
- [11] T. Blumensath, "Accelerated iterative hard thresholding," *Signal Process.*, vol. 92, no. 3, pp. 752–756, Mar. 2012.
- [12] Z. Dong and W. Zhu, "Homotopy methods based on ℓ_0 -norm for compressed sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1132–1146, Apr. 2018.
- [13] B. Dong and Y. Zhang, "An efficient algorithm for ℓ_0 minimization in wavelet frame based image restoration," *J. Sci. Comput.*, vol. 54, pp. 350–368, Feb. 2013.
- [14] S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau, "Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1405–1419, Oct. 2016.
- [15] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *Ann. Statist.*, vol. 44, no. 2, pp. 813–852, 2016.
- [16] *IBM ILOG CPLEX V12.8.0*. Accessed: Jun. 1, 2018. [Online]. Available: <http://www.ibm.com/products/ilog-cplex-optimization-studio>
- [17] S. Boyd and J. Mattingley, "Branch and bound methods," Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, Course Notes EE364b, Mar. 2018. [Online]. Available: https://web.stanford.edu/class/ee364b/lectures/bb_notes.pdf
- [18] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 644–658, Mar. 2015.
- [19] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," *J. Sci. Comput.*, vol. 66, no. 3, pp. 889–916, Mar. 2016.
- [20] A. K. Yadav, R. Ranjan, U. Mahbub, and M. C. Rotkowitz, "New methods for handling binary constraints," in *Proc. 54th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2016, pp. 1074–1080.
- [21] N. Souto and R. Dinis, "MIMO detection and equalization for single-carrier systems using the alternating direction method of multipliers," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1751–1755, Dec. 2016.
- [22] R. Takapoui, N. Moehle, S. Boyd, and A. Bemporad, "A simple effective heuristic for embedded mixed-integer quadratic programming," *Int. J. Control*, Apr. 2017, doi: 10.1080/00207179.2017.1316016.
- [23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.
- [24] M. F. Duarte *et al.*, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.
- [25] D. B. Phillips *et al.*, "Adaptive foveated single-pixel imaging with dynamic supersampling," *Sci. Adv.*, vol. 3, no. 4, p. e1601782, Apr. 2017.
- [26] G. Satat, M. Tancik, and R. Raskar, "Lensless imaging with compressive ultrafast sensing," *IEEE Trans. Comput. Imag.*, vol. 3, no. 3, pp. 398–407, Sep. 2017.
- [27] G. Huang, H. Jiang, K. Matthews, and P. Wilford, "Lensless imaging by compressive sensing," in *Proc. 20th IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, VIC, Australia, Sep. 2013, pp. 2101–2105.
- [28] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Mar. 2000, pp. 523–541.
- [29] S. J. Park and H.-N. Lee. (2018). *ADM-MIQP*. Accessed: Oct. 1, 2018. [Online]. Available: <https://github.com/infonetGIST/infonetcompressedensing>
- [30] J. E. Mitchell, "Branch-and-cut algorithms for combinatorial optimization problems," in *Handbook of Applied Optimizations*. Oxford, U.K.: Oxford Univ. Press, 2000.
- [31] B. Boyd and L. Vandenberghe, "Localization and cutting-plane method," Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, Course Notes EE364b, Apr. 2008. [Online]. Available: https://see.stanford.edu/materials/Isocoe364b/05-localization_methods_notes.pdf
- [32] M. V. Afonso, J.-M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Sep. 2010.



SANGJUN PARK received the B.S. degree in computer engineering from Chungnam National University, Daejeon, South Korea, in 2009. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. His research interests include information theory, numerical optimization, and compressed sensing.



HEUNG-NO LEE (SM'13) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California at Los Angeles, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively. From 1999 to 2002, he was with HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member. From 2002 to 2008, he was an Assistant Professor with the University of Pittsburgh, Pittsburgh, PA, USA. Since 2009, he has been with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. His areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future Internet, and brain-computer interface. He has received several prestigious national awards, including the Top 100 National Research and Development Award in 2012, the Top 50 Achievements of Fundamental Researches Award in 2013, and the Science/Engineer of the Month (January 2014).

• • •

Research Article

Highly Reliable Decision-Making Using Reliability Factor Feedback for Factory Condition Monitoring via WSNs

Zafar Iqbal ¹, Heung-No Lee ², and Saeid Nooshabadi³

¹Department of Computer Science, Michigan Technological University, Houghton, MI 49931, USA

²School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea

³Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, MI 49931, USA

Correspondence should be addressed to Heung-No Lee; heungno@gist.ac.kr

Received 22 May 2018; Revised 27 August 2018; Accepted 12 September 2018; Published 1 October 2018

Academic Editor: André de Almeida

Copyright © 2018 Zafar Iqbal et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cooperation among sensors in a wireless sensor network, deployed for industrial monitoring in an indoor scenario, is a topic of interest in the smart factory and smart city research. The indoor wireless communication channel is very harsh and the observations of all the sensors cannot be sent reliably to the base station. Failure to transmit correct sensing results to the base station may result in false alarms or missed detection of events. Therefore, we propose a cooperation scheme for the wireless sensors to send the data reliably to the base station. Our aim is to increase the reliability of the received information, reduce the probability of error, lower the overall power consumption, and keep the latency to an acceptable low level. We propose a reliability factor feedback algorithm to adjust the weight of unreliable sensors in the decision-making process. The proposed scheme is analyzed based on its latency, power consumption, and packet delivery ratio. Our results show significant improvement in the reliability of the received data, improved packet delivery, and reduced false alarm ratio for full repetition and cluster head-based cooperation. The power consumption and latency in data transmission are also kept to an acceptable low level.

1. Introduction

With the advancement in Internet-of-Things (IoT) and the drive towards smart factory goal, industrial wireless sensor networks (IWSNs) are becoming increasingly important in monitoring the indoor industrial area. The wireless communication link plays a very important role in transmitting the sensed information to a processing unit located in the base station (BS). A broken communication link or a fault in the sensor leads to false alarms or missed detection of events at the BS. This situation may also cause the nodes to repeat transmissions of the data or use higher transmit power leading to higher energy consumption and lower overall throughput of the network. The energy consumption per bit of the network is also affected negatively by the amount of data transmitted by the network nodes and the processing required at the receiver.

In order to improve the reliability of the received information at the BS, a number of methods have been proposed,

including cooperation among wireless nodes to reduce the error due to bad channel conditions. These methods include network coding [1–3], packet loss issues in wireless traffic [4], and relay selection mechanism in networked control system (NCS) for successful cooperative transmission in industrial environments [5]. The work in [6] proposes an energy-efficient scheme to improve packet delivery by using a reliable reactive routing enhancement (R3E) protocol. For the amplify-and-forward (AF) based cooperative communication systems, an adaptive-gain M-relay AF scheme was proposed in [7] in order to achieve good error-rate performance. A solution for machine condition monitoring (MCM) in large factories, which reduces the energy consumption and improves the network throughput, was proposed in [8]. In order to reduce the probability of false alarms sent by the sensors, the work in [9] presents an IWSN-based MCM system. When cooperation among sensor nodes is used in a network consisting of multiple sensors, the aggregation and processing of data at the intermediate sensor nodes play an

important role in the performance and energy consumption of the cooperative multihop communication system. Since all the packets are addressed to a single destination and the size of data packets is usually small, therefore, a reduction in the size of control packet overhead and the number of transmissions can improve the energy efficiency, and throughput, of the system [10–12].

Our recently proposed solution [13] reduces the probability of error in the received information at the BS by using cooperation and data aggregation at the relay nodes. However, [13] uses full repetition of the aggregated data in the cooperation group, which results in an unnecessary redundancy and leads to significant reduction in the throughput, which may be critical to the performance of the network. An improvement to [13] was proposed in [14] by performing partial repetition of the aggregated data at the intermediate nodes with the help of cluster heads. This method reduces the amount of transmissions required to transmit the same information to the BS and also reduces the latency at the expense of some reduction in performance. In the schemes in [13, 14], all the sensors in a cooperation group only share their observations with each other in the first phase unlike [1–3] in which the data is also received at the BS in the first phase. In the second phase, the cooperative information is sent to the BS by either using a full repetition mechanism [13] or using a selected number of cluster head (CH) nodes [14]. In these methods, the relays only detect the received symbols and do not need to decode the symbols, unlike the method in [1–3]. The detected symbols are then used in the cooperation phase even if not correctly received. In order to ensure successful packet delivery, schemes like [5, 6] incur the extra overhead of retransmission but our proposed scheme does not require retransmission. Therefore, it simplifies the hardware and signal processing requirements of the relay node.

This work combines the data aggregation and cooperation mechanisms to improve the reliability of the information received at the BS as well as keeping the redundancy overhead to a certain limit in order to perform with low latency. In this paper, we introduce a parameter called reliability factor, which keeps track of the reliability of information received from a sensor. The reliability factor is obtained by comparing the received information from a sensor with the final result, which is obtained after fusion of the information received from all the sensors within the cooperation group. Based on the reliability factor, we propose an algorithm called reliability factor feedback algorithm (RFFA) to improve the reliability of the final result by adjusting the weights of each participating sensor node in the fusion and decision-making process. A comparison of the latency in data communication and power consumption has been presented between the noncooperation, full repetition (F-Rep.), and cluster head-based cooperation schemes. Also, the packet delivery rate and false alarm rate of the proposed scheme have been compared with some previous related works.

The rest of the paper is organized as follows. Section 2 presents the system model. Section 3 describes the operation of the network. Section 4 presents the performance analysis of the proposed scheme. Section 5 presents simulation results, and Section 6 concludes the paper.

2. System Model

A WSN with indoor non-line-of-sight (NLOS) configuration is considered. The sensor nodes are organized into different cooperation groups based on their geographic proximity to each other. Cooperative transmission is performed within each cooperation group, $\mathcal{V} = \{V_i\}_{i=1}^N$, where N is the maximum number of nodes in a cooperation group. Each node in a cooperation group is able to communicate with the BS in a dual-hop manner. The channels from a source node to an intermediate node, β , and the channels from an intermediate node to the destination, α , are modeled as lognormal distributed Rayleigh fading channels.

2.1. Sensor Coverage and Connectivity. For the purpose of condition monitoring inside an industrial building, the sensors need to be deployed in the form of a static grid or may be deployed randomly. For the deployment of sensor network, we used the static-triangular grid deployment method [15]. The minimum number of sensors required to provide 1-coverage to an area of length, l , and breadth, b , is $\mathcal{N} = 2lb/(r^2\sqrt{27})$, where 1-coverage means that any point in the respective area is covered by at least one sensor and r represents the sensing radius of a sensor. This results in an optimal and regular deployment of sensor nodes making a triangular grid structure. In the resulting triangular grid, every three nodes with intersecting sensing ranges form an equilateral triangle with each side equal to $r\sqrt{3}$.

The minimum number of necessary and sufficient neighbor nodes of a sensor node, required to ensure the connectivity of the network, is given as $\Theta(\log \mathcal{N})$ and ranges between $0.074 \log \mathcal{N}$ and $5.1774 \log \mathcal{N}$ [16]. Accordingly, each node in a cooperation group is assumed to be able to communicate with a minimum of 6 to a maximum of 20 neighbor nodes in this paper. A node decodes the information received only from its neighbor nodes and discards the rest.

2.2. Path-Loss and Shadowing in a Factory Area. As the signal propagates through the walls, machines, and other installations inside a factory area, it creates a shadowing effect which results in the attenuation of the transmit power, referred to as path-loss, and is expressed as a ratio between the transmitted and received power. Path-loss is used to measure the received signal strength (RSS) at the receiver.

In order to find the RSS at each sensor from all other sensors in the cooperation group, we use the lognormal shadowing model. This is a generic model used to predict the propagation loss for a wide range of environments including free space and indoor factory environments [17]. The path-loss measured in dB at a distance d from the transmitter is given by

$$PL_{dB}(d) = PL_{dB}(d_0) + 10\eta \log_{10} \left(\frac{d}{d_0} \right) + X_{\sigma,dB}, \quad (1)$$

where PL_{dB} is the path-loss in dB, η is the path-loss exponent indicating the rate of decay of the mean signal with respect to distance, d_0 is a reference distance, and $X_{\sigma,dB}$ is a zero-mean Gaussian random variable with standard deviation σ

representing the shadowing effect. In (1), $PL_{dB}(d_0)$ is the path-loss in dB at a reference distance d_0 , which is calculated using the Friis free-space propagation model [18]. It is used to model the line-of-sight (LOS) path-loss incurred in the channel, given as

$$P_r(d_0) = P_t \frac{G_t G_r \lambda^2}{(4\pi d_0)^2 L}, \quad (2)$$

where $P_r(d_0)$ is the received signal power in Watts, P_t is the transmitted signal power in Watts, and G_t and G_r are the gains of transmitter and receiver, respectively. λ is the wavelength of the carrier in meters, and L is the system losses which are not associated with propagation loss. Generally, it is more convenient to work in log domain because the transmitted and received powers are usually available in dBm and the antenna gains in dBi. Therefore, the Friis free-space equation is given in log domain as

$$\begin{aligned} PL_{dB}(d_0) &= P_{t,dB} + 10 \log_{10}(G_t) + 10 \log_{10}(G_r) \\ &\quad + 20 \log_{10}(\lambda) - 20 \log_{10}(4\pi d_0) \\ &\quad - 10 \log_{10}(L). \end{aligned} \quad (3)$$

In (3), G_t , G_r , and L are taken equal to 1 as we consider unit gain antennas and the internal system losses are considered as 1, whereas the reference distance d_0 is taken as 1 m. Using (3) in (1) and the parameters suitable for indoor factory non-line-of-sight (NLOS) environments, we can compute the RSS at a receiving node as

$$\begin{aligned} PL_{dB}(d) &= P_{t,dB} + 10 \log_{10}(G_t) + 10 \log_{10}(G_r) \\ &\quad + 20 \log_{10}(\lambda) - 20 \log_{10}(4\pi d_0) \\ &\quad - 10 \log_{10}(L) + 10\eta \log_{10}\left(\frac{d}{d_0}\right) + X_{\sigma,dB} \end{aligned} \quad (4)$$

3. Network Operation

The operation of the network is controlled by using the organize and operate protocol (OOP) [13] and also OOP with cluster heads (OOP-CH) [14]. In these protocols, the nodes are first organized into cooperation groups. Then in OOP-CH, the BS chooses a number of cluster heads based on the received signal strength information (RSSI). After this, the normal operation of sensing and sending the data to the BS by using a two-phase cooperation mechanism starts. In the case of OOP, the sensing and transmit operations start after the nodes are organized into cooperation groups as there are no CH nodes used. A sequence flow diagram of the OOP-CH protocol is presented in Figure 1. The OOP has a similar flow except the controls necessary for CH-based cooperation. Upon receiving the cooperative packets from all the intermediate nodes, the BS performs majority voting-based fusion and makes a decision based on the received information.

3.1. Phase 1. In this phase, a sensor first senses the surrounding area for the intended information. After this, every

sensor in the cooperation group shares its sensed information with the intermediate nodes (\mathcal{C} nodes in the case of CH-based cooperation, as shown in Figure 2) present in its neighborhood by referring to its neighbor list, using BPSK modulation and TDMA scheme. The received signal $r_{i,j}$ at node V_j , from node V_i , in phase 1 is

$$r_{i,j} = \sqrt{E_{s1}} v_i \beta_{i,j} + n_{i,j} \quad (5)$$

where E_{s1} is the transmitted symbol energy in phase 1, v_i is the binary information sent from node V_i , and $n_{i,j}$ is the AWGN with power spectral density, N_0 . The data packet in this phase contains the floor number, sensor ID, time-of-origin (TOO), and the sensed alarm information. See [13] for detail of the data packet.

3.2. Phase 2. In this phase, each intermediate node (\mathcal{C} node in the case of CH-based cooperation, as shown in Figure 2) V_j makes a cooperative data packet by combining the information received from the cooperating nodes within its cooperation group, \mathcal{V} , during the first phase. Then the cooperative data packet denoted by x_j at a node j , which is formed by the aggregation of the received and amplified packets, is transmitted to the BS in a TDMA manner. The received signal at the BS, $y_{j,D}$, can be written as

$$y_{j,D} = \sqrt{E_{s2}} x_j \alpha_{j,D} + n_{j,D} \quad (6)$$

where $\alpha_{j,D}$ represents the lognormal fading channel coefficient from node V_j to the BS and E_{s2} is the transmitted symbol energy in phase 2. $n_{j,D}$ is the AWGN at destination D from node j , with power spectral density, N_0 . The signal from the source node i , relayed via the relay node j and received at the destination D , can be written as

$$y_{j,D} = \sqrt{E_{s1} E_{s2}} \zeta_{i,j} \alpha_{j,D} \beta_{i,j} v_i + n'_{j,D} \quad (7)$$

where $\zeta_{i,j} = 1/\sqrt{E_{s1} |\beta_{i,j}|^2 + N_{0,i,j}}$ represents the amplification factor used at relay node V_j with a corresponding source node V_i and $n'_{j,D} = (\sqrt{E_{s2}}/\sqrt{E_{s1} |\beta_{i,j}|^2 + N_{0,i,j}}) \alpha_{j,D} n_{i,j} + n_{j,D}$. Since the noise terms $n_{i,j}$ and $n_{j,D}$ can be assumed independent, then the equivalent noise $n'_{j,D}$ is a zero-mean complex Gaussian random variable with variance given as $N'_0 = ((E_{s2} |\alpha_{j,D}|^2)/(E_{s1} |\beta_{i,j}|^2 + N_{0,i,j}) + 1) N_0$.

3.3. Reliability Factor Feedback Algorithm. The base station receives the information from the intermediate nodes (either \mathcal{C} CH nodes or all N node in the cooperation group), decodes the information, and combines it at the fusion center by using majority rule decision. In the case of OOP-CH, a majority rule decision, which consists of votes from \mathcal{C} CH nodes in the cooperation group \mathcal{V} , is mathematically represented as

$$R(i) = \arg \max_X \sum_{j=1}^{\mathcal{C}} w_j I(y_j(i) = X) \quad (8)$$

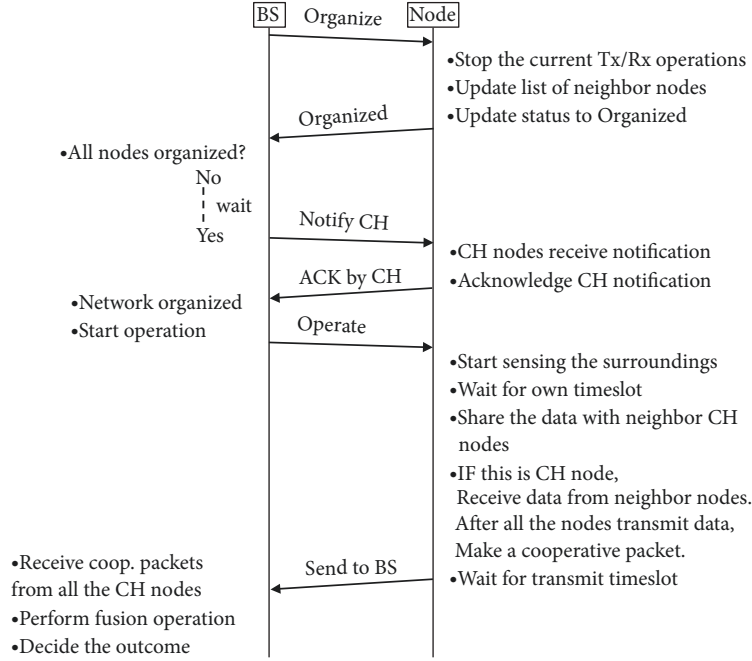


FIGURE 1: The proposed organize and operate protocol with cluster heads (OOP-CH) for WSN.

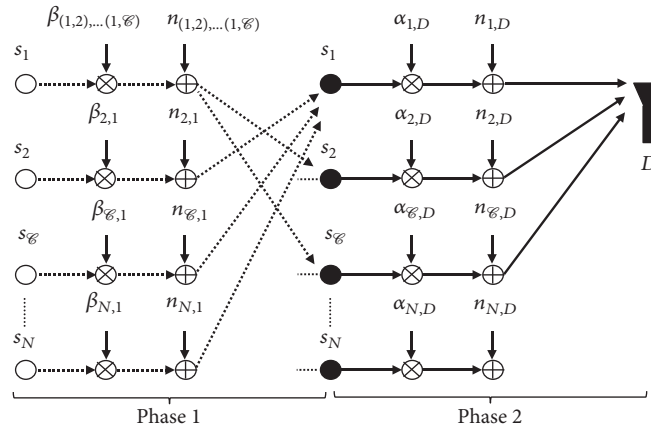


FIGURE 2: Depiction of the two-phase cooperative communication system. In Phase 1, for example, the sensor s_1 sends its information to all CH sensors ($s_2, s_3, \dots, s_{\mathcal{C}}$) during its allocated time slot. Similarly, all the other sensors transmit their information to CH sensors. In Phase 2, only the CH sensors then make a cooperative packet of the aggregated data and transmit it to the destination, D .

where $y_j(i)$ is the i th cooperative symbol received from a sensor j , w_j is the weight associated with the reliability of information received from each intermediate node, and $I(\cdot)$ is an indicator function. In the case of OOP, the summation in (8) is taken over all N nodes in the cooperation group.

In order to achieve a highly reliable result after fusion of the received information, we propose a *reliability factor feedback algorithm*. In this algorithm we compute a reliability factor for each of the intermediate nodes (\mathcal{C} nodes in the case of CH-based cooperation and N nodes in the case of F-Rep. cooperation) by using the result obtained after information fusion at the fusion center. The reliability factor is then fed back to the majority rule fusion and used as the weight w_j of each intermediate node involved in the fusion

process. Figure 3 shows a block diagram of the proposed fusion mechanism with the RFFA for OOP-CH scheme. In case of OOP, the number of sensors are N instead of \mathcal{C} .

3.3.1. Computing the Reliability Factor. The following steps are taken in order to compute the reliability factor, γ for each sensor.

Definition 1. An error report, ε , is defined as a reported observation by a sensor which is different from the final decision after majority rule fusion.

- (1) Find all the error reports made by each sensor in one transmission.

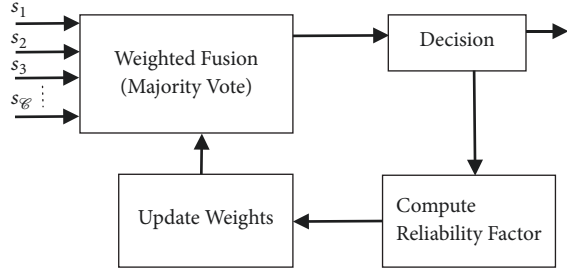


FIGURE 3: Block diagram of the reliability factor feedback algorithm.

- (2) Divide the number of error reports of each sensor by the total number of sensors in a cooperation group.
- (3) Subtract the computed value from 1 to get the reliability factor.

$$\gamma_{i,j} = 1 - \left(\frac{\epsilon_{i,j}}{N} \right) \quad (9)$$

3.3.2. *Updating the Weights.* After the reliability factor has been computed, it is used as the weight of each sensor in the decision-making process. A high reliability factor value of a sensor results in heavier weight of the corresponding sensor in the voting and decision-making process.

- (1) Take an average of the reliability factors of each sensor in the cooperation group, received from the intermediate sensors involved in the fusion process. Divide the computed value by the total number of intermediate sensors (\mathcal{C} nodes in the case of CH-based cooperation and N nodes in the case of F-Rep. cooperation).

$$\begin{aligned} w_{j(F-Rep.)} &= \frac{\sum_{i=1}^N \gamma_{i,j}}{NN} \\ w_{j(ch)} &= \frac{\sum_{i=1}^N \gamma_{i,j}}{N\mathcal{C}} \end{aligned} \quad (10)$$

- (2) Feedback the computed weights w_j to the fusion process given by (8) and as shown in Figure 3.

4. Performance Analysis

In this section, we analyze the performance of the proposed CH cooperation scheme with the help of symbol error rate (SER), latency in transmission, and power consumption.

4.1. *Symbol Error Rate.* The proposed system is a dual-hop communication system with multiple branches. Each relay has multiple branch inputs and repeats the symbols for its neighbor nodes, in a single branch output by using AF scheme, in a TDMA manner to ensure orthogonality of the transmission. The resulting SER can be approximated by

the following equation derived in our previous work [13, Theorem 1],

$$\begin{aligned} P_s(\gamma_{eq,j,D}) &= F \left(1 + \frac{g_{PSK}}{N_0^2 \sin^2 \theta} \left(\frac{\sigma_{j,D}^2 \prod_{i=1}^{N-1} \sigma_{i,j}^2}{\prod_{i=1}^{N-1} \sigma_{i,j}^2 + \sigma_{j,D}^2 + 1} \right) \right) \end{aligned} \quad (11)$$

where $F(x(\theta)) = (1/\pi) \int_0^{(M-1)\pi/M} (1/x(\theta)) d\theta$, M is the modulation symbol size, $g_{PSK} = \sin^2(\pi/M)$, $\gamma_{eq,j,D}$ represents the instantaneous SNR per relay node at the destination, and $\sigma_{i,j}^2$, $\sigma_{j,D}^2$ are the variances of the Rayleigh fading channel coefficients $\beta_{i,j}$ and $\alpha_{j,D}$, respectively. For the proof of the result in (11), please see [13] Appendix A.

From (11), we get P_s as the probability of error in the information received from an intermediate sensor. Since this is a cooperative system with multiple nodes sending information about the same event, therefore, the number of votes needed to decide the final outcome, i.e., the majority, is $l = \lceil (N+1)/2 \rceil$ in the case of F-Rep. cooperation and $l = \lceil (\mathcal{C}+1)/2 \rceil$ in the case of CH-based cooperation. The respective probability of error in the consensus can be computed by using the Binomial theorem,

$$P_{eF-Rep.}(N) = \sum_{m=l}^N \binom{N}{m} P_s^m (1-P_s)^{N-m} \quad (12)$$

$$P_{eCH}(\mathcal{C}) = \sum_{m=l}^{\mathcal{C}} \binom{\mathcal{C}}{m} P_s^m (1-P_s)^{\mathcal{C}-m} \quad (13)$$

4.2. *Latency.* For the sake of a fair comparison between noncooperative and cooperative systems, we assume a traditional relay-based scheme with dual-hop communication for the noncooperative method. In this scheme, a relay node forwards the data from a source node in the second hop towards the BS without any cooperative mechanism involved. Let B represent the number of bits per symbol, and the symbol duration is given by $T_s = 1/f_s$, where f_s is the symbol rate. Then, the throughput in case of noncooperation (T_{nc}), F-Rep. cooperation ($T_{F-Rep.}$), and CH-based cooperation (T_{ch}) dual-hop communication can be written as

$$\begin{aligned} T_{nc} &= \frac{NB}{NT_s + NT_s} \text{bps} \\ T_{F-Rep.} &= \frac{NB}{NT_s + NNT_s} \text{bps} \\ T_{ch} &= \frac{NB}{NT_s + \mathcal{C}NT_s} \text{bps} \end{aligned} \quad (14)$$

where the time taken by two hops to transmit the symbol to BS is represented by the addition in the denominator. Since in F-Rep. cooperation and CH cooperation, each intermediate node relays the data of N or \mathcal{C} nodes in the second phase, respectively, it results in the additional N or \mathcal{C} in the denominator for $T_{F-Rep.}$ and T_{ch} . The delay incurred in transmitting N packets to the BS in the case of noncooperation (\mathcal{D}_{nc}) F-Rep.

cooperation ($\mathcal{D}_{F-Rep.}$) and CH cooperation (\mathcal{D}_{ch}) schemes can then be computed as

$$\begin{aligned}\mathcal{D}_{nc} &= \frac{N \times \text{size of data packet (bits)}}{T_{nc} \text{ (bps)}} \\ \mathcal{D}_{F-Rep.} &= \frac{N \times \text{size of data packet (bits)}}{T_{F-Rep.} \text{ (bps)}} \\ \mathcal{D}_{ch} &= \frac{N \times \text{size of data packet (bits)}}{T_{ch} \text{ (bps)}}\end{aligned}\quad (15)$$

4.3. Power Consumption. In this subsection, we compute the power consumption of the proposed system. In order to simplify our analysis, we do not take into account the power consumed by each sensor during sensing, and the power consumed by the usual processing operations at the BS and the intermediate nodes as these power consumption operations are common among all the schemes compared here in this paper. Hence, we will compute the power consumed in transmitting the information to the BS and the information fusion operation at the BS and compare the noncooperation, F-Rep. cooperation, and CH cooperation schemes. We first compute the energy consumed by these operations and then convert it to power in dBm units as it is easy to visualize. Let E_t , E_i , and E_r represent the energy consumed by the transmit operation by a sensor, idle listening, and reception at a sensor node/BS, respectively. In the case of noncooperative dual-hop communication, each node transmits with energy E_t in phase 1 and the other $N - 1$ nodes receive this information with energy E_r . In phase 2, each relay node transmits with energy E_t to the BS while the other $N - 1$ nodes remain idle, and the BS receives each node's data with energy E_r . Thus the total power consumed (\mathcal{P}_{nc}) is given as

$$\begin{aligned}\mathcal{P}_{nc} &= 10 \log_{10} \left(\frac{1000}{NT_s \times 1W} \times (N (E_t + (N - 1) E_r) \right. \\ &\quad \left. + N (E_t + (N - 1) E_i + E_r)) \right).\end{aligned}\quad (16)$$

For computing the power consumption of the cooperative dual-hop communication, let E_f represent the energy consumed by the fusion operation at the BS. The total power consumed by F-Rep. cooperation ($\mathcal{P}_{F-Rep.}$) and CH cooperation (\mathcal{P}_{ch}) is given as

$$\begin{aligned}\mathcal{P}_{F-Rep.} &= 10 \log_{10} \left(\frac{1000}{NT_s \times 1W} \right. \\ &\quad \times (N (E_t + (N - 1) E_r) \\ &\quad \left. + N (E_t + (N - 1) E_i + E_r) + NNE_f) \right) \\ \mathcal{P}_{ch} &= 10 \log_{10} \left(\frac{1000}{NT_s \times 1W} \times (N (E_t + (N - 1) E_r) \right. \\ &\quad \left. + \mathcal{C} (E_t + (N - 1) E_i + E_r) + \mathcal{C}NE_f) \right)\end{aligned}\quad (17)$$

(18)

TABLE 1: Simulation parameters.

Parameter	Value
Total area	100 m × 100 m
No. of cooperation nodes, N	12, 18
No. of cluster head nodes, \mathcal{C}	3, 5
Carrier frequency	2.4 GHz (ISM Band)
Transmit power, E_{s1}, E_{s2}	1 mW
Standard deviation, σ	7 (Indoor NLOS)
Path-loss exponent, η	3 (Indoor NLOS)
Sensing radius of each sensor, r	18 m

where NN and $\mathcal{C}N$ terms in the numerator of (17) and (18) represent the number of multiply-and-accumulate operations performed to compute the fusion result for either N or \mathcal{C} cooperative packets each containing N number of observations as given in (8). Also, in the second term in the numerator of (18), N is replaced by \mathcal{C} as there are \mathcal{C} CH nodes transmitting to the BS instead of all the N relay nodes.

5. Simulation Results

In order to carry out simulations, we assumed an indoor communication environment with an area of 100 m × 100 m. The indoor area is assumed to contain heavy machines and hard partitioned walls. Rayleigh fading with NLOS lognormal shadowing channel parameters (standard deviation $\sigma = 7$, path-loss exponent $\eta = 3$) is used to model the indoor factory environment [19]. The ISM band carrier frequency of 2.4 GHz is used with a transmit power of 1 mW. Suppose that a fault in the operation or state of the machine at a certain location is evident from higher temperature at that location. We use Gaussian random fields to model this information over the entire area. As the field varies from high temperature to low, four different kinds of alarms, i.e., Danger, Warning, Caution, and OK, are generated, respectively. For the simulation, we choose a cooperation group of either 18 or 12 nodes with 5 or 3 CH nodes, respectively, and the results are averaged over 10,000 sensing operations. To observe the advantage of using RFFA clearly, we deliberately introduced error in the transmission from 3 of the 18 nodes for F-Rep. cooperation and 1 of the 5 CH nodes for CH cooperation in the second phase. The simulation parameters are summarized in Table 1.

5.1. SER. In order to verify our approximated numerical result in (13), we simulate a cooperation group of 18 nodes with 3 CH nodes. Figure 4 shows the plot of the result obtained in (13) compared with the SER obtained from simulation of the CH cooperation scenarios. A similar result for F-Rep. cooperation (12) was published in our previous work [13] and is not shown here. The result shows that the approximation works well to predict the performance of the proposed scheme.

5.2. Latency and Power Consumption. In order to compute the latency and power consumption of the proposed scheme in a practical scenario, we take the example of a Zigbee

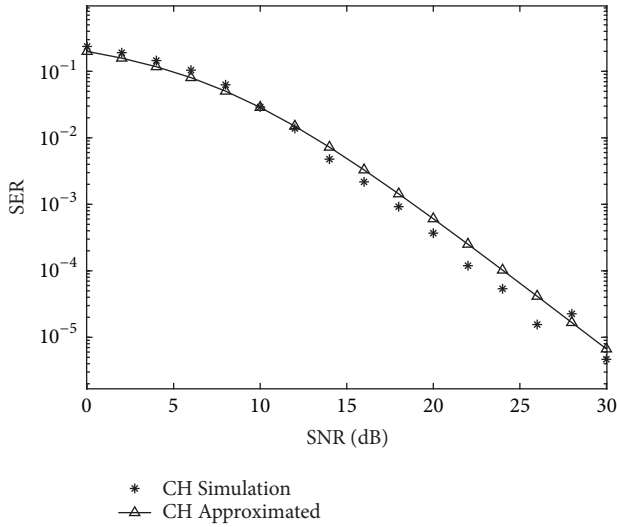


FIGURE 4: Comparison of CH simulation and the approximated result, given in (13). In this experiment, 3 CH nodes were chosen from a cooperation group of 18.

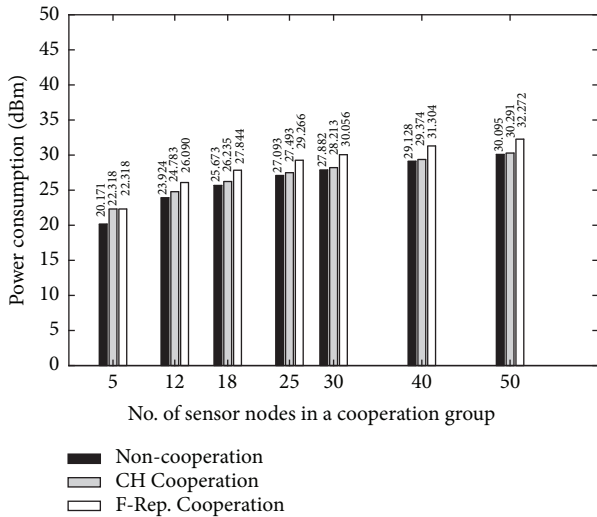


FIGURE 5: Comparison of power consumption results of the proposed CH scheme, F-Rep. scheme, and no cooperation scheme.

(IEEE 802.15.4) based implementation [20]. For the sensor nodes, we utilize the data from Silicon Lab's EFR32 Mighty Gecko Mesh Networking Wireless SoC, which can be used to implement a Zigbee, Bluetooth, Thread, or a proprietary 2.4 GHz wireless sensor network [21]. Therefore, we take $T_s = 50 \mu s$ [20], $E_t = 0.05 \mu J$, $E_i = 0.14 nJ$, $E_r = 1.02 \mu J$ [21], and $E_f = 0.665 \mu J$ [22]. The latency and power consumption results of the F-Rep. cooperation (each node transmits a cooperative packet in phase 2 to the BS, as in [13]), CH cooperation (each CH node transmits a cooperative packet in phase 2 to the BS, as in [14]), and the relayed transmission (a relay node forwards the data for a source node without any cooperation mechanism) are shown in Figures 5 and 6, respectively. The results show increased latency and power

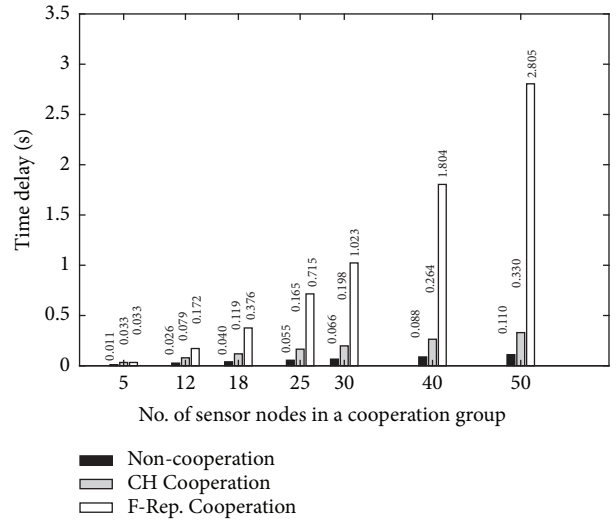


FIGURE 6: Comparison of latency results of the proposed CH scheme, F-Rep. scheme, and no cooperation scheme.

consumption in the case of F-Rep. cooperation and CH cooperation. However, the increase in power consumption in case of F-Rep. cooperation, reported in [13], has been reduced from ~ 2 dB to ~ 0.8 dB for $N = 12$ and from ~ 2.2 dB to ~ 0.2 dB for $N = 50$. The reduction is because the number of relay nodes in the second phase has been reduced from N to \mathcal{C} and only \mathcal{C} CH nodes now transmit to the BS in the second phase instead of all the N nodes. The latency, shown in Figure 6, has been reduced from ~ 145 ms to ~ 53 ms for $N = 12$ and from ~ 2.7 s to ~ 220 ms for $N = 50$, by using CH cooperation method. This is also because of the reduction in the number of relay nodes in the second phase from N to \mathcal{C} . As the number of relay nodes reduce to a suitable number in the form of CH nodes, necessary to obtain cooperation benefit, it helps reduce the power consumption as well the time required to transmit all the information to the BS in order for the BS to be able to make a decision. The results of CH cooperation show a significant improvement in the latency and energy consumption on that reported in [13] and this will be helpful in achieving the low-latency design goal of future communication systems.

5.3. False Alarm and Packet Delivery Rates. We have used the false alarm rate (FAR) and packet delivery rate (PDR) metrics to compare our results with some of the previous works including our own work in [13]. The FAR and PDR were calculated and averaged over a range of SNR (0 to 30 dB) with a total of 10,000 packets for $N = 12$ (F-Rep. cooperation) and $\mathcal{C} = 5$ (CH cooperation). In order to keep the comparison fair, we use the PDR result of [5], when no relay selection mechanism is used, and the PDR result for IWSN given by [6]. As shown in Table 2, the CH cooperation scheme shows significant improvement in the FAR when compared to [9], performs better than [8] and shows increased FAR from that reported in [13]. This work shows an increased FAR than that of [13] because the benefit of cooperation has been reduced from full repetition (N nodes) to partial repetition (\mathcal{C} nodes)

TABLE 2: Comparison with related works.

Performance Metrics	[4]	[5]	[6]	[8]	[9]	[13] F-Rep.	[14] CH
FAR	–	–	–	3.8%	10.5%	1.8%	3.1%
PDR	~84%	~73%	~70%	–	–	~86%	~80%

TABLE 3: Effect of RFFA on performance.

Performance Metrics	F-Rep.	WF-Rep.	CH	WCH
FAR	3.1%	2.9%	7.1%	5.3%
PDR	~78%	~81%	~70%	~73%

in the second phase. The PDR of the CH cooperation scheme is higher than that reported in [5, 6] but lower than the PDR reported in [4, 13]. The reason for this is [4–6] use mechanisms of retransmission, guide-path discovery, and relay selection, respectively, which increases the overhead significantly. In contrast, our work does not involve these overheads and therefore, our results show a higher PDR and lower FAR as compared to these works. Again, the PDR is lower than that of [13] because the benefit of cooperation has been reduced from full repetition (N nodes) to partial repetition (\mathcal{C} nodes) in the second phase.

Table 3 shows the FAR and PDR results of our proposed RFFA scheme in both F-Rep. ($N=18$) and CH ($\mathcal{C}=5$) cooperation methods. Notice that the FAR and PDR of both F-Rep. and CH cooperation schemes drop down significantly from that in Table 2 because of the deliberately induced error in the cooperating nodes (3 nodes in case of F-Rep. and 1 node in case of CH cooperation). Using RFFA mitigates this problem by using the reliability factor associated with each node and improves the FAR and PDR as shown by WF-Rep. and WCH columns in Table 3. This result shows that our proposed RFFA helps in increasing the reliability of the final decision even in the presence of adversely affected sensor nodes because of communication link failure or node failure. The reason for improved FAR and PDR in both F-Rep. and CH cooperation methods is because the proposed algorithm helps reduce errors in the final decision at the BS by disregarding the information from the compromised sensors in the cooperation group.

5.4. Packet Error Rate. Figure 7 compares the packet error rate (PER) of the proposed CH, F-Rep., relayed, and direct transmission schemes. The results show that, using the proposed RFFA, the error induced in either the intermediate nodes (CH cooperation) or any of the cooperating nodes (F-Rep. cooperation) is successfully mitigated, as shown by the dashed lines for both WCH and WF-Rep. cooperation. The F-Rep. cooperation and CH cooperation schemes achieve, on average, 10^{-2} probability of error at almost 20 dB and 12 dB lower SNR compared with the direct (noncooperation) schemes, respectively. Despite the extra energy (~ 0.5 dB for CH and ~ 2 dB for F-Rep., 18 nodes, Figure 5) spent by the network in performing cooperation, the amount of energy saving that can be achieved by using the CH cooperation and F-Rep. cooperation is ~ 11 dB and ~ 18 dB, respectively. The reduced energy saving in CH scheme is a result of the

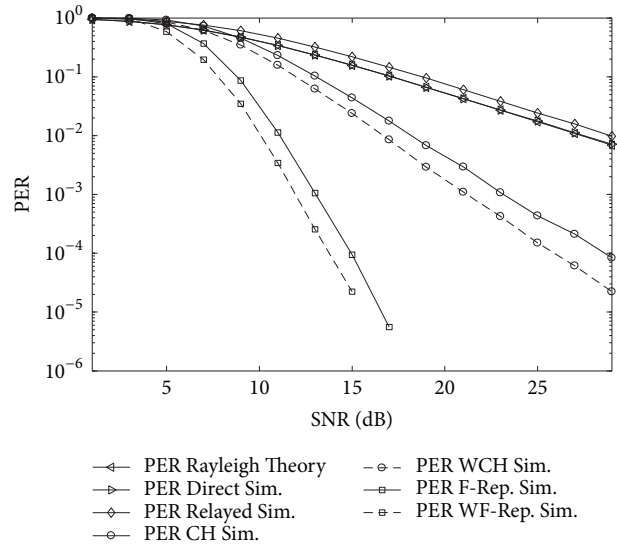


FIGURE 7: Comparison of the PER for direct, relayed, F-Rep. and CH cooperation showing the proposed weighted CH and weighted F-Rep. simulation results.

loss in performance due to using fewer nodes to relay the cooperative packet as compared to F-Rep. cooperation. Using the proposed RFFA mechanism, a further improvement of ~ 2 dB and ~ 1 dB is achieved in energy savings for CH and F-Rep. cooperation, respectively, at 10^{-3} BER. Thus, the CH cooperation scheme is able to reduce the latency and energy consumption of the network at the expense of some performance benefits. Using RFFA (WCH and WF-Rep. cooperation) allows us to save further energy and at the same time, improve the PER of the received data.

6. Conclusion

In this paper, we have proposed an algorithm called reliability factor feedback algorithm to improve the reliability of decisions made after sensor data fusion in relay-based cooperative WSNs to monitor the indoor industrial environment. We have analyzed the SER, power consumption, and latency of the proposed scheme. With the proposed algorithm, the reliability of the final decision has been increased significantly at the fusion center. Along with the increased reliability, significant energy savings have been achieved, which can be very beneficial in increasing the lifetime of the sensors.

Data Availability

The code used to model the above described network and generate the provided results can be found at <https://infonet.gist.ac.kr>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MSIP) (NRF-2018RIA2A1A19018665) and in part by the Institute of Computing and Cybersystems (ICC) at Michigan Technological University.

References

- [1] X. Bao and J. Li, "Adaptive Network Coded Cooperation (ANCC) for wireless relay networks: Matching code-on-graph with network-on-graph," *IEEE Transactions on Wireless Communications*, vol. 7, no. 2, pp. 574–583, 2008.
- [2] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, 2003.
- [3] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.
- [4] A. Ulusoy, O. Gurbuz, and A. Onat, "Wireless model-based predictive networked control system over cooperative wireless network," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, pp. 41–51, 2011.
- [5] N. Marchenko, T. Andre, G. Brandner, W. Masood, and C. Bettstetter, "An experimental study of selective cooperative relaying in industrial wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 3, pp. 1806–1816, 2014.
- [6] J. Niu, L. Cheng, Y. Gu, L. Shu, and S. K. Das, "R3E: reliable reactive routing enhancement for wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 784–794, 2014.
- [7] Y. M. Khattabi and M. M. Matalgah, "Performance analysis of multiple-relay AF cooperative systems over rayleigh time-selective fading channels with imperfect channel estimation," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 1, pp. 427–434, 2016.
- [8] J. Neuzil, O. Kreibich, and R. Smid, "A distributed fault detection system based on IWSN for machine condition monitoring," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1118–1123, 2014.
- [9] O. Kreibich, J. Neuzil, and R. Smid, "Quality-based multiple-sensor fusion in an industrial wireless sensor network for MCM," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 9, pp. 4903–4911, 2014.
- [10] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-network aggregation techniques for wireless sensor networks: a survey," *IEEE Wireless Communications Magazine*, vol. 14, no. 2, pp. 70–87, 2007.
- [11] P. Jesus, C. Baquero, and P. S. Almeida, "A Survey of Distributed Data Aggregation Algorithms," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 381–404, 2015.
- [12] H. Harb, A. Makhoul, R. Tawil, and A. Jaber, "Energy-efficient data aggregation and transfer in periodic sensor networks," *IET Wireless Sensor Systems*, vol. 4, no. 4, pp. 149–158, 2014.
- [13] Z. Iqbal, K. Kim, and H.-N. Lee, "A cooperative wireless sensor network for indoor industrial monitoring," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 482–491, 2017.
- [14] Z. Iqbal and H. Lee, "Low-Latency and High-Reliability Cooperative WSN for Indoor Industrial Monitoring," in *Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, Sydney, NSW, June 2017.
- [15] R. Williams, *The Geometrical Foundation of Natural Structure. A Source Book of Design*, Dover, New York, NY, USA, 1979.
- [16] F. Xue and P. R. Kumar, "The Number of Neighbors Needed for Connectivity of Wireless Networks," *Wireless Networks*, vol. 10, no. 2, pp. 169–181, 2004.
- [17] Y. Ai, M. Cheffena, and Q. Li, "Radio frequency measurements and capacity analysis for industrial indoor environments," in *Proceedings of the 9th European Conference on Antennas and Propagation (EuCAP 2015)*, p. 1, May 2015.
- [18] A. F. Molisch, *Wireless communications*, John Wiley Sons, New York, NY, USA, 2011.
- [19] W. H. Tranter, K. S. Shanmugan, T. S. Rappaport, and K. L. Kosbar, *Principles of Communication Systems Simulation with Wireless Applications*, Prentice Hall, Upper Saddle River, NJ, USA, 2004.
- [20] IEEE 802.15.4-2011, *Part 15.4: Low-rate wireless personal area networks (LR-WPANs)*, IEEE-SA Standards Board, 2011.
- [21] "Silicon Labs, EFR32 Mighty Gecko Mesh Networking Wireless SoC," <https://www.silabs.com/>.
- [22] K. Wu, C. Liang, K. Yu, and S. Kuang, "Multiple-mode floating-point multiply-add fused unit for trading accuracy with power consumption," in *Proceedings of the 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*, pp. 429–435, Niigata, Japan, June 2013.






Hindawi

Submit your manuscripts at
www.hindawi.com



Article

Reduced Computational Complexity Orthogonal Matching Pursuit Using a Novel Partitioned Inversion Technique for Compressive Sensing

Seonggeon Kim ^{1,†} , Uihyun Yun ^{1,†}, Jaehyuk Jang ¹ , Geunsu Seo ², Jongjin Kang ³, Heung-No Lee ¹ and Minjae Lee ^{1,*} 

¹ The School of Electrical and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea; okkskso@gist.ac.kr (S.K.); qnrmsajfl@naver.com (U.Y.); jjh2014@gist.ac.kr (J.J.); heungno@gist.ac.kr (H.-N.L.)

² The LG Innotek, Ansan 426791, Korea; seogs89@naver.com

³ The Hanwha System, Seongnam 13524, Korea; jongjin.kang@hanwha.com

* Correspondence: minjae@gist.ac.kr; Tel.: +82-062-715-2205

† Both authors contributed equally to this work.

Received: 22 August 2018; Accepted: 14 September 2018; Published: 18 September 2018



Abstract: This paper reports a field-programmable gate array (FPGA) design of compressed sensing (CS) using the orthogonal matching pursuit (OMP) algorithm. While solving the least-squares (LS) problem in the OMP algorithm, the complexity of the matrix inversion operation at every loop is reduced by the proposed partitioned inversion that utilizes the inversion result in the previous iteration. By the proposed matrix ($n \times n$) inversion method inside the OMP, the number of operations is reduced down from $O(n^3)$ to $O(n^2)$. The OMP algorithm is implemented with a Xilinx Kintex UltraScale. The architecture with the proposed partitioned inversion involves 722 less DSP48E compared with the conventional method. It operates with a sample period of 4 ns, signal reconstruction time of 27 μ s, and peak signal to noise ratio (PSNR) of 30.26 dB.

Keywords: compressed sensing (CS); field programmable gate array (FPGA); high-level synthesis (HLS); partitioned inversion; orthogonal matching pursuit (OMP)

1. Introduction

The compressed sensing (CS) can effectively acquire and reconstruct sparse signals with significantly less samples than that required from the Nyquist–Shannon sampling theorem [1]. The reconstruction process in CS finds the best solution to an underdetermined system, with a linear equation of the form $\mathbf{y} = \mathbf{F}\mathbf{x}$, where we know the measurement matrix, \mathbf{F} , that model the sampling system and the measurement vector, \mathbf{y} , while the original signal, \mathbf{x} , remains to be determined.

Various algorithms have been proposed to reconstruct the signal \mathbf{x} from the compressively sensed samples. Generally, two algorithms, the greedy pursuit [2,3] and the convex relaxation [4,5], are mainly selected for sparse signal reconstruction. The greedy pursuit is a more useful CS algorithm than the convex relaxation, because it uses floating point operations [6]. Orthogonal matching pursuit (OMP) is one of the representative greedy-type solvers for CS, which finds columns of the measurement matrix, \mathbf{F} , that are mostly correlated with the current estimate, $\tilde{\mathbf{x}}$, of the original signal for m -iterations, where m is the sparse level, and updates an advanced signal estimate from a least-squares (LS) method.

In OMP, one of the major problems in the LS step is the matrix inversion, because it results in a high computational complexity per iteration [7]. Several inversion methods for the OMP algorithm have been proposed, such as the QR decomposition [8] and Cholesky-based factorization [7,9], to improve the computation efficiency of the matrix inversion. However, OMP that utilizes the partitioned

inversion has not been presented, to the best of authors' knowledge, because the conventional partitioned inversion may result in a low efficiency for computation. Thus, we propose a novel matrix inversion with a better computational efficiency, based on the incremental computation of the partitioned inversion targeting the OMP.

In this paper, we utilize three properties of the input matrix for the inversion in each OMP iteration, namely: conjugate symmetry, positive definiteness, and overlapped regions. It is found that the properties reduce the computation complexity by re-utilizing the computation results obtained in the previous OMP iteration. In terms of the comparison of the computation complexity, the novel partitioned inversion method improves the complexity over the conventional Cholesky-based inversion and the conventional partitioned inversion based on our derived equations. In addition, multiple measurement vectors (MMV) are applied to the OMP algorithm to improve the sparse signal recovery compared to the single measurement vector (SMV) method, and it is called simultaneous OMP (SOMP) [10]. Lastly, we have implemented SOMP with the proposed matrix inversion method in the field-programmable gate array (FPGA) and measured the running time. The experiments show that the total hardware utilization is significantly reduced compared with the conventional partitioned inversion, and the reconstruction time is 27 μ s. Section 2 introduces the overview of the SOMP algorithm, and the conditions of the input matrix in LS problem are described in Section 3. Section 4 proposes the novel partitioned inversion, and the experimental results are presented in Section 5. Finally, Section 6 provides a conclusion.

2. Overview of SOMP Algorithm

2.1. Description of SOMP Algorithm

In the SOMP, the linear equation is defined by the following:

$$\mathbf{y} = \mathbf{F}\mathbf{x}, \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^{M \times L}$, $\mathbf{F} \in \mathbb{R}^{M \times N}$, and $\mathbf{x} \in \mathbb{R}^{N \times L}$. The SOMP process given in Algorithm 1 [11] consists of the optimization problem (1 and 2) and the LS problem (3 and 4). For this process, the residue, \mathbf{r}^0 (when $i = 1$) is initially set to \mathbf{y} . During the i th iteration, the optimization problem chooses one of the columns of \mathbf{F} , which is strongly correlated to the residue of \mathbf{y} , and then searches the position, k , of this column. The \mathbf{F}_k is the sub-matrix including the column according to the k , and \mathbf{F}^i is updated by summing \mathbf{F}_k with the previous sub-matrix. The LS problem removes the contributed column for the new estimate, $\tilde{\mathbf{x}}$, and then computes a new residue, \mathbf{r} . Finally, when the m -iteration is achieved, the final estimate of the original signal is computed.

Algorithm 1. Simultaneous orthogonal matching pursuit (SOMP).

Input:

- $\mathbf{F} \in \mathbb{R}^{M \times N}$: The measurement matrix
- $\mathbf{y} \in \mathbb{R}^{M \times L}$: The multiple measurement vector

Output:

- $\tilde{\mathbf{x}} \in \mathbb{R}^{N \times L}$: The estimate of original signal

Variable:

- m : The sparsity level of original signal \mathbf{x}
- $\mathbf{r} \in \mathbb{R}^{M \times L}$: The residue

Initialize: $\mathbf{r}^0 = \mathbf{y}$, $\tilde{\mathbf{x}} = 0$

For i th iteration:

1. $k^i = \underset{j}{\operatorname{argmax}} \|\langle \mathbf{r}^{i-1}, \mathbf{F}_j \rangle\|_2$, where $1 \leq j \leq N$, and k is an index
2. $\mathbf{F}^i = [\mathbf{F}_{k^{i-1}} \ \mathbf{F}_{k^i}]$
3. $\tilde{\mathbf{x}}^i = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{F}^i \mathbf{x}\|_2$
4. $\mathbf{r}^i = \mathbf{y} - \mathbf{F}^i \tilde{\mathbf{x}}^i$

Repeat process until $i = m$ to generate the final estimate of the \mathbf{x} .

2.2. Least Square (LS) Problem

In the SOMP algorithm, the recovered signal, $\tilde{\mathbf{x}}$, gradually becomes similar to the input signal, \mathbf{x} , with more iterations, and the equation for the new residue in the LS problem is as follows:

$$\mathbf{r}_i = \mathbf{y} - \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{F}^i \tilde{\mathbf{x}}^i = \mathbf{y} - \mathbf{F}^i (\mathbf{U}^i)^{-1} (\mathbf{F}^i)^H \mathbf{y} \tag{2}$$

In Equation (2), the computation complexity reduction of the inverse of the input matrix, \mathbf{U} , is a challenge in the OMP algorithm.

3. Conditions of the Input Matrix in LS Problem

The input matrix, \mathbf{U} , in the LS problem presents the various conditions, as below. The \mathbf{U} is always symmetric, as follows:

$$\mathbf{U}^T = (\mathbf{F}^T \mathbf{F})^T = \mathbf{F}^T (\mathbf{F}^T)^T = \mathbf{F}^T \mathbf{F} \tag{3}$$

which is a positive definite, and is expressed by the following:

$$\mathbf{x}^T (\mathbf{F}^T \mathbf{F}) \mathbf{x} = (\mathbf{F}\mathbf{x})^T (\mathbf{F}\mathbf{x}) = \mathbf{F}\mathbf{x}_2^2 > 0, \text{ If } \mathbf{x} \neq 0, \text{ then } \mathbf{F}\mathbf{x} \neq 0 \tag{4}$$

Equation (4) indicates that the eigenvalue, $\mathbf{F}\mathbf{x}$, is always greater than zero. That is, $\mathbf{U}(=\mathbf{F}^T \mathbf{F})$ satisfies the positive definite condition, while following the characteristics: (1) \mathbf{U}^{-1} exists, (2) $c_{tt} > 0$ for all t in $\text{diag}(c_{11}, c_{22}, c_{33}, \dots, c_{tt})$ and (3) principle sub-matrices, are also positive definite.

In step 2 of Algorithm 1, \mathbf{F}^i is added one column per loop, and then the input matrix, \mathbf{U} , is determined. As the loop progresses, the next \mathbf{U}^i includes the matrix \mathbf{U}^{i-1} for the previous loop, as shown in Table 1. For example, \mathbf{U}^2 is an expanded matrix of \mathbf{U}^1 of the previous iteration. Thus, \mathbf{U}^i is divided into a previously obtained part and a newly added part.

Table 1. Overlapped region in input matrix.

Measurement Matrix	Input Matrix
$\mathbf{F}^1 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$	$\mathbf{U}^1 = (\mathbf{F}^1)^H \mathbf{F}^1 = [c_{11}]$
$\mathbf{F}^2 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$	$\mathbf{U}^2 = (\mathbf{F}^2)^H \mathbf{F}^2 = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} [U^1] & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$
$\mathbf{F}^3 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$	$\mathbf{U}^3 = (\mathbf{F}^3)^H \mathbf{F}^3 = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} & & c_{13} \\ [U^2] & & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$

4. Proposed Partitioned Inversion

By leveraging the properties of the input matrix as examined in Section 3, we propose a novel partitioned inversion method after explaining the conventional partitioned inversion in this section.

4.1. Conventional Partitioned Inversion

The conventional partitioned inversion is a method to obtain the inverse matrix by dividing the input matrix into four parts. This inversion is robust against noise, because of a low number of conditions, and operates using lower parts of the input matrix in symmetric and positive definite (SPD) characteristics. The step for the inverse matrix of the conventional way is introduced in the left column of in Table 2. However, as the matrix size increases, the number of computations grows significantly.

Table 2. Conventional partitioned inversion versus proposed partitioned inversion.

Input Matrix $U = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$	Inversion of Input Matrix $U^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$
Conventional Partitioned Inversion	Proposed Partitioned Inversion
Step 1. A^{-1} 2. CA^{-1} 3. $A^{-1}B$ 4. $CA^{-1}B$ 5. $D - CA^{-1}B$ 6. $H = (D - CA^{-1}B)^{-1}$ 7. $F = -A^{-1}B(D - CA^{-1}B)^{-1}$ 8. $G = -(D - CA^{-1}B)^{-1}CA^{-1}$ 9. $A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$ 10. $E = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$	Step 1. CA^{-1} 2. $CA^{-1}C^T$ 3. $D - CA^{-1}C^T$ 4. $H = (D - CA^{-1}C^T)^{-1}$ 5. $G^T, F = -(CA^{-1})^T(D - CA^{-1}C^T)^{-1}$ 6. $(CA^{-1})^T(D - CA^{-1}C^T)^{-1}CA^{-1}$ 7. $E = A^{-1} + (CA^{-1})^T(D - CA^{-1}C^T)^{-1}CA^{-1}$

4.2. Proposed Partitioned Inversion

In order to reduce the inversion computational complexity, the following relationships are obtained by utilizing the input matrix properties presented in Section 3.

$$A = A^T \rightarrow A^{-1} = (A^{-1})^T, D = D^T \rightarrow D^{-1} = (D^{-1})^T, B = C^T \tag{5}$$

As the input matrix, U , is symmetric, the constituent matrices of the input matrix (A, B, C , and D) satisfy Equation (5), which is used in the following optimization processes.

Step 1 in the conventional partitioned inversion in Table 2 is eliminated by the following observation. In i th OMP iteration, U^i is portioned into four sub-matrixes containing U^{i-1} , as shown in Figure 1. Unlike the conventional partitioned inversion, this step contains U^{i-1} , of which the inversion is already available in the previous OMP iteration. Therefore, step 1 is unnecessary, by utilizing $(U^{i-1})^{-1}$, shown in gray in Figure 1.

Steps 2 and 3 are decreased down to a single step based on Equation (5).

$$(A^{-1}B)^T = (A^{-1}C^T)^T = C(A^{-1})^T = CA^{-1} \tag{6}$$

Steps 7 and 8 are also further simplified by Equation (5), as below.

$$F = -A^{-1}B(D - CA^{-1}B)^{-1} \tag{7}$$

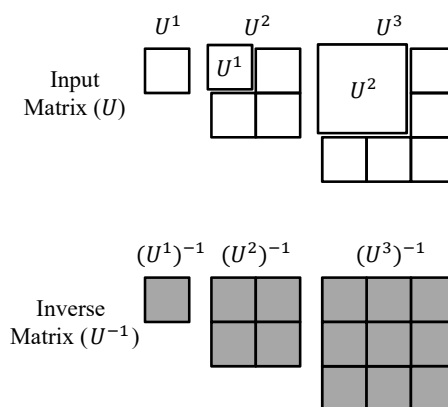


Figure 1. Reduction of computation complexity by copying the inverse matrix from previous iteration.

$$G^T = \left(-(D - CA^{-1}B)^{-1}CA^{-1} \right)^T = (A^{-1})^T C^T \left((D - CA^{-1}B)^{-1} \right)^T = -A^{-1}B(D - CA^{-1}B)^{-1} \quad (8)$$

From Equations (5)–(8), the number of steps in the inversion process is reduced from 10 to 7, as shown in Table 2.

4.3. Computational Complexity for Proposed Partitioned Inversion

Table 3 shows the complexity comparison between three inversion techniques. The Cholesky-based and conventional partitioned inversion continually calculates the inversion of the size-increased matrix per iteration, whereas our inversion method computes the extra sub-matrices except for prior iteration calculation. Accordingly, the proposed method improves the computational complexity for multiplication, addition/subtraction, and division, depending on the input matrix size, where the matrix size increases from 1×1 to $m \times m$ per loop. To visualize the improved complexity, we executed a MATLAB (2013b, The MathWorks, Natick, MA, USA) simulation for a computational comparison of the total number between the Cholesky-based inversion, conventional partitioned inversion, and proposed partitioned, depending on the matrix size ($m \times m$), as depicted in Figure 2.

Table 3. Computation Complexity Comparison of input matrix inversion (for $m \times m$ matrix U and a single supporter system).

Inversion Operation	Cholesky-Based [7]	Conventional Partitioned	Proposed Partitioned
Multiplication	$\frac{4m^3+3m^2-7m}{6}$	$m^3 - m$	$2m^2 - 2m$
Add/sub	$\frac{m^3-m^2}{2}$	$m^3 - 2m^2 + m$	$2m^2 - 4m + 2$
Division	m	m	1

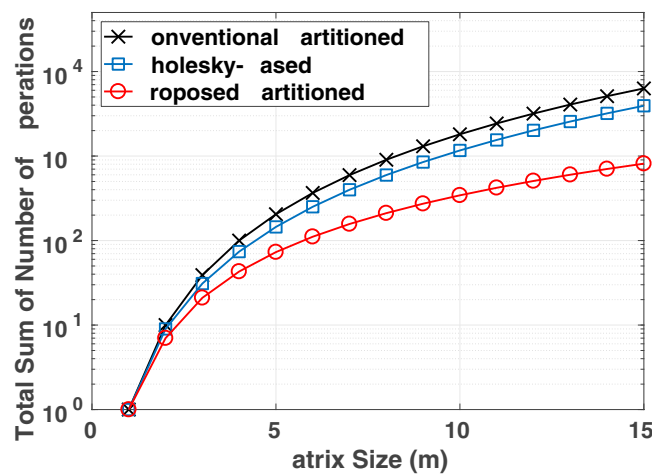


Figure 2. The total sum of number of operations for the Cholesky-based inversion, conventional partitioned inversion, and proposed partitioned inversion, depending on matrix size ($m \times m$).

4.4. Proposed Partitioned Inversion for Multiple Supporter System

Computation complexity functions in Table 3 are only applied to a single supporter system; however, our inversion method is also used in the multiple supporter system, and the complexity equations are changed as Equations (9)–(11).

- Multiplication:

$$(2m^2 - 1)n - 2mn^2 + n^3 \quad (9)$$

- Add/sub:

$$(2m^2 - 2m + 1)n - 2mn^2 + n^3 \tag{10}$$

- Division:

$$n \tag{11}$$

where n is the number of the supporter. Although the computation complexity becomes larger as n increases, the proposed inversion method can be applied to the multiple supporter system.

4.5. SOMP Structure with the Proposed Partitioned Inversion

The SOMP structure with the proposed partitioned inversion is organized as shown in Figure 3. The correlation matching block (CRMB) finds the location, k , of the most correlated column of the measurement matrix, F , and F^i is updated by the summing of F^{i-1} (which is saved by the F memory block at the previous iteration) and F_{k^i} . Generating the U block (GUB) identifies the row of the measurement matrix corresponding to k , and generates the matrix U by the inner-product. The inverse of U is generated by the inversion U block (IUB), which is stored in the U^{-1} memory block, and the memory offers the previous calculated inverse matrix of U . Thus, IUB can reduce the computational complexity by using the previously computed inverse matrix of U . The least square block (LSB) calculates the estimate, \tilde{x} , of the original signal, and the residual, r^i , is generated by the residual block (RB). Finally, the final estimate signal is obtained by the system at m -iteration.

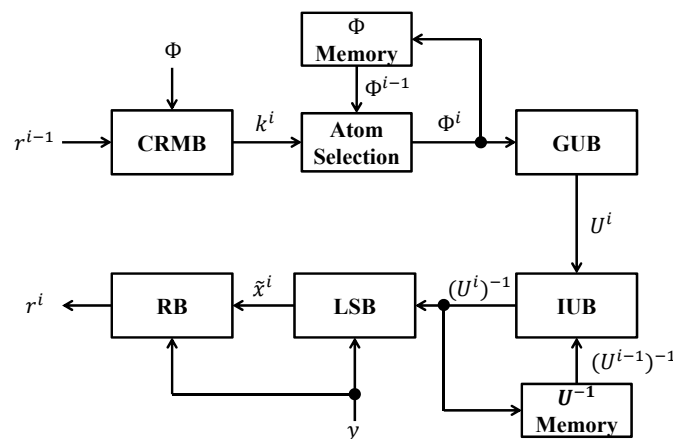


Figure 3. Simultaneous orthogonal matching pursuit (SOMP) structure with the proposed partitioned inversion.

5. Experiment Results

5.1. FPGA Implementation Approach

In our experiment, the Xilinx Kintex UltraScale board and the XCKU115-FLVA2104-2-I chipset was used to demonstrate the proposed inverse matrix and the operation of the entire SOMP algorithm. The board is advantageous to the design of a large algorithm, because it has a large number of DSP slices.

The high-level synthesis (HLS) tool provided by Xilinx for the FPGA design offers many advantages, namely: (1) arbitrary precision data type, math, IP, linear algebra, and many other libraries; (2) the register-transfer level (RTL) can be extracted automatically, and RTL verification is easy, because of the provided co-simulation; (3) as C synthesis provides design resources, such as clock, area, and I/O port description, a user can easily change the design according to their intention; (4) it enables high-performance hardware design with little hardware knowledge, because it facilitates its utilization for the IP module in other design tools, such as the system generator. By leveraging these

advantages, we convert the C/C++ code into the hardware description language (HDL), such as Verilog, which is subsequently synthesized to the gate level for the optimization of our design architecture.

Our device is based on a 16- and 32-bit fixed-point complex operation. Most of the calculations were performed with 16-bit, but the partitioned inversion process relies on 32-bit, because the divide operation requires more precise computation.

The parameters used in our experiment are as follows. The number of physical channels is four and the dimensions of the measurement matrix ($M \times N$) are 32 and 128, respectively. The channel reuse factor is 8, MMV extension is 16, Nyquist bandwidth is 4 GHz, sampling frequency is 250 MHz, F_p is 31.25 MHz, and FFT-point is 128.

5.2. Additional Optimisation in FPGA Implementation

In the SOMP algorithm, a large matrix inner product is needed. To express the matrix inner product in the code, we require three *for* loops. We apply a pipeline to the second *for* loop to perform real-time signal reconstruction. Although the number of DSP increases in proportion to the number of the innermost *for* loop, the latency is greatly reduced, because the calculation can be parallel.

Most of the operations that occurred inside it are complex types, and is thus accompanied by an increase in the computational complexity. Even if the partitioned inversion device requires complex division, it can be replaced by two real divisions because the denominator is always a positive real number.

5.3. SOMP Hardware Utilization

Various components are required for hardware realization of the SOMP algorithm. The block ram (BRAM) effectively stores the vectors and matrices, namely: the measurement vector, \mathbf{y} , measurement matrix, \mathbf{F} , and residue, \mathbf{r} . The DSP is the pre-built multiply-accumulate (MAC) circuit in the FPGA and is an essential hardware, because it quickly performs the multiplication and addition/subtraction operations. Flip-flop (FF) is the shift register used to synchronize the logic. The total hardware utilization for the proposed partitioned inversion is reduced compared to the conventional partitioned inversion, as shown in Table 4. However, only the BRAM for our inversion method is increased compared to the conventional approach because both the \mathbf{U} matrix and previous \mathbf{U} matrix must be stored simultaneously.

Table 4. OMP hardware utilization comparison between conventional partitioned inversion and proposed partitioned inversion. BRAM—block ram; FF—flip-flop.

Xilinx Kintex UltraScale XCKU115	Conventional	Proposed
BRAM	283 (13.1%)	307 (14.2%)
DSP48E	2754 (49.9%)	2032 (36.8%)
FF	225,337 (17%)	210,577 (15.9%)
LUT	190,742 (28.7%)	153,447 (23.1%)

5.4. Signal Reconstruction

We executed the MATLAB simulation and HLS to validate the recovered signal from the input sparse signal, and data precision is converted from 16(0.12) to 32(0.24) bits, where the data format p.(f) with the precision (p) and fractional bits (f). Figure 4a presents the input sparse signal, which is mixed by the random carrier signals. The signal is well reconstructed by both the simulation and HLS, as shown in Figure 4b,c, respectively.

From our experiment, we obtained the peak signal to noise ratio (PSNR) for an objective evaluation [7], where PSNR is as follows:

$$\text{PSNR} = 20 \log_{10} \left(\frac{\text{MAX}}{\sqrt{\text{MSE}}} \right) \quad (12)$$

$$\text{MSE} = \frac{1}{N} \times \sum_i [x^i - \tilde{x}^i]^2 \quad (13)$$

where MAX is the maximum possible value of the signal x , N is the total number of samples, \tilde{x}^i is the sample value at i in the reconstructed signal, and x^i is the sample value at i in the original signal. The PSNR of our SOMP structure was determined to be 30.26 dB for 16- and 32-bit data precision.

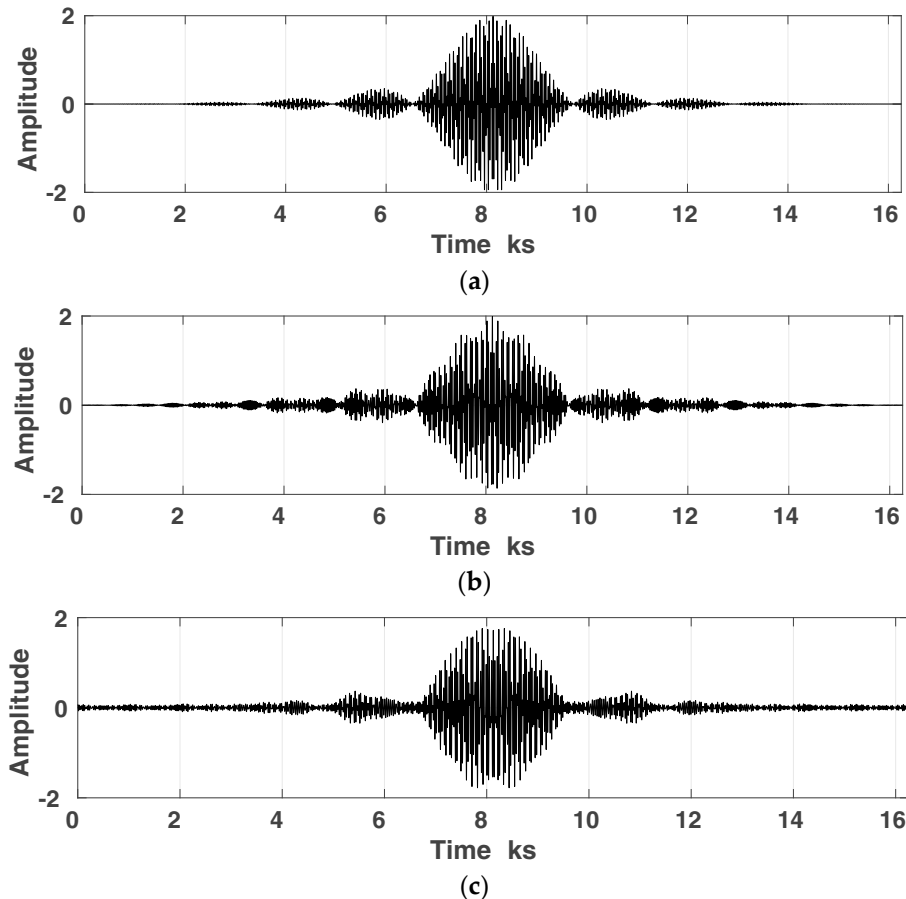


Figure 4. Reconstruction of the sparse modulated by random carrier frequencies for (a) input signal, (b) reconstruction signal by MATLAB, and (c) reconstruction signal by high-level synthesis (HLS) for 16- and 32-bit data precision.

5.5. Performance Comparison

Table 5 provides a comparison of the performance of existing sparse signal recovery algorithm devices. Although the other works summarized in Table 5 focus on the OMP algorithm rather than the inversion method, an objective comparison is required to highlight our proposed inversion method. For this reason, we selected papers with the identical measurement matrix size (32×128), while applying the different inversion methods (i.e., QR decomposition-based and Cholesky-based inversions). Although our work uses a higher clock frequency than other works in Table 5, our reconstruction time is sufficiently fast with larger sparsity than in the literature [7,12,13].

Table 5. Performance comparison for the 32×128 measurement matrix size.

	Sparsity	Clock Frequency	Reconstruction Time	Inversion Type	Data Format
Intel Core Duo [7]	5	2.8 GHz	606 μ s	Cholesky-based	32-bit fixed-point real data
FPGA Virtex 7 [12]	5	0.165 GHz	18.3 μ s	QR decomposition-based	32-bit fixed-point hybrid complex data
FPGA Virtex 5 [13]	5	0.039 GHz	24 μ s	Cholesky-based	32-bit fixed-point real data
FPGA Kintex UltraScale [This Work]	8	0.25 GHz	27 μ s	Proposed partitioned inversion	16- and 32-bit fixed-point complex data

6. Conclusions

The SOMP algorithm generally results in a high computational complexity for the LS problem, because of a very large inner product. In addition, the SMV model causes an increase of this complexity. To reduce the complexity, we have proposed the SOMP algorithm for the MMV models, while applying a novel partitioned inversion that omits the recalculation for the inversion of the input matrix at the previous loop. Accordingly, our proposed inversion method restores a signal with less hardware compared with the conventional partitioned inversion techniques. By using the Kintex UltraScale board and HLS, we verified the signal reconstruction with a high PSNR and obtained a fast reconstruction time. Therefore, in comparison with the conventional partitioned inversion and Cholesky-based factorization, our proposed inversion is suitable to apply the diverse applications requiring the high-level restoration with less hardware, such as bio-signals, bio-medical imaging, and radar.

Author Contributions: Conceptualization, M.L.; Formal analysis, U.Y. and G.S.; Investigation, S.K. and J.J.; Methodology, U.Y.; Project administration, J.K.; Supervision, H.-N.L. and M.L.; Validation, J.J.; Visualization, S.K.; Writing—original draft, S.K.

Funding: This work was supported by a grant-in-aid of HANWHA SYSTEMS.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Candes, E.; Wakin, M. An introduction to compressive sampling. *IEEE Signal Process. Mag.* **2008**, *25*, 21–30. [[CrossRef](#)]
2. Pati, Y.; Rezaiifar, R.; Krishnaprasad, P. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Proceedings of the IEEE Record of The Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 1–3 November 1993; pp. 40–44.
3. Tropp, J.; Gilbert, A. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **2007**, *53*, 4655–4666. [[CrossRef](#)]
4. Chen, S.; Donoho, D.; Saunders, M. Atomic decomposition by basis pursuit. *SIAM Rev.* **2001**, *43*, 129–159. [[CrossRef](#)]
5. Khan, I.; Singh, M.; Singh, D. Compressive sensing-based sparsity adaptive channel estimation for 5G massive MIMO systems. *Appl. Sci.* **2018**, *8*, 2076–3417. [[CrossRef](#)]
6. Sahoo, S.; Makur, A. Signal recovery from random measurements via extended orthogonal matching pursuit. *IEEE Trans. Signal Process.* **2015**, *63*, 2572–2581. [[CrossRef](#)]
7. Rabah, H.; Amira, H.; Mohanty, B.; Almaadeed, S.; Meher, P. FPGA implementation of orthogonal matching pursuit for compressive sensing reconstruction. *IEEE Trans. Very Large Scale Integr. Syst.* **2015**, *23*, 2209–2220. [[CrossRef](#)]
8. Bai, L.; Maechler, P.; Muehlberghuber, M.; Kaeslin, H. High-speed compressed sensing reconstruction on FPGA using OMP and AMP. In Proceedings of the 19th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), Seville, Spain, 9–12 December 2012; pp. 53–56.
9. Blache, P.; Rabah, H.; Amira, A. High level prototyping and FPGA implementation of the orthogonal matching pursuit algorithm. In Proceedings of the 11th IEEE International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012; pp. 1336–1340.

10. Tropp, J.; Gilbert, A.; Strauss, M. Algorithms for simultaneous sparse approximation. Part 1: Greedy pursuit. *Signal Process.* **2006**, *86*, 572–588. [[CrossRef](#)]
11. Determe, J.; Louveaux, J.; Jacques, L.; Horlin, F. On the exact recovery condition of simultaneous orthogonal matching pursuit. *IEEE Signal Process. Lett.* **2016**, *23*, 164–168. [[CrossRef](#)]
12. Quan, Y.; Li, Y.; Gao, X.; Xing, M. FPGA implementation of realtime compressive sensing with partial Fourier dictionary. *Int. J. Antennas Propag.* **2016**, *2016*, 1671687. [[CrossRef](#)]
13. Septimus, A.; Steinberg, R. Compressive sampling hardware reconstruction. In Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Paris, France, 30 May–2 June 2010; pp. 3316–3319.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Intentional Aliasing Method to Improve Sub-Nyquist Sampling System

Jehyuk Jang , Sanghun Im, and Heung-No Lee , *Senior Member, IEEE*

Abstract—A modulated wideband converter (MWC) has been introduced as a sub-Nyquist sampler that exploits a set of fast alternating pseudo random (PR) signals. Through parallel analog channels, an MWC compresses a multiband spectrum by mixing it with PR signals in the time domain, and acquires its sub-Nyquist samples. Previously, the ratio of compression was fully dependent on the specifications of PR signals. That is, to further reduce the sampling rate without information loss, faster and longer-period PR signals were needed. However, the implementation of such PR signal generators results in high power consumption and large fabrication area. In this paper, we propose a novel aliased modulated wideband converter (AMWC), which can further reduce the sampling rate of MWC with fixed PR signals. The main idea is to induce intentional signal aliasing at the analog-to-digital converter (ADC). In addition to the first spectral compression by the signal mixer, the intentional aliasing compresses the mixed spectrum once again. We demonstrate that AMWC reduces the number of analog channels and the rate of ADC for lossless sub-Nyquist sampling without needing to upgrade the speed or the period of PR signals. Conversely, for a given fixed number of analog channels and sampling rate, AMWC improves the performance of signal reconstruction.

Index Terms—Sub-Nyquist sampling, modulated wideband converter, sampling efficiency, intentional aliasing, compressed sensing, random filter.

I. INTRODUCTION

APPLICATIONS of electronic warfare (EW) systems, electronic intelligence (ELINT) systems, or cognitive radios are demanding the observation of a multiband signal, i.e., a collection of multiple narrow-band signals, each with different center frequencies, scattered across a wide frequency range up to tens of gigahertz (GHz). The Nyquist sampling rate is twice the maximum frequency of the wide range. When a multiband signal is *sparse*, i.e. consists of a few narrow bands, the signal can be sampled without information loss at a sub-Nyquist rate far less than the Nyquist rate. The theoretical lower limit of the rate required for lossless sub-Nyquist sampling is the sum of

the bandwidths, known as the Landau rate, when the spectral locations of all the narrow-band signals are known [1]. When spectral locations are unknown, the lower limit is doubled [2].

The modulated wideband converter (MWC) proposed by Mishali *et al.* [3] is a lossless sub-Nyquist sampler that aims at achieving the theoretical lower limit of sampling rate. Similar to other sub-Nyquist samplers proposed in [4]–[6], MWC exploits pseudo-random (PR) signals, which periodically output pulsed patterns. MWC has multiple analog channels, each of which consists of a PR signal generator, signal mixer, low-pass filter (LPF) for anti-aliasing, and low-rate analog-to-digital converter (ADC) in sequence. The system compresses a multiband spectrum through the mixing and LPF procedures, following which it samples at a sub-Nyquist rate. The reconstruction of the input multiband spectrum is guaranteed under some conditions of the compressed sensing (CS) theory [7]–[11]. With the help of CS reconstruction algorithms in [2], [12] developed for the MWCs, it has been proved that an MWC can achieve the theoretical lower limit of the lossless sub-Nyquist sampling rate.

However, to achieve the lower limit of the lossless sub-Nyquist sampling rate, the previously proposed MWC by Mishali *et al.* relied on a high-end PR signal generator, since it was the only spectral compressor. The ratio of spectral compression was fully dependent on the *oscillation speed* and *length of the pulsed patterns* within a single period of the PR signals. Specifically, to improve the compression ratio for a sparser multiband signal, PR signals with a greater pattern length were required. In addition, the oscillation speed should be faster than the Nyquist rate for a lossless compression. Unfortunately, increasing the pattern length of a PR signal generator with tens of GHz-range switching speed leads to difficult research problems in the field of chip engineering, such as high power consumption and large fabrication area due to the high chip speed [13], [14], which hinder the commercial availability of such a PR signal generator chip.

Recently, efforts to reduce the rate for lossless sub-Nyquist sampling with MWC closer to the theoretical lower limit without upgrading the PR signal generators have been made in [15], [16]. In [15], the authors proposed a method that channelizes the multiband spectrum into few orthogonal subbands before mixing with the PR signals. Since the channelized signals have a lower Nyquist rate than the original input, for a given oscillation speed and pattern length of PR signals, the method achieves a higher ratio of spectral compression. Although the method led to a further reduction of the lossless sub-Nyquist sampling rate, it requires additional hardware resources for the channelization,

Manuscript received September 28, 2017; revised February 7, 2018 and April 1, 2018; accepted April 1, 2018. Date of publication April 9, 2018; date of current version May 18, 2018. The associate editor coordinating the review of the manuscript and approving it for publication was Prof. Xin Wang. This work was supported by a grant-in-aid of HANWHA SYSTEMS. (Corresponding author: Heung-No Lee.)

J. Jang and H.-N. Lee are with the Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: jjh2014@gist.ac.kr; heungno@gist.ac.kr).

S. Im is with the Hanwha Systems, Seongnam 13524, South Korea (e-mail: sh.im@hanwha.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSP.2018.2824257

such as band-pass filters, local oscillators, and a greater number of independent PR signal generators proportional to the number of subbands. In [16], a method similar to that proposed in [15] was presented, in which the input signal was divided into in-phase (I) and quadrature (Q) channels before mixing it with PR signals. The lossless sub-Nyquist sampling rate can be reduced by the same principle as in [15], although the authors did not mention this point. However, the system also required additional hardware resources for the I-Q division.

In this paper, we propose an aliased MWC (AMWC), which reduces the lossless sub-Nyquist sampling rate for given practical PR signals. The main idea of AMWC is to break the anti-aliasing rule and induce *intentional aliasing* at the ADC of each spatial channel by setting the bandwidth of the prior LPF to be greater than the ADC sampling rate. In addition to the first spectral compression by the mixing and LPF procedures, this intentional aliasing leads to another spectral compression under a certain relation between the ADC sampling rate and bandwidth of the prior LPF. Through the two spectral compression procedures, the compression ratio is improved without faster or longer PR signals. Consequently, for a given and fixed PR signal generator, the lossless sub-Nyquist sampling rate of AMWC is closer to the lower limit than that of MWC.

The proposed AMWC achieves the same effect as in previous works [15], [16], i.e., reduction in the lossless sub-Nyquist sampling rate without upgrading the PR signal generators, and requires no additional hardware components. To our knowledge, AMWC is *novel* in that no study has thus far improved the sub-Nyquist sampling capability of MWC by improving the utilization efficiency of given hardware resources.

In [17], [18], variations of MWC similar to AMWC that include aliasing at the ADC have been investigated for analyzing *channel capacity*. Their main results indicate that suppressing non-active subbands before spectral compression minimizes the loss of information rate incurred by aliasing the noise spectrum. Interestingly, the authors of [18] introduced a rule for determining the sampling rate of each spatial channel similar to that of AMWC (see Section III-A for details). However, the rule was designed to make a fair comparison with other filterbank-based systems by flexibly controlling the bandwidth of subbands, rather than to exploit the aliasing at the ADC to reduce the lossless sub-Nyquist sampling rate. Additionally, according to our results, the rule in [18] is insufficient and aliasing at the ADC may lead to information loss.

Our main contribution is that the anti-aliasing rule of MWC is shown to be unnecessary for lossless sub-Nyquist sampling. We reveal a certain relationship between the ADC sampling rate and bandwidth of the prior LPF so that AMWC can avoid the loss of signal information during the additional spectral compression. We demonstrate that, for given oscillation speed and pattern length of PR signals, the sampling rate and analog channels of AMWC required for the reconstruction of a multiband signal are further reduced. For given sampling rate and number of analog channels, we show that the reconstruction performance of AMWC for a multiband signal with a given sparsity is improved.

Additionally, we show that the benefits from intentional aliasing can be further strengthened using a non-flat LPF. The

non-flat frequency response of LPF results in a different input-output relationship for each frequency component of the sub-Nyquist samples of AMWC. Simulation results show that the reduction of lossless sub-Nyquist sampling rate is boosted when the filter response is samples of a random distribution as the input-output relationships of different frequency components become independent.

The remainder of this paper is organized as follows. In Section II, we briefly introduce MWC with the anti-aliasing rule and then define the goal of this paper. In Section III, we propose AMWC and derive its input-output relationship. The relationship between the sampling rate of ADC and bandwidth of LPF to avoid information loss is also provided. In Section IV, a revised input-output relationship of AMWC corresponding to the use of a non-ideal LPF is provided. Simulation results are provided in Section V. Section VI concludes the paper.

II. BACKGROUND AND PROBLEM FORMULATION

The modulated wideband converter (MWC) is a sub-Nyquist sampling system for multiband signals. A signal $x(t)$ is a multiband signal if its spectrum $X(f)$ is composed of K_B disjoint continuous bands of maximum bandwidth B [2], [3]. We assume that the maximum frequency of a target multiband signal does not exceed f_{\max} , i.e., $X(f) = 0$ for $f \in \mathcal{F}_{NYQ}^c$, where $\mathcal{F}_{NYQ} \triangleq [-f_{\max}, f_{\max}]$, and \mathcal{F}_{NYQ}^c is the complementary set of \mathcal{F}_{NYQ} . We denote the Nyquist rate by $f_{NYQ} \triangleq 2f_{\max}$.

A. System Constitution and Parameters

MWC consists of M analog channels in parallel (see Fig. 1-(a)). Each channel consists of a PR signal generator, a mixer, an LPF, and an ADC in sequence. Each PR signal $p_i(t)$ for channel index i is T_p -periodic and outputs chips of an odd length L within a single period T_p . Each chip lasts for a chip duration $T_c = T_p L^{-1}$. We denote the chip speed by $f_c \triangleq T_c^{-1}$ and the repetition rate of the PR signal by $f_p \triangleq T_p^{-1}$. The LPF has a cut-off frequency $W_{LPF}/2$, where W_{LPF} denotes the bandwidth of the filter including the negative frequency. The LPF bandwidth is set to $W_{LPF} = qf_p$, where q is the channel-trading parameter, an odd positive integer. Finally, we denote the sampling rate, which is equal at every channel, by f_s . The total sampling rate is the sum of sampling rates of all channels, defined by $f_{s,total} \triangleq Mf_s$.

MWC first compresses the input multiband spectrum using PR signals. After that, nonzero subbands of the multiband spectrum are recovered by CS recovery algorithms. For the successful CS recovery, all spectral components within the Nyquist range \mathcal{F}_{NYQ} of each PR signal are needed to be independent, which requires a fast chip speed $f_c \geq f_{NYQ}$ [3]. Throughout this paper, we set $f_c = f_{NYQ}$.

B. Conventional Modulated Wideband Converters

In the original paper [3] by Mishali *et al.*, for lossless sub-Nyquist sampling, the ADC followed the *anti-aliasing* rule, i.e., $f_s \geq W_{LPF}$. This conventional rule has sufficed for lossless

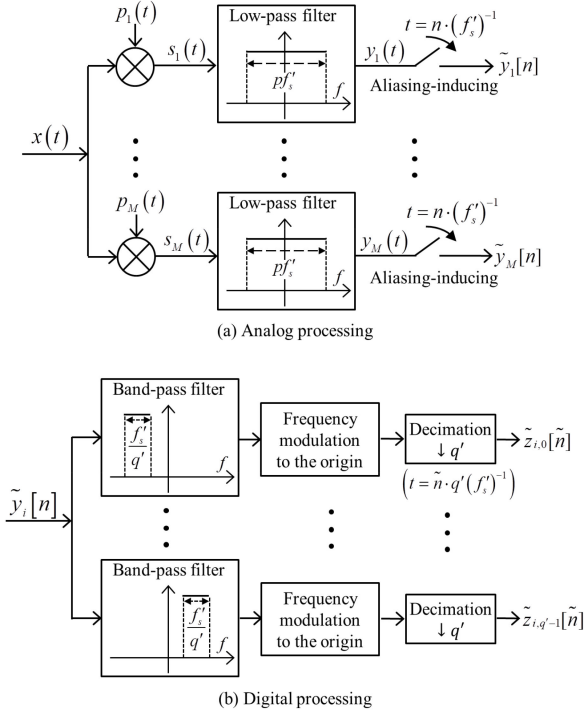


Fig. 1. Sampling system of AMWC. (a) Analog processing. (b) Digital processing. The system is equivalent to cMWC when $p = 1$ and $q' = q$. In AMWC, the sampling rate is p -times lower than the filter bandwidth with $p > 1$ to intentionally induce aliasing.

sub-Nyquist sampling. We refer to MWC that follows the anti-aliasing rule as *conventional MWC* (cMWC).

The input-output relationship of cMWC is given in [3]. The input $x(t)$ at the i -th channel is first mixed with the T_p -periodic PR signal $p_i(t)$ that periodically outputs a sequence of L mixing chips. By the periodicity, the Fourier transform (FT) of $p_i(t)$ is an impulse train. The FT of the mixed signal $s_i(t) = x(t)p_i(t)$ is the convolution $*$ of the two spectra:

$$\begin{aligned} S_i(f) &\triangleq \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \\ &= P_i(f) * X(f) \\ &= \sum_{l=-\infty}^{\infty} c_{i,l} X(f - lf_p), \end{aligned} \quad (1)$$

where $c_{i,l}$ for $l = -\infty, \dots, \infty$ are the Fourier series coefficients of $p_i(t)$. The mixed signal $s_i(t)$ and $X(f - lf_p)$ in (1) are filtered by the LPF $H(f)$. We let $H(f) = 1$ for $f \in \mathcal{F}_{LPP}$, and otherwise, $H(f) = 0$, where $\mathcal{F}_{LPP} \triangleq [-W_{LPP}/2, W_{LPP}/2]$. Since $X(f)$ is band-limited by \mathcal{F}_{NYQ} , the infinite-order summation in (1) is reduced to a finite order as follows:

$$\begin{aligned} Y_i(f) &= S_i(f) H(f) \\ &= \sum_{l=-(L_0+q_0)}^{L_0+q_0} c_{i,l} X(f - lf_p), \text{ for } f \in \mathcal{F}_{LPP}, \end{aligned} \quad (2)$$

where L_0 is computed by $L_0 = (L - 1)/2$ [3], and $q_0 \triangleq (q - 1)/2$. Next, the ADC of rate $f_s = T_s^{-1}$ takes samples of $y_i(t)$, i.e., $y_i[n] = y_i(t)|_{t=nT_s}$. By the conventional anti-aliasing rule, we set $f_s = W_{LPP}$. Then, the discrete-time FT (DTFT) of $y_i[n]$ preserves the spectrum of (2).

In (2), every subband $X(f - lf_p)$ is spectrally correlated with nearby $q - 1$ subbands, since the bandwidth W_{LPP} is wider than the shifting interval f_p . To make them spectrally orthogonal, the samples $y_i[n]$ are modulated and low-pass filtered in parallel through q digital channels by

$$z_{i,s}[\tilde{n}] = [(y_i[n] e^{-j2\pi s f_p T_s n}) * h_{f_p}[n]]|_{n=\tilde{n}q} \quad (3)$$

for $s = -q_0, \dots, q_0$, where $h_{f_p}[n]$ is a digital LPF with the cut-off frequency of $f_p/2$ and a flat passband response. The DTFT of (3) is

$$Z_{i,s}(e^{j2\pi f q T_s}) = \sum_{l=-L_0}^{L_0} c_{i,l+s} X(f - lf_p) \text{ for } f \in \mathcal{F}_p, \quad (4)$$

where $\mathcal{F}_p \triangleq [-f_p/2, f_p/2]$. The subbands $X(f - lf_p)$ in (4) are spectrally orthogonal to each other, since the bandwidth equals the shifting interval. As $X(f)$ is a multiband signal, only a few subbands in (4) have nonzero values. If $f_p \geq B$, the upper bound on the sparsity K of the subbands is $K \leq 2K_B$, since the uniform grid of interval f_p splits each band into two pieces at most.

Consequently, each analog channel outputs q different sequences, and therefore, cMWC obtains totally Mq equations for input reconstruction. Depending on the number of equations, it was shown in [3] that the input spectrum can be perfectly reconstructed. Previously, to obtain more equations for a fixed number of channels M and for a given specification f_p for PR signal generation, cMWC has to rely on the increased sampling rate $f_s = qf_p$ by controlling the *channel-trading* parameter q . In this paper, we aim to show there is another way to obtain more equations and improve the input reconstruction performance, without the cost intensive ways of increasing the total sampling rate $f_{s,\text{total}} = Mf_s$ or reducing f_p , or both.

C. Sampling Efficiency

In (4), MWC splits the input spectrum into many subbands along a uniform grid of a *splitting interval*, and it then takes samples of the weighted sum of subbands. We denote the splitting interval by f_I . Note that the splitting interval of cMWC $f_{I,\text{cMWC}}$ equals f_p . From the samples, a CS recovery algorithm (e.g., [11], [12], [19], [20]) finally recovers the K nonzero subbands containing the split pieces of the K_B multibands. Consequently, the total sampling rate is consumed to take samples of K nonzero subbands of bandwidth f_I . This indicates that the total sampling rate required for lossless sampling by an MWC would be at least $f_{s,\text{total}} \geq 2Kf_I$, where the factor of 2 arises from the unknown supports of the nonzero subbands. In contrast, a result in [2] states that, for a general sub-Nyquist sampling system, the minimum requirement for lossless sampling of a multiband signal is $f_{s,\text{total}} \geq 2K_B B$, where $K_B B$ is the upper bound of the *actual spectral occupancy* of a multiband signal. That is,

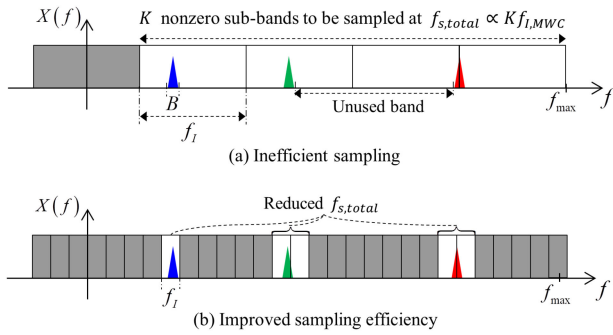


Fig. 2. Illustration of the *sampling efficiency* in the relation between the maximum bandwidth B and splitting interval f_I . (a) Inefficient sampling. (b) Improved sampling efficiency. (a) When $f_I \gg B$, MWC wastes a portion of the total sampling rate because of the unused band in the nonzero subbands. (b) The regulated f_I improves the sampling efficiency.

when f_I is far greater than B , MWC consumes a portion of the total sampling rate inefficiently. Specifically, f_I greater than B yields a higher probability for the K nonzero subbands to be comprised of unused bands, i.e., zeros. The inefficient use of total sampling rate is illustrated in Fig. 2.

Ideally, when the splitting interval f_I becomes finer and closer to B while satisfying $f_I \geq B$, the sampling efficiency is improved, as shown in Fig. 2. The efficiency is maximized when $K f_I = K_B B$. Based on this observation, we define the *sampling efficiency* α of MWC as the ratio between the actual spectral occupancy of the multiband signal and the total bandwidth of the recovered subbands, i.e.,

$$\alpha \triangleq \frac{K_B B}{K f_I}. \quad (5)$$

Note that, by the definition of K , $\alpha \leq 1$ always holds.

In summary, improving α has two advantages. First, for the lossless sampling of a given multiband signal, it would reduce the required total sampling rate $f_{s,total}$ closer to the theoretical minimum requirement $f_{s,total} \geq 2K_B B$. By the definition, the higher α closer to 1 indicates that a portion of $f_{s,total}$ inefficiently consumed for taking samples of the unused bands in Fig. 2 is reduced. By the reduced $f_{s,total}$, the number of channels M or the sampling rate f_s of ADC at each channel is reduced. Secondly, for given and fixed $f_{s,total}$, we will show throughout the rest of paper that improving α yields more independent equations for signal reconstruction, and thus, more complex multiband signals with higher K_B can be recovered perfectly.

D. Limitations in Conventional MWC

For cMWC, the sampling efficiency depends entirely on the hardware capabilities of PR signal generators, which may result in severe implementation problems. The sampling efficiency of cMWC depends on the specifications of PR signal generators since $f_{I,cMWC}$ is fixed to f_p . By the definition, the only way to improve the sampling efficiency α_{cMWC} of cMWC has been to make the repetition rate f_p of the PR signals closer to B . As discussed, the chip speed f_c of PR signals should not be less

than the Nyquist rate, i.e., $f_c \geq f_{NYQ}$. Thus, from the relation $f_p = f_c L^{-1}$, the chip length L is the only free parameter to control f_p . Since B is usually far smaller than f_{NYQ} , to fit f_p closer to B , a very long L is needed. However, in applications where f_{NYQ} reaches tens of gigahertz, due to the extremely high chip speed f_c , implementing PR signal generators having a high chip length L poses problems in terms of power consumption and fabrication area [13], [14]. Hence, other means to improve α without relying on the chip length L of the PR signals are very important.

For example, suppose one is observing on-air radar signals of bandwidth up to $B = 30$ [MHz] over an extremely wide observation frequency scope $f_{max} = 40$ [GHz]. This setting is reasonable in radar systems [21], [22]. We discussed that the chip speed should not be less than the Nyquist rate, i.e., $f_c \geq f_{NYQ}$, where $f_{NYQ} = 80$ [GHz]. In this example, to achieve $f_p \approx B$, the chip length needs to be $L = 2^{11} - 1$. Although hardware implementations of such PR signal generators having $f_c = 80$ [GHz] and chip length greater than $L = 2^{11} - 1$ were proposed in the literature [23], [24], they require very large fabrication areas and high power consumption, which has hindered practical uses thus far.

E. Problem Formulation

The goal of this paper is to introduce the proposed sampling system, Section III, which aims to improve the sampling efficiency α with given and fixed specifications f_p , f_c , and L for PR signal generation. Throughout this paper, we assume small L and B and a large $f_{NYQ} = f_c$, which implies f_p large enough compared to B and makes room for improving α . That is, $f_p \geq pB$ for a natural number $p > 1$. Then, improving α can be made without upgrading the PR signal generators and causing the said implementation issues such as higher power consumption and larger fabrication area discussed in the previous subsection. Thus, very wideband signals can be losslessly sampled using commercially available PR signal generators and ADCs, while this was not possible in the past with the conventional cMWC system.

III. ALIASED MWC

We propose *Aliased modulated wideband converters* (AMWC). AMWC renders the anti-aliasing rule $f_s \geq W_{LPF}$ used in cMWC unnecessary, as revealed later. Instead, AMWC intentionally induces and exploits aliasing at the ADC to regulate the splitting interval f_I and improve α without relying on the specification of PR signals.

In this section, we first discuss our method to induce controlled aliasing at the ADC and derive revised input-output relationships of AMWC. We then investigate how to control the aliasing for lossless sampling. Finally, we compare the sampling efficiency of AMWC with that of cMWC.

A. Intentional Aliasing Method

The AMWC system is depicted in Fig. 1. As mentioned already, compared to cMWC, AMWC is designed to not satisfy the

TABLE I
PARAMETER COMPARISONS BETWEEN AMWC AND cMWC

Multiband model		
$f_{\text{NBQ}}=18$ [GHz]	$B=30$ [MHz]	$K_B=10$
System specification		
$L=2^7-1$	$f_p=142$ [MHz]	$M=3$
Parameters	cMWC	AMWC (with $p=4$)
Channel-trading parameter	$q=5$	$q'=19$
Sampling rate [MHz]	$f_s=f_pq=710$	$f'_s=f_pq'p^{-1}=674.5$
Splitting interval [MHz]	$f_I=f_p=142$	$f_I=f_p p^{-1}=35.5$
Sparsity	$K \leq 2K_B=20$	$K \leq 2K_B=20$
Number of rows of \mathbf{X}	$N=L=127$	$N=Lp=508$
Total number of equations	$Mq=15$	$Mq'=57$

anti-aliasing rule at the ADC; rather, it is designed to induce *intentional* aliasing by setting the bandwidth of LPF greater than the sampling rate. In fact, in both cMWC and AMWC, an aliasing is introduced first by the mixer. The effect of this first aliasing is shown in (2), where the mixer shifts, gives weights, and has the signal spectrum $X(f)$ overlapped with shifted versions of itself at intervals of f_p . By the second aliasing at the ADC, the overlapped spectrum is aliased again at intervals of new sampling rate of AMWC f'_s , which is smaller than the filter bandwidth. By adjusting the relationship between f_p and f'_s , the splitting interval f_I , which is the interval at which $X(f)$ is split in the outputs of AMWC, is regulated.

Specifically, we set the new sampling rate f'_s of AMWC:

$$f'_s = \frac{q'}{p} f_p, \quad (6)$$

where q' is the new channel trading parameter for AMWC and an odd number. The bandwidth of LPF is $W_{LPF} = q' f_p$, and therefore, $W_{LPF} = p f'_s$ for the integer *aliasing parameter* $p > 1$. We will show that coprime p and q' with $q' > p$ is necessary for no information loss of $X(f)$. The new sampling rate induces additional aliasing and regulates the splitting interval f_I to improve the sampling efficiency. We let

$$f'_p \triangleq \frac{f_p}{p} \quad (7)$$

denote the *least common shifting interval* (LCS), which will become the splitting interval of AMWC, i.e., $f_{I,AMWC} = f'_p$.

With the introduction of new sampling rate f'_s in (6), it becomes easier to compare AMWC with cMWC. Specifically, with the sampling rate fixed, the number of equations for the input reconstruction obtained by cMWC and that by AMWC can be compared; with the number of equations fixed, the sampling rates for the two can be compared. For a given sampling rate $f'_s = q' f_p / p$, we will show in this section, the number of equations obtained by AMWC is Mq' . For a given sampling rate $f_s = q f_p$, from Section II-B, the number of equations obtained by cMWC is Mq . With the sampling rate fixed the same, i.e., $f_s = f'_s$, we note that $q' = qp$. This implies that AMWC has p -times more equations than that of cMWC. Table I presents an example of the increase in the number of equations of AMWC. With the number

of equations fixed, i.e., $Mq = Mq'$, on the other hand, AMWC requires p -times smaller sampling rate than cMWC does.

In [18], a variation of MWC using a sampling rate similar to (6) was considered, to analyze the noise factor incurred by the aliasing of subbands. There appear coprime relations between p and q' similar to that in this paper. However, the purpose of using coprime p and q' in [18] was completely different from that of this paper, i.e., they regulated the splitting interval of the subbands to make a fair comparison with other filterbank-based sampling systems with regard to the effect of noise. No relation between p and q' for lossless sampling and improving sampling efficiency was studied in [18].

To support intentional aliasing, AMWC requires an ADC with an operating bandwidth wider than its sampling rate. Such an ADC can be implemented by using a wideband track-and-hold amplifier (THA) developed by Hittite Corp. for the applications of EW and ELLINT in [25]. This THA has an 18 GHz bandwidth and can be integrated at the front end of commercially available ADCs of sampling rate up to 4 giga-samples per second.

To show that the AMWC obtains Mq' equations, we observe the input-output relationships of the aliased samples $\tilde{y}_i[n]$ in Fig. 1. Without loss of generality, we assume $q' = q$ and $f_s = p f'_s$. By the sampling theorem, the DTFT of $\tilde{y}_i[n]$ is the sum of shifts of $Y_i(f)$:

$$\begin{aligned} \tilde{Y}_i(e^{j2\pi f T'_s}) &= \sum_{r=-\infty}^{\infty} Y_i(f - r f'_s) \\ &= \sum_{r=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} c_{i,l} X(f - r f'_s - l f_p) H(f - r f'_s), \end{aligned} \quad (8)$$

where $T'_s \triangleq (f'_s)^{-1}$ and $Y_i(f)$ given in (2) is the spectrum of the output of the LPF $H(f)$. Within only a single period of $\tilde{Y}_i(e^{j2\pi f T'_s})$ in (8), i.e., $\mathcal{F}'_s(f_0) \triangleq [f_0, f_0 + f'_s)$ for any $f_0 \in \mathbb{R}$, because the bandwidth of $Y_i(f)$ is limited by the LPF $H(f)$, most of the shifts $Y_i(f - r f'_s)$ for sufficiently large $|r|$ are zeros. In other words, there exist (f_0, R_1, R_2) such that the infinite order of the outer summation in (8) is reduced to a finite order, i.e.,

$$\tilde{Y}_i(e^{j2\pi f T'_s}) = \sum_{r=R_1}^{R_2} \sum_{l=-\infty}^{\infty} c_{i,l} X(f - r f'_s - l f_p) H(f - r f'_s) \quad (9)$$

for $f \in \mathcal{F}'_s(f_0)$. Assuming $H(f) = 1$ for $f \in \mathcal{F}_{LPF}$, if f_0, R_1 , and R_2 satisfy the conditions of Lemma 1, the LPF responses in (9) are replaced with $H(f - r f'_s) = 1$ for $f \in \mathcal{F}'_s(f_0)$. Note that, when $p = 1$, i.e., no aliasing exists at the ADC, $R_1 = R_2$, which is equivalent to cMWC.

Lemma 1: Equation (9) is equivalent to (8) if f_0, R_1 , and R_2 with $R_1 < R_2 \in \mathbb{Z}$ satisfy

$$R_2 - R_1 = p - 1, \quad (10)$$

and

$$f_0 = \left(R_2 - \frac{p}{2}\right) f'_s. \quad (11)$$

Proof: See Appendix A.

We represent the shifting indices $rf'_s + lf'_p$ in (9) in terms of the LCS f'_p . Then,

$$\tilde{Y}_i \left(e^{j2\pi f T'_s} \right) = \sum_{r=R_1}^{R_1+p-1} \sum_{l=-\infty}^{\infty} c_{i,l} X \left(f - (rq' + lp) f'_p \right) \quad (12)$$

for $f \in \mathcal{F}'_s(f_0)$. To merge the inner and outer summations in (12), we use Lemma 2.

Lemma 2: If p and q' are coprime, the linear combination $rq' + lp$ for $r \in \mathcal{P} \triangleq \{R_1, \dots, R_1 + p - 1\}$ and $l \in \mathbb{Z}$ spans every integer.

Proof: We consider the following congruent relationship

$$k \equiv rq' \pmod{p}. \quad (13)$$

By modular arithmetic, if p and q' are coprime, there always exists one-to-one correspondence between r and k in the least residue system modulo p . Since $|\mathcal{P}| = p$, $rq' \pmod{p}$ for $r \in \mathcal{P}$ in (13) spans every number in the least residue system of modulo p . Hence, for $r \in \mathcal{P}$ and $l \in \mathbb{Z}$, $rq' + lp = k \pmod{p} + lp$ spans every integer. ■

By denoting $k = rq' + lp$ in (12), we have the equivalent relationship

$$\tilde{Y}_i \left(e^{j2\pi f T'_s} \right) = \sum_{k=-\infty}^{\infty} d_{i,k} (R_1, p, q') X \left(f - kf'_p \right) \quad (14)$$

for $f \in \mathcal{F}'_s(f_0)$, where $d_{i,k}(R_1, p, q')$ are the new sensing coefficients of AMWC. Proposition 3 provides the rule to obtain the coefficients $d_{i,k}$ from the Fourier coefficients $c_{i,l}$ of PR signals.

Proposition 3: For coprime p and q' , let us define

$$I(k; R_1, p, q') \triangleq \frac{1}{p} \left\{ k - q' \left[\left((q')^{-1} k - R_1 \right) \pmod{p} + R_1 \right] \right\}, \quad (15)$$

where $(q')^{-1} \pmod{p}$ is the multiplicative inverse of q' modulo p . Equation (14) is equivalent to (12) if

$$d_{i,k} (R_1, p, q') = c_{i, I(k; R_1, p, q')}. \quad (16)$$

Proof: See Appendix B.

In (14), the bandwidth of the subbands $X(f - kf'_p)$ for $f \in \mathcal{F}'_s(f_0)$ equals f'_s and is q' times wider than their shifting interval f'_p . Therefore, every subband is correlated with the closest $q' - 1$ subbands. By making these subbands spectrally orthogonal, the M relationships for $i = 1, \dots, M$ are expanded to Mq' equations to enhance the input reconstruction performance. A similar work was done for cMWC through (3) to (4), which further divides the observing frequency domain $\mathcal{F}'_s(f_0)$ (14) into q' tiny domains. Specifically, for $u = 0, \dots, q' - 1$, the u -th tiny frequency domain is defined by $\mathcal{F}'_p(f_0 + uf'_p)$, where

$$\mathcal{F}'_p(f_0) \triangleq [f_0, f_0 + f'_p]. \quad (17)$$

Then, the corresponding divided outputs have relationships

$$\begin{aligned} \tilde{Y}_i^{(u)} \left(e^{j2\pi f T'_p} \right) \\ = \sum_{k=-\infty}^{\infty} d_{i,k} (R_1, p, q') X \left(f - kf'_p \right) \text{ for } f \in \mathcal{F}'_p \left(f_0 + uf'_p \right), \end{aligned} \quad (18)$$

for $u = 0, \dots, q' - 1$. Finally, we define the output $\tilde{Z}_{i,u}(e^{j2\pi f T'_p})$ of AMWC as follows:

$$\begin{aligned} \tilde{Z}_{i,u} \left(e^{j2\pi f T'_p} \right) &\triangleq \tilde{Y}_i^{(u)} \left(e^{j2\pi f T'_p} \right) \Big|_{f=f+uf'_p} \\ &= \sum_{k=-\infty}^{\infty} d_{i,k+u} (R_1, p, q') X \left(f - kf'_p \right) \end{aligned} \quad (19)$$

for $f \in \mathcal{F}'_p(f_0)$. The final output $\tilde{z}_{i,u}[\tilde{n}]$ in the discrete-time domain can be obtained by performing digital frequency modulation and low-pass filtering on $\tilde{y}_i[n]$, as similarly done for cMWC in (3). The specific design of the digital processing system is shown in Fig. 1-(b).

Consequently, in (19), the input $X(f)$ is split into spectrally orthogonal subbands at intervals of f'_p . Therefore, the splitting interval of AMWC equals the LCS f'_p :

$$f_{I,AMWC} = f'_p \triangleq \frac{f_p}{p}, \quad (20)$$

which is p times lower than $f_{I,cMWC}$. By reducing the splitting interval by controlling the aliasing parameter p , the sampling efficiency of AMWC in (5) is improved. Fig. 3 illustrates how AMWC regulates the splitting interval and improves the sampling efficiency. In contrast, as discussed in Section II-D, regulating the splitting interval of cMWC requires a very costly solution of advanced PR signal generators with a larger chip length. Consequently, both cMWC and AMWC obtain $Mq = Mq'$ equations for input reconstruction, although AMWC consumes a p -times lower total sampling rate (6). In Section III-C, we show that the Mq' equations of AMWC are independent.

B. Matrix Form of Input–Output Relationship

For convenience of analyzing and solving linear simultaneous Mq' (19), we cast them as a matrix equation. To this end, we first reduce the infinite summation in (19) to be finite. We then discretize the continuous spectra to form a matrix with a finite number of columns.

Since $X(f)$ is band-limited to $f \in \mathcal{F}_{NYQ}$, within the limited frequency range $f \in \mathcal{F}'_p(f_0)$, the infinite summation order in (19) is reduced to a finite order as follow:

$$\begin{aligned} \tilde{Z}_{i,u} \left(e^{j2\pi f T'_p} \right) \\ = \sum_{k=N_1}^{N_2} d_{i,k+u} (R_1, p, q') X \left(f - kf_{I,AMWC} \right) \text{ for } f \in \mathcal{F}'_p(f_0), \end{aligned} \quad (21)$$

where N_1 and N_2 are, respectively, the smallest and largest index k of the subbands $X(f - kf_{I,AMWC})$ that contain some

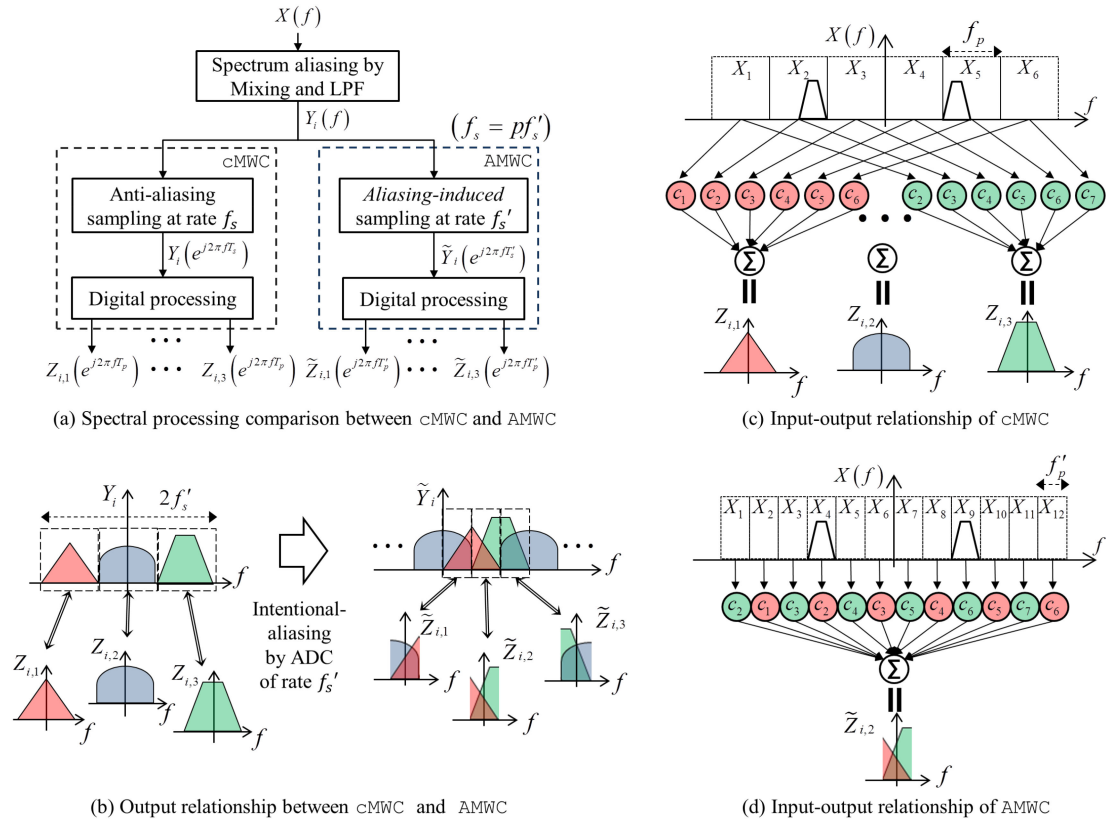


Fig. 3. Principle of improving the sampling efficiency by AMWC at a single analog channel is illustrated, with setting $q = 3$, $q' = q$, $p = 2$, and $M = 3$. (a) Spectral processing comparison between cMWC and AMWC. (b) Output relationship between cMWC and AMWC. (c) Input-output relationship of AMWC. At the first stage, the input spectrum $X(f)$ is aliased by mixing it with the PR signal and low-pass filtering it. This aliased-version of $X(f)$ is depicted as $Y_i(f)$. In (a), the main difference between cMWC and AMWC is how to take time-samples of $Y_i(f)$. cMWC prevents the spectrum from being aliased in taking time-samples. AMWC, on the contrary, aims to make the spectrum $Y_i(f)$ intentionally aliased once again, as depicted as $\tilde{Y}_i(f)$ in (b). In (c), as a result, the splitting-interval of cMWC is f_p , whereas in (d), that of AMWC is halved to f'_p . Thus, the sampling efficiency of AMWC becomes doubled (as $p = 2$).

active value of $X(f)$ within $f \in \mathcal{F}_{NYQ}$. Namely, these indices N_1 and N_2 indicate $X(f - kf_{I,AMWC}) = 0$ for $k < N_1$ and $k > N_2$, and thus help us obtain a matrix equation of (21) with finite dimensions. To mathematically define N_1 and N_2 , note that the k -th subband $X(f - kf_{I,AMWC})$ in (21) observes the frequency range

$$\mathcal{F}_k \triangleq [f_0 - kf_{I,AMWC}, f_0 - kf_{I,AMWC} + f'_p) \quad (22)$$

of $X(f)$. Then, the indices N_1 and N_2 are defined by

$$\begin{aligned} N_1 &\triangleq \min \{k \in \mathbb{Z} : \mathcal{F}_k \cap \mathcal{F}_{NYQ} \neq \emptyset\} \\ &= \min \{k \in \mathbb{Z} : f_0 - kf_{I,AMWC} < f_{\max}\} \end{aligned} \quad (23)$$

and

$$\begin{aligned} N_2 &\triangleq \max \{k \in \mathbb{Z} : \mathcal{F}_k \cap \mathcal{F}_{NYQ} \neq \emptyset\} \\ &= \max \{k \in \mathbb{Z} : f_0 - kf_{I,AMWC} + f'_p > -f_{\max}\}, \end{aligned} \quad (24)$$

respectively. Using the parameters and relations given in Table II and Lemma 1, the two problems (23) and (24) turn into

$$N_1 = \min \left\{ k \in \mathbb{Z} : R_2 q' - \frac{(q' + L)p}{2} < k \right\} \quad (25)$$

and

$$N_2 = \max \left\{ k \in \mathbb{Z} : R_2 q' - \frac{(q' - L)p}{2} + 1 > k \right\} \quad (26)$$

respectively. As both q' and L are odd positive integers, the solutions of two problems (25) and (26) are determined as follow:

$$N_1 = R_2 q' - \frac{(q' + L)p}{2} + 1, \quad (27)$$

and

$$N_2 = R_2 q' - \frac{(q' - L)p}{2}. \quad (28)$$

Finally, the output spectrum $\tilde{Z}_{i,u}(e^{j2\pi f T'_p})$ in (21) turns into a linear combination of unknown subbands $X(f - kf_{I,AMWC})$ for $f \in \mathcal{F}'_p(f_0)$. The matrix-multiplication form $\mathbf{Z} = \mathbf{D}\mathbf{X}$ of (21) is provided in (30) as shown at the bottom of the next page. We denote the number of subbands, i.e., the dimension of matrix \mathbf{X} , by N , which equals

$$\begin{aligned} N &= N_2 - N_1 + 1 \\ &= Lp. \end{aligned} \quad (29)$$

TABLE II
SUMMARY OF AMWC PARAMETERS (CMWC WHEN $p = 1$ AND $q' = q$)

f_{\max}	maximum frequency of multiband signal
f_{NYQ}	Nyquist rate of multiband signal, $f_{\text{NYQ}} \triangleq 2f_{\max}$
B, K_B	maximum bandwidth and number of the narrow bands in a multiband signal
K	number of nonzero subbands (sparsity), $K \leq 2K_B$ if $f'_p \geq B$.
M	number of analog channels
L	length of PR chips within a single period
f_c	chip speed of PR signals, $f_c = f_{\text{NYQ}}$
f_p	repetition rate of PR signals, $f_p = f_c L^{-1}$
q', P	channel-trading parameter, aliasing parameter
W_{LPF}	bandwidth of LPF, $W_{\text{LPF}} = q' f_p$
f'_p	least common shifting interval, $f'_p \triangleq f_p p^{-1} \geq B$
f'_s	sampling rate of an ADC, $f'_s = W_{\text{LPF}} p^{-1}$
$f_{s,\text{total}}$	total sampling rate, $f_{s,\text{total}} \triangleq M f'_s$
$f_{I,\text{AMWC}}$	splitting interval, $f_{I,\text{AMWC}} = f'_p$
α_{AMWC}	sampling efficiency, $\alpha_{\text{AMWC}} \triangleq \frac{K_B B}{K f_{I,\text{AMWC}}}$

Since $X(f)$ consists of K_B narrow bands over the wide Nyquist range, only a few of its subbands $X(f - k f_{I,\text{AMWC}})$ for $f \in \mathcal{F}'_p(f_0)$ have nonzero values. Therefore, the matrix \mathbf{X} in (30) is row-wise sparse with a sparsity K related to K_B .

To draw a relationship between the analytic result (30) and actually acquired samples $\tilde{z}_{i,u}[\tilde{n}]$, we convert the DTFT (30) to the DFT of $\tilde{z}_{i,u}[\tilde{n}]$ by taking the frequency samples of the infinite columns of \mathbf{Z} and \mathbf{X} . When the input is observed for a finite duration T_o , taking samples of the spectrum (21) at frequency intervals of $\Delta f = T_o^{-1}$ does not cause any information loss. The samples of spectrum $\tilde{Z}_{i,u}(e^{j2\pi f T'_p})$ is obtained by taking the DFT of the actually acquired time-samples $z_{i,u}[\tilde{n}]$. Consequently, for a finite observation time $T_o = 2W T'_p$ for a sample length $2W$, we rewrite the matrix-multiplication form (30) as

$$\mathbf{Z}_{2W} = \mathbf{D}\mathbf{X}_{2W}, \quad (31)$$

where columns of $\mathbf{Z}_{2W} \in \mathbb{C}^{Mq' \times 2W}$ and $\mathbf{X}_{2W} \in \mathbb{C}^{N \times 2W}$ are sub-columns of \mathbf{Z} and \mathbf{X} , respectively, at frequency intervals of Δf . This concept will be exploited in Section IV to derive a revised input-output relationship of AMWC for using LPF with a non-flat frequency response.

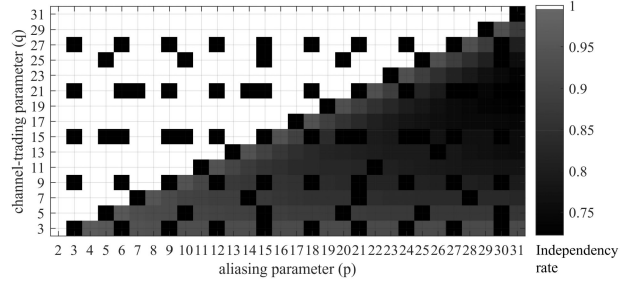


Fig. 4. Independency rates under various p and q' for which randomly selected Mq' columns of the sensing matrix $\mathbf{D} \in \mathbb{C}^{Mq' \times N}$ of AMWC are independent. When p and q' are coprime and $q' > p$, every selection of Mq' columns is linearly independent.

C. Choosing the Aliasing Parameter

For a given total sampling rate, AMWC obtains more equations used for input reconstruction than CMWC does. What remains is to check if the extended equations provide independent information. We reveal a condition on the aliasing parameter p that necessitates the linear system (30) to be well-posed for every K -sparse signal matrix \mathbf{X} .

Proposition 4: There exists the unique solution of (30) for every K -sparse signal \mathbf{X} only if p and q' are coprime and $q' > p$.

Proof: See Appendix C.

Proposition 4 gives a condition $q' < p$ for coprime p and q' that makes AMWC an ill-posed system. This indicates that, within the set of coprime $q' > p$, there may be a subset that makes AMWC guarantees the existence of unique solution of (30) for every K -sparse signal matrix \mathbf{X} .

In [11], a CS result states there exist the unique solution of a multiple measurement vector (MMV) CS equation $\mathbf{Z} = \mathbf{D}\mathbf{X}$ for every K -sparse signal \mathbf{X} if

$$2K < \text{spark}(\mathbf{D}) - 1 + \text{rank}(\mathbf{X}), \quad (32)$$

where *spark* is the minimum number of linearly dependent columns in \mathbf{D} . Meanwhile, the spark of an Mq' -by- N matrix is upper bounded to $Mq' + 1$ by the Singleton bound [26]. Based on these results, we find a sufficient condition on p and q' from Monte Carlo experiments in Section V-A (Fig. 4) that maximizes the spark of \mathbf{D} .

Main Result 5. Let $Mq' \geq 2K$. For every K -sparse signal \mathbf{X} , there exists the unique solution of (30), and therefore, AMWC

$$\underbrace{\begin{pmatrix} Z_{1,0} (e^{j2\pi f T'_p}) \\ \vdots \\ Z_{1,q'-1} (e^{j2\pi f T'_p}) \\ Z_{2,0} (e^{j2\pi f T'_p}) \\ \vdots \\ Z_{M,q'-1} (e^{j2\pi f T'_p}) \end{pmatrix}}_{\triangleq \mathbf{Z} \in \mathbb{C}^{Mq' \times \infty}} = \underbrace{\begin{pmatrix} d_{1,N_1} & d_{1,N_1+1} & \cdots & d_{1,N_2} \\ \vdots & \vdots & & \vdots \\ d_{1,N_1+(q'-1)} & d_{1,N_1+(q'-1)+1} & \cdots & d_{1,N_2+(q'-1)} \\ d_{2,N_1} & d_{2,N_1+1} & \cdots & d_{2,N_2} \\ \vdots & \vdots & & \vdots \\ d_{M,N_1+(q'-1)} & d_{M,N_1+(q'-1)+1} & \cdots & d_{M,N_2+(q'-1)} \end{pmatrix}}_{\triangleq \mathbf{D} \in \mathbb{C}^{Mq' \times N}} \underbrace{\begin{pmatrix} X(f - N_1 f'_p) \\ X(f - (N_1 + 1) f'_p) \\ \vdots \\ X(f - N_2 f'_p) \end{pmatrix}}_{\triangleq \mathbf{X} \in \mathbb{C}^{N \times \infty}} \quad (30)$$

does not lose any information of K -sparse signal \mathbf{X} , if p and q' are coprime and $q' > p$.

Meanwhile, we choose p to minimize the maximum of the sparsity K , which is the number of nonzero subbands of $X(f)$ at splitting intervals $f_{I,AMWC} = f'_p$. The sparsity K is dependent on the center frequencies of K_B multibands and their maximum bandwidth B . When $f_{I,AMWC} \geq B$, every multiband occupies at most two subbands, which implies $K \leq 2K_B$. On the other hand, when $f_{I,AMWC} < B$, some multibands may occupy more than two subbands, which provides an opportunity to increase K beyond $2K_B$. Hence, we recommend choosing the aliasing parameter p as

$$p \leq \left\lfloor \frac{f_p}{B} \right\rfloor. \quad (33)$$

D. Sampling Efficiency Analysis

We compare the sampling efficiencies of AMWC, α_{AMWC} , and cMWC, α_{cMWC} , defined in (5). The sampling efficiencies are functions of the sparsity K , which is a random variable in general. We denote the sparsity of cMWC and AMWC by K_{cMWC} and K_{AMWC} , respectively. To make them deterministic, we put assumptions on K_{cMWC} and K_{AMWC} that in both cMWC and AMWC, the K_B bands in $X(f)$ respectively occupies exactly one subband, i.e., $K_{cMWC} = K_{AMWC} = K_B$. This occurs with high probability when $f_p p^{-1} \gg B$ and the center frequencies of multibands are far enough apart from each other with a small K_B .

Under the assumption above, the sampling efficiencies of cMWC and AMWC are obtained by

$$\alpha_{cMWC} = \frac{K_B B}{K_{cMWC} f_{I,cMWC}} = \frac{B}{f_p}, \quad (34)$$

and

$$\alpha_{AMWC} = \frac{K_B B}{K_{AMWC} f_{I,AMWC}} = \frac{pB}{f_p}, \quad (35)$$

respectively. Note that if $p = 1$, AMWC and cMWC are completely identical, and therefore $\alpha_{AMWC} = \alpha_{cMWC}$. When $p > 1$, the intentional aliasing of AMWC takes effect and improves the sampling efficiency proportionally to p .

IV. NON-IDEAL LOW-PASS FILTERS

The input-output relationship in the previous section is based on the ideal LPF $H(f)$ having a flat pass-band response. However, in real applications, the pass-band response of an LPF significantly fluctuates. In the case of cMWC, a post digital-processing technique to equalize the effects of non-flat filter responses was proposed in [27]. Unfortunately, owing to the aliasing at ADC, the equalizations cannot be applied to AMWC. In this section, we instead provide a revised input-output relationship of AMWC based on the fluctuated LPF $G(f)$. Without loss of generality, we assume all analog channels use the same LPF. We assume that the response $G(f)$ is nonzero and known within the pass-band $f \in \mathcal{F}_{LPF}$ and is zero for $f \in \mathcal{F}_{LPF}^C$. We derive a revised input-output relationship reflecting the

effect of $G(f)$. Paradoxically, our empirical results in Section V conclude that, for a given sampling efficiency, an irregularly fluctuated filter response is helpful to further decrease the total sampling rate required for lossless sub-Nyquist sampling.

The derivation starts from substituting $H(f)$ in the input-output relations of (8)–(12) with $G(f)$. Without loss of generality, we assume $q' = q$ and $f_s = p f'_s$. Equation (9) then turns into

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{r=R_1}^{R_2} \sum_{l=-\infty}^{\infty} c_{i,l} X(f - (rq' + lp) f'_p) G(f - rq' f'_p) \end{aligned} \quad (36)$$

for $f \in \mathcal{F}'_s(f_0)$, where R_1 and R_2 are chosen from Lemma 1. By Lemma 2, we substitute $rq' + lp = k$ and merge the outer and inner summations:

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{k=N_1}^{N_2} d_{i,k}(R_1, p, q') X(f - k f'_p) G(f - \gamma_p(k) f'_p) \end{aligned} \quad (37)$$

for $f \in \mathcal{F}'_s(f_0)$, where the sensing coefficients $d_{i,k}(R_1, p, q')$, N_1 , and N_2 are, respectively, computed from Proposition 3, (27), and (28). We define the function γ_p of k that maps k in (37) to the corresponding rq' in (36) so that the two equations are equivalent. Lemma 6 reveals the mapping rule for $\gamma_p(k)$.

Lemma 6: Under the conditions of Lemma 1 and Lemma 2, (36) and (37) are equivalent if the mapping rule of γ_p is assigned by

$$\gamma_p(k) = k - pI(k; R_1, p, q'), \quad (38)$$

where the picking regularity $I(k; R_1, p, q')$ is defined in (15).

Proof: See Appendix B.

As done in (14) to (19), the final outputs $\tilde{z}_{i,u}[\tilde{n}]$ for $u = 0, \dots, q' - 1$ are obtained by processing the time-samples $y_i[n]$ of the spectrum (37) using the digital system given in Fig. 1-(b). Then, those spectra $Z_{i,u}(e^{j2\pi f T'_p})$ have the following input-output relationships:

$$\begin{aligned} \tilde{Z}_{i,u} \left(e^{j2\pi f T'_p} \right) &= \sum_{k=N_1}^{N_2} d_{i,k+u}(R_1, p, q') X(f - k f'_p) G(f + u f'_p - \gamma_p(k+u) f'_p) \\ &= \sum_{k=N_1}^{N_2} d_{i,k+u}(R_1, p, q') G(f - \gamma'_p(k, u) f'_p) X(f - k f'_p) \end{aligned} \quad (39)$$

for $f \in \mathcal{F}'_p(f_0)$, where $\gamma'_p(k, u) \triangleq \gamma_p(k+u) - u$.

Consequently, the linear coefficients on the subbands $X(f - k f'_p)$ in (39) become frequency-selective. To numerically solve (39), we discretize the continuous frequency, as

discussed in Section III-B. We assume that the signal is observed for the finite duration $T_o = 2WT'_p$, where $2W$ is the length of the discretized signal. Then, the samples of spectrum are defined by

$$\begin{aligned} \tilde{Z}_{i,u}[w] &\triangleq \tilde{Z}_{i,u} \left(e^{j2\pi f T'_p} \right) \Big|_{f=wT_o^{-1}} \\ &= \sum_{k=N_1}^{N_2} d_{i,k+u} [G(f - \gamma'(k, u) f'_p) X(f - kf'_p)]_{f=wT_o^{-1}} \\ &= \sum_{k=N_1}^{N_2} b_{(i,u),k}[w] X(f - kf'_p) \Big|_{f=wT_o^{-1}} \end{aligned} \quad (40)$$

for $w \in \mathcal{W} \triangleq \{f_0 T_o, \dots, (f_0 + f'_p) T_o - 1\}$, where the frequency-selective sensing coefficients $b_{(i,u),k}[w]$ are defined as

$$b_{(i,u),k}[w] \triangleq d_{i,k+u} G(f - \gamma'(k, u) f'_p) \Big|_{f=wT_o^{-1}} \quad (41)$$

for $w \in \mathcal{W}$. Note that, by the relation between DFT and DTFT, the spectrum samples (40) are obtained by taking the DFT as follows:

$$Z_{i,u}[w] = \sum_{\tilde{n}=0}^{2W-1} z_{i,u}[\tilde{n}] e^{j2\pi \frac{\tilde{n}}{2W} (w \bmod 2W)} \text{ for } w \in \mathcal{W}, \quad (42)$$

where $z_{i,u}[\tilde{n}]$ are the output sequences of AMWC.

For convenience, we represent the input-output relation of (40) for $w \in \mathcal{W}$ in a vector form as

$$\mathbf{Z}[w] = \mathbf{B}[w] \mathbf{X}[w], \quad (43)$$

where the elements of the output column vector $\mathbf{Z}[w] \in \mathbb{C}^{Mq'}$ are $Z_{i,u}[w]$ for row indices $i = 1, \dots, M$ and $u = 0, \dots, q' - 1$. The unknown column vector $\mathbf{X}[w] \in \mathbb{C}^N$ consists of $X(f - kf'_p) \Big|_{f=wT_o^{-1}}$ for row indices $k = N_1, \dots, N_2$. The frequency-selective sensing matrix $\mathbf{B}[w] \in \mathbb{C}^{Mq' \times N}$ consists of $b_{(i,u),k}[w]$ with row indices i and u and column index k . The CS model (43) is called MMV with different sensing matrices, for which many numerical solvers have been developed [9], [28].

The existence of unique solution of (43) depends on the spark of sensing matrix $\mathbf{B}[w]$. Note that from (41), the elements of $\mathbf{B}[w]$ are multiplications of the elements of \mathbf{D} and the samples of the low pass filter $G(f)$. In [29], Davies *et al.* proved that the spark of a matrix from an independent continuous distribution achieves the Singleton bound with probability one. When the filter response $G(f)$ is designed to be irregular, i.e., its samples are drawn from an independent random distribution, the spark of $\mathbf{B}[w]$ after multiplication with the samples of $G(f)$ should grow closer to achieving the Singleton bound. When the spark of $\mathbf{B}[w]$ indeed achieves the Singleton bound and the condition (32) holds, for every K -sparse signal \mathbf{X} the unique solution to (43) always exists.

V. SIMULATION

A. Spark of Sensing Matrix

To support Main Result 5, the sufficiency of lossless sub-Nyquist sampling by AMWC, we demonstrate that the sensing matrix \mathbf{D} with coprime parameters $q' > p$ achieves the Singleton bound.

Monte Carlo experiments were performed under various settings of p and q' . With $L = 127$, we used the maximum length sequences of length L as the chip values of PR signal for each channel $i = 1, \dots, M$. We set the number of analog channels to $M = 3$. For 5×10^5 independent trials, we randomly selected Mq' columns of \mathbf{D} and counted the rate for which the selected columns are linearly independent.

Fig. 4 shows how the linear independency of columns in \mathbf{D} varies as p and q' change. The white points in the plot indicate the pairs of p and q' where every selection of Mq' columns of \mathbf{D} is linearly independent. The dark points indicate that at least one selection of Mq' columns has linear dependency. The upper triangular area indicates the region of (p, q') with $q' > p$ where all points except for the points that p and q' are not coprime belong to the white set. That is, for coprime $q' > p$, all the selections of Mq' columns are linearly independent, and thus the spark of \mathbf{D} achieves the Singleton bound. This result is consistent with Proposition 4 and supports Main Result 5.

B. Reduction of Total Sampling Rate

We demonstrate that, with the improved sampling efficiency, AMWC indeed reduces the total sampling rate required for lossless sub-Nyquist sampling for given specifications of PR signals. Additionally, when the frequency response of low-pass filters is drawn at random, the reduction of total sampling rate is boosted. The reduction of total sampling rate reduces the number of channels as well as the sampling rate of each channel.

For simulation, we generated real-valued multiband inputs $x(t)$ as the sum of K_B narrow band signals of bandwidth $B = 5$ [MHz]. The energies of narrow bands are equal. The center frequencies of narrow band signals were drawn at random, while those spectra were not overlapped with each other. The maximum frequency of $x(t)$ does not exceed $f_{\max} = 10$ [GHz]. The signals last for the duration $T_o = 2WT'_p$ seconds with $W = 15$. The parameters of PR signals were $L = 127$, $f_p = 2f_{\max} L^{-1} \simeq 157.48$ [MHz]. We used maximum length sequences with different initial seeds as the chip values of PR signals for channel indices $i = 1, \dots, M$. We expressed the continuous signals in simulation on a dense discrete-time grid with intervals of $(2q' f_{NYQ})^{-1}$ seconds. The bandwidth of low-pass filters and the sampling rate followed the parameter relations of AMWC, i.e., $W_{LPF} = q' f_p$ and $f'_s = p^{-1} W_{LPF}$. We considered the ideal LPF $H(f)$ with a flat passband response and the non-ideal LPF $G(f)$ with an irregular passband response. In simulation, the impulse response of $G(f)$ was drawn initially from the normal distribution, windowed to limit the filter bandwidth, and then held fixed throughout the whole simulation. We call $G(f)$ the random LPF with this irregular passband response. Under various settings of p , q' , and K_B with coprime $q' > p$, we

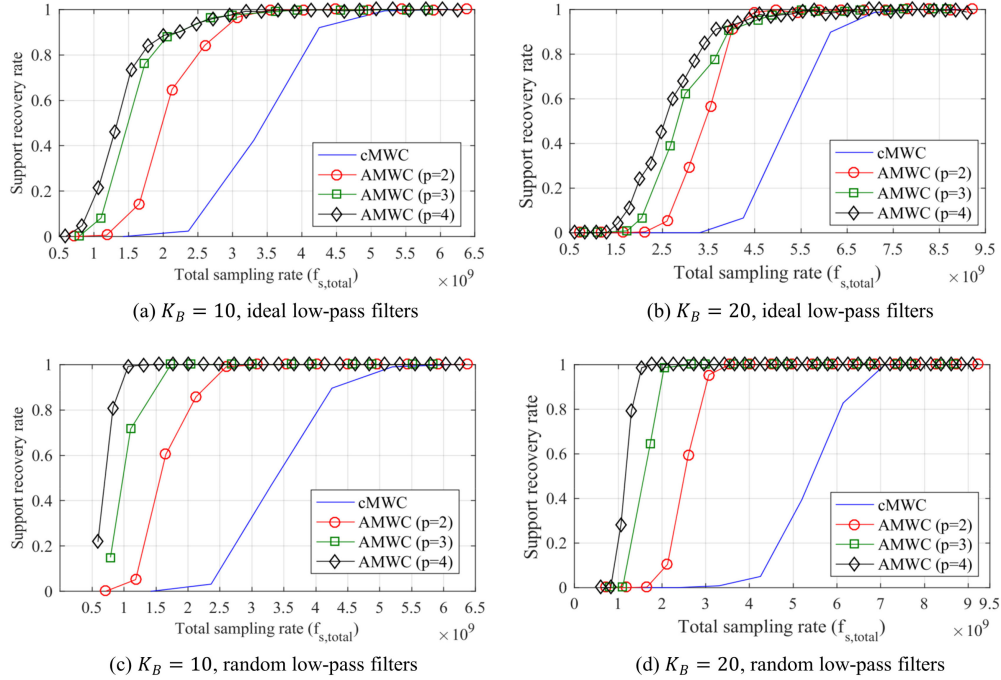


Fig. 5. Rate of successful support recovery of cMWC and AMWC as a function of total sampling rate for various aliasing parameters p and multibands K_B . The number of channels was fixed to $M = 3$. Ideal ((a)-(b)) and random ((c)-(d)) low-pass filters were used. (a) $K_B = 10$, ideal low-pass filters. (b) $K_B = 20$, ideal low-pass filters. (c) $K_B = 10$, random, low-pass filters. (d) $K_B = 20$, random low-pass filters.

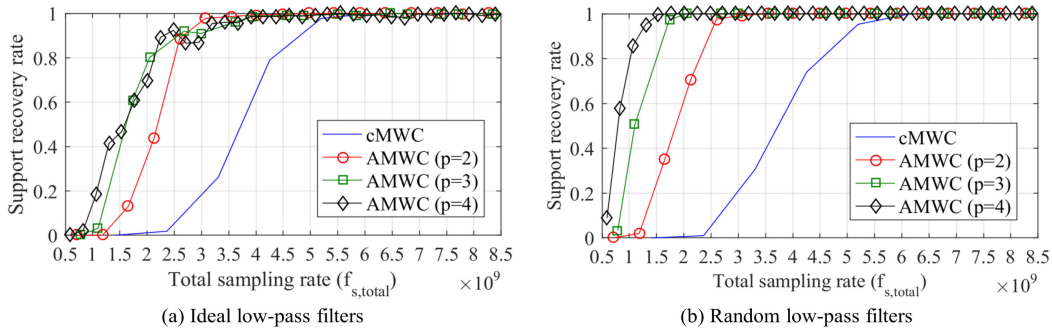


Fig. 6. Rate of successful support recovery of cMWC and AMWC as a function of total sampling rate when SNR=3 [dB]. The number of channels was fixed to $M = 3$, and the number of multibands in $X(f)$ is fixed to $K_B = 10$. (a) Ideal low-pass filters. (b) Random low-pass filters.

measured the rate of successful recovery of the supports of \mathbf{X} by the distributed CS orthogonal matching pursuit (DCS-SOMP) algorithm [28]. For single supports estimation, DCS-SOMP was run for $2K_B$ iterations. It aimed to find one distinct support per each iteration out of K supports, given $K \leq 2K_B$. Once the supports are found, x can be reconstructed by the least squares. The successful support recovery was declared if $\mathcal{S} \subseteq \hat{\mathcal{S}}$, where \mathcal{S} and $\hat{\mathcal{S}}$ are, respectively, the true and found supports. The support recovery rate in simulations was defined as the number of successful support recovery divided by total 500 trials with randomly regenerated $x(t)$.

Fig. 5 shows the support recovery rate of AMWC as a function of total sampling rate when $M = 3$. We set $K_B = \{10, 20\}$. Plots (a) and (b) are results of using the ideal LPF $H(f)$. It is demonstrate that compared to cMWC, AMWC reduces the total sampling rate required for reconstruction of given multiband signals. Inversely, for a given total sampling rate, AMWC takes

TABLE III
THE TOTAL SAMPLING RATE REQUIRED FOR 90% SUPPORT RECOVERY RATE WITH VARIOUS SNR AND VALUES OF p

SNR [dB]	LPF	$p=1$ (cMWC)	$p=2$ (AMWC)	$p=3$ (AMWC)	$p=4$ (AMWC)
-6	Ideal	6.142	4.016	3.622	3.898
	Random	6.142	3.543	2.677	2.244
-3	Ideal	6.142	3.543	3.622	3.425
	Random	6.142	3.071	2.047	1.535
0	Ideal	5.197	3.071	2.677	2.953
	Random	5.197	2.598	1.732	1.535
3	Ideal	5.197	3.071	2.677	2.480
	Random	5.197	2.598	1.732	1.299
12	Ideal	5.197	3.071	2.677	2.244
	Random	5.197	2.126	1.732	1.063

The floating numbers in cells indicate the minimal total sampling rate in GHz which achieves the support rate recovery of 90%. The number of analog channels and multibands were set to $M = 3$ and $K_B = 10$, respectively.

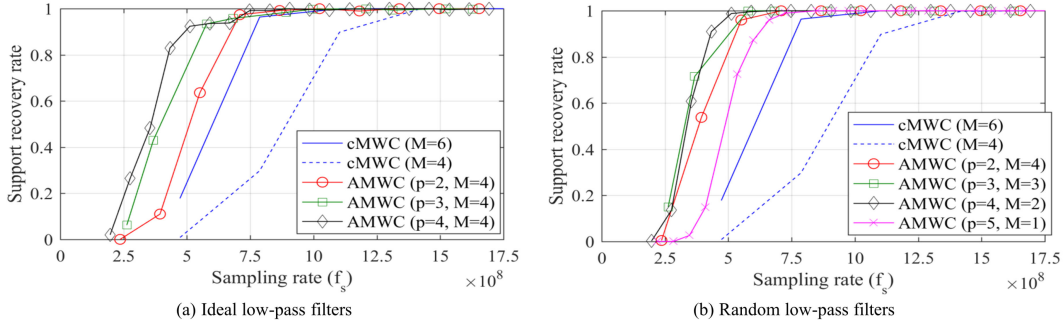


Fig. 7. Rate of successful support recovery of cMWC and AMWC as a function of sampling rate of each channel for various aliasing parameters p and the number of channels M . The number of multibands was fixed to $K_B = 10$. (a) Ideal low-pass filters. (b) Random low-pass filters.

sub-Nyquist samples of more multibands than cMWC does, without information loss.

However, when p increases, although the sampling efficiency is improved proportionally to p from (35), the total sampling rate does not decrease anymore. This is caused by the lack of degrees of freedom in the sensing matrix \mathbf{D} . The elements of \mathbf{D} are made of the Fourier coefficients $c_{i,l}$ of the PR signals, and most elements are repeatedly reused. Although it was demonstrated in the previous sub-section that \mathbf{D} has the maximum spark and well preserves the sparse signal \mathbf{X} , recovering \mathbf{X} by non-optimal CS algorithms requires \mathbf{D} to have a large degrees of freedom [10]. This limitation is overcome by using the random LPF $G(f)$.

Plots (c) and (d) are the results of using the random LPF $G(f)$. It is shown that AMWC further reduces the total sampling rate required for successful support recovery as the sampling efficiency improves. Consequently, the random response of $G(f)$ enhances the degrees of freedom of sensing matrices $\mathbf{B}[w]$ for different frequency indices w and improves the recovery performance by the non-optimal algorithm DCS-SOMP. This enhancement cannot be applied for cMWC, since the effect of random response becomes removable by equalization [27].

In Fig. 6, additive white Gaussian noise $n(t)$ of SNR = 3 [dB] was considered, where the signal-to-ratio noise (SNR) in decibel is defined as $\text{SNR} \triangleq 10 \log_{10}(\|x\|^2/\|n\|^2)$. We fixed $K_B = 10$. Plots (a) and (b) are the results for using the ideal LPF and the random LPF, respectively. Despite the additive noise, the results show that AMWC still reduces the total sampling rate or improves the recovery performance. Including the results in Fig. 6, we conducted more simulations under various SNR = $\{-6, -3, 0, 3, 12\}$ [dB] but omitted to repeat the plots as the graphs exhibit the similar pattern. Instead, we summarized the minimal sampling point results in Table III, where the minimal sampling point is defined as the minimal total sampling rate which achieves the support recovery rate of 90%. In the results, as p and/or SNR increase, the minimal sampling point gets smaller, which is expected.

Fig. 7 demonstrates that AMWC reduces the number of channels required for the support recovery. We set $K_B = 10$ and compared the support recovery rates of cMWC and AMWC for various M and given sampling rate of each channel. In plot (a), the support recovery rate of AMWC slightly outperforms cMWC, although AMWC uses fewer channels with a lower sampling rate of each channel than cMWC. Additionally, in plot (b), when the

random low-pass filter is used, AMWC using a single channel outperforms cMWC using six channels.

As the increase in the number of rows in \mathbf{Z} in (30) or in (43) by p -times, the performance of AMWC is improved but the computational complexity (CC) for the support recovery with AMWC inevitably increases as well. The CC of a compressed sensing algorithm depends on the sizes of matrices in the linear inverse problem $\mathbf{Z} = \mathbf{D}\mathbf{X}$. Let Q_{equation} , Q_{sample} , and Q_{subband} denote the number of rows and columns of \mathbf{Z} and the number of rows of \mathbf{X} for cMWC problem, respectively. We make note of the report that the CC of DCS-SOMP with cMWC is $O(Q_{\text{equation}}^2 Q_{\text{subband}} Q_{\text{sample}})$ [28]. When the two total sampling rates $f_{s,\text{total}}$ of cMWC and $f'_{s,\text{total}}$ of AMWC are equal to each other, the number of rows of \mathbf{Z} of AMWC becomes pQ_{equation} and that of \mathbf{X} becomes pQ_{subband} , respectively, as discussed in Section III-A. In addition, since the bandwidth of the subbands of AMWC is p -times narrower than that of cMWC, the number of columns of \mathbf{Z} becomes $p^{-1}Q_{\text{sample}}$. Thus, the CC of DCS-SOMP with AMWC is $O(p^2 Q_{\text{equation}}^2 Q_{\text{subband}} Q_{\text{sample}})$.

VI. CONCLUSION

We proposed a new MWC system called AMWC which improves the sampling efficiency by intentionally inducing an aliasing at the ADC. We showed that the improved sampling efficiency leads to reduction on the sampling rate and number of channels required for obtaining a certain number of equations for signal reconstruction. We provided conditions that the sensing matrix of the equations obtained by AMWC achieves the Singleton bound, and thus no loss from sampling is guaranteed. In summary, the improved sampling efficiency of AMWC reduces the total sampling rate required for lossless sampling. In other words, with fewer channels and less sampling rate of each channel than those of the conventional MWCs, a multiband signal can be captured without information loss by AMWC. Conversely, for given hardware resources, the input reconstruction with AMWC outperforms the conventional MWCs. Extensive simulation demonstrated that AMWC indeed reduces the total sampling rate or improves the reconstruction performance significantly. Additionally, it was demonstrated that the benefits of AMWC are maintained in various SNRs. Moreover, use of LPF with random passband response, it was shown, further improves the sampling efficiency.

APPENDIX A
PROOF OF LEMMA 1

With the relationship $f_{LPF} = pf'_s$, the pass-band frequency of $H(f - rf'_s)$ in (8) is given by $f \in [rf'_s - \frac{pf'_s}{2}, rf'_s + \frac{pf'_s}{2}]$. When we observe (8) only for a single period $\mathcal{F}'_s(f_0)$, since $W_{LPF} > f'_s$, some of $H(f - rf'_s)$, the pass bands of which include the frequency domain $\mathcal{F}'_s(f_0)$, can be replaced by the constant frequency response. Without loss of generality, we set the pass-band response to one, i.e., $H(f) = 1$ for $f \in \mathcal{F}_{LPF}$. Then, for $r \in \mathbb{Z}$ satisfying

$$rf'_s - \frac{pf'_s}{2} \leq f_0 \quad (44)$$

and

$$rf'_s + \frac{pf'_s}{2} \geq f_0 + f'_s, \quad (45)$$

the shifts of filter responses in (8) are replaced with $H(f - rf'_s) = 1$ within $f \in \mathcal{F}'_s(f_0)$. Let R_1 and R_2 be the minimum and maximum integers r satisfying (44) and (45), respectively. Additionally, for (8) and (9) to be equivalent, we add some conditions on R_1 and R_2 such that the pass bands of $H(f - rf'_s)$ for r smaller than R_1 and greater than R_2 have no intersection with $f \in \mathcal{F}'_s(f_0)^C$. In other words, we have following conditions on R_1 and R_2 :

$$(R_2 + 1)f'_s - \frac{pf'_s}{2} \geq f_0 + f'_s \quad (46)$$

and

$$(R_1 - 1)f'_s + \frac{pf'_s}{2} \leq f_0 \quad (47)$$

so that $H(f - rf'_s) = 0$ within $f \in \mathcal{F}'_s(f_0)$ for $r < R_1$ or $r > R_2$. By combining (44) and (46), we have a condition on R_2 that

$$R_2f'_s - \frac{pf'_s}{2} = f_0, \quad (48)$$

and from (45) and (47), we have a condition on R_1 that

$$R_1f'_s + \frac{pf'_s}{2} = f_0 + f'_s. \quad (49)$$

Finally, combining (48) and (49) provides the conditions of Lemma 1. ■

APPENDIX B
PROOFS OF PROPOSITION 3 AND LEMMA 6

A. Proof of Proposition 3

We track the input-output relation starting from (12):

$$Y_i \left(e^{j2\pi f T'_s} \right) = \sum_{r=R_1}^{R_2} \sum_{l=-\infty}^{\infty} c_{i,l} X \left(f - (lp + rq') f'_p \right)$$

for $f \in \mathcal{F}'_s(f_0)$, where R_1, R_2 , and f_0 satisfy Lemma 1. Alternatively, by using $r' = r - R_1$, we have

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{r'=0}^{R_2-R_1} \sum_{l=-\infty}^{\infty} c_{i,l} X \left(f - (lp + (r' + R_1)q') f'_p \right) \\ &= \sum_{r'=0}^{p-1} \sum_{l=-\infty}^{\infty} c_{i,l} X \left(f - (lp + (r' + R_1)q') f'_p \right) \end{aligned} \quad (50)$$

for $f \in \mathcal{F}'_s(f_0)$, where $R_2 - R_1 = p - 1$ by Lemma 1. We replace the term $(r' + R_1)q'$ in (50) by a combination of its quotient $\mu_p(r'; q', R_1)$ and remainder $\rho_p(r'; q', R_1)$ by divisor p , which are, respectively, defined by

$$\mu_p(r'; q', R_1) \triangleq \left\lfloor \frac{(r' + R_1)q'}{p} \right\rfloor \quad (51)$$

and

$$\rho_p(r'; q', R_1) \triangleq ((r' + R_1)q') \bmod p. \quad (52)$$

By substituting $(r' + R_1)q' = p \cdot \mu_p(r'; q', R_1) + \rho_p(r'; q', R_1)$ into (50), we have

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{l=-\infty}^{\infty} \sum_{r'=0}^{p-1} c_{i,l} X \left(f - (lp + p \cdot \mu(r') + \rho(r')) f'_p \right) \\ &= \sum_{l=-\infty}^{\infty} \sum_{r'=0}^{p-1} c_{i,l-\mu(r')} X \left(f - (lp + \rho(r')) f'_p \right) \end{aligned} \quad (53)$$

for $f \in \mathcal{F}'_s(f_0)$, where the notations $\mu_p(r'; q', R_1)$ and $\rho_p(r'; q', R_1)$ are simplified to $\mu(r')$ and $\rho(r')$, respectively. When p and q' are coprime, by modular arithmetic, there exists one-to-one correspondence between $\rho(r')$ and r' modulo p . We arrange the order of inner summation of (53) by introducing a utility variable $v \triangleq \rho(r') \in \{0, \dots, p-1\}$:

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{l=-\infty}^{\infty} \sum_{v=0}^{p-1} c_{i,l-\mu(\rho_p^{-1}(v; q', R_1))} X \left(f - (lp + v) f'_p \right) \end{aligned} \quad (54)$$

for $f \in \mathcal{F}'_s(f_0)$, where the inverse $\rho_p^{-1}(v; q', R_1)$ of the remainder $\rho_p(r; q', R_1)$ modulo p is computed by

$$\rho_p^{-1}(v; q', R_1) \triangleq \left(v(q')^{-1} - R_1 \right) \bmod p, \quad (55)$$

where $(q')^{-1} \bmod p$ is the multiplicative inverse of q' modulo p . We simplify the expression $\rho_p^{-1}(v; q', R_1)$ to $\rho^{-1}(v)$. From Lemma 2, we can merge the inner and outer summations of (54) as follows:

$$Y_i \left(e^{j2\pi f T'_s} \right) = \sum_{k=-\infty}^{\infty} c_{i, \lfloor \frac{k}{p} \rfloor - \mu(\rho^{-1}(k \bmod p))} X \left(f - kf'_p \right) \quad (56)$$

for $f \in \mathcal{F}'_s(f_0)$.

We now simplify the picking regularity of the coefficients $c_{i,J(\cdot)}$ in (56), which is defined by

$$\begin{aligned} J(k; R_1, p, q') &\triangleq \left\lfloor \frac{k}{p} \right\rfloor - \mu(\rho^{-1}(k \bmod p)) \\ &= \left\lfloor \frac{k}{p} \right\rfloor - \mu(\rho^{-1}(k)). \end{aligned} \quad (57)$$

Meanwhile, by the definitions of the quotient $\mu(\cdot)$ and remainder $\rho(\cdot)$, we have

$$\begin{aligned} \mu(\rho^{-1}(k)) &= \left\lfloor \frac{(\rho^{-1}(k) + R_1)q'}{p} \right\rfloor \\ &= \frac{1}{p} \left((\rho^{-1}(k) + R_1)q' - ((\rho^{-1}(k) + R_1)q') \bmod p \right) \\ &= \frac{1}{p} \left((\rho^{-1}(k) + R_1)q' - \rho(\rho^{-1}(k)) \right) \\ &= \frac{1}{p} \left((\rho^{-1}(k) + R_1)q' - k \bmod p \right). \end{aligned} \quad (58)$$

By substituting (58) into (57),

$$\begin{aligned} J(k; R_1, p, q') &= \left\lfloor \frac{k}{p} \right\rfloor + \frac{k \bmod p}{p} - \frac{(\rho^{-1}(k) + R_1)q'}{p} \\ &= \frac{k}{p} - \frac{(\rho^{-1}(k) + R_1)q'}{p} \\ &= \frac{1}{p} \left\{ k - q' \cdot \left[(k(q')^{-1} - R_1) \bmod p + R_1 \right] \right\} \\ &= I(k; R_1, p, q'). \end{aligned} \quad (59)$$

Thus, the proof is completed.

B. Proof of Lemma 6

We track the input-output relation starting from (36):

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{r=R_1}^{R_2} \sum_{l=-\infty}^{\infty} c_{i,l} X(f - (rq' + lp)f'_p) G(f - rq'f'_p) \end{aligned}$$

for $f \in \mathcal{F}'_s(f_0)$. Under the conditions of Lemma 1 and Lemma 2, by using $r' \triangleq r - R_1$, we have

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{r'=0}^{R_2-R_1} \sum_{l=-\infty}^{\infty} c_{i,l} X(f - (lp + (r' + R_1)q')f'_p) \\ &\quad G(f - (r' + R_1)q'f'_p) \\ &= \sum_{r'=0}^{p-1} \sum_{l=-\infty}^{\infty} c_{i,l} X(f - (lp + (r' + R_1)q')f'_p) \\ &\quad G(f - (r' + R_1)q'f'_p) \end{aligned} \quad (60)$$

for $f \in \mathcal{F}'_s(f_0)$. As done in (50) to (54), we introduce a utility variable $v \triangleq \rho(r')$ and substitute $(r' + R_1)q' = p \cdot$

$\mu(\rho^{-1}(v)) + v$ into the inputs of X and G in (60). It then follows

$$\begin{aligned} Y_i \left(e^{j2\pi f T'_s} \right) &= \sum_{l=-\infty}^{\infty} \sum_{v=0}^{p-1} \left(c_{i,J(k; R_1, p, q')} X(f - (lp + v)f'_p) \right) \\ &\quad \cdot G(f - (p\mu(\rho^{-1}(v)) + v)f'_p) \end{aligned} \quad (61)$$

for $f \in \mathcal{F}'_s(f_0)$. After merging the inner and outer summations based on Lemma 2, we obtain (37)

$$\begin{aligned} \tilde{Y}_i \left(e^{j2\pi f T'_s} \right) &= \sum_{k=-\infty}^{\infty} d_{i,k}(R_1, p, q') X(f - kf'_p) G(f - \gamma_p(k)f'_p) \end{aligned}$$

for $f \in \mathcal{F}'_s(f_0)$, where $\gamma_p(k)$ is defined by

$$\begin{aligned} \gamma_p(k) &\triangleq p\mu(\rho^{-1}(k \bmod p)) + k \bmod p \\ &= p\mu(\rho^{-1}(k)) + k \bmod p. \end{aligned} \quad (62)$$

By (58) and the definition of $\rho^{-1}(k)$ in (55), (62) turns into

$$\begin{aligned} \gamma_p(k) &= (\rho^{-1}(k) + R_1)q' \\ &= q' \left[(kq^{-1} - R_1) \bmod p + R_1 \right]. \end{aligned} \quad (63)$$

By the definition of $I(k; R_1, p, q')$ in (15), we finally have

$$\gamma_p(k) = k - pI(k; R_1, p, q'). \quad (64)$$

Thus, the proof is completed. \blacksquare

APPENDIX C

PROOF OF PROPOSITION 4

We first show that if $p > q'$ for coprime p and q' , at least two columns of \mathbf{D} are identical. Then, from a result in [11], this violates a necessary condition for the unique existence of a K -sparse solution.

We first mathematically formulate the meaning of two columns of \mathbf{D} being identical. From Proposition 3 and (30), the entries $d_{i,k+u}(R_1, p, q')$ of \mathbf{D} are picked from $c_{i,I(k; R_1, p, q')}$, where k and u in $d_{i,k+u}$ represent the column and row position, respectively. To search for identical columns in \mathbf{D} , we investigate the existence of pairs (k^*, ω^*) of a column index k^* and shift index ω^* such that $d_{i,k^*+u} = d_{i,k^*+u+\omega^*}$ for every row index $u \in \mathcal{Q} \triangleq \{0, \dots, q' - 1\}$. In other words, we find pairs (k^*, ω^*) satisfying

$$I(k^* + \omega^* + u; R_1, p, q') = I(k^* + u; R_1, p, q'). \quad (65)$$

for every $u \in \mathcal{Q}$, where the function I is defined in (15). We use a computation result of $I(k) \triangleq I(k; R_1, p, q')$ in the second line of (59):

$$I(k) = \frac{k}{p} - \frac{(\rho^{-1}(k) + R_1)q'}{p}, \quad (66)$$

where $\rho^{-1}(k) \triangleq \rho_p^{-1}(k; q', R_1)$ is a function modulo p defined in (55) by $\rho_p^{-1}(k; q', R_1) \triangleq (k(q')^{-1} - R_1) \bmod p$. By

substituting (66) into (65), we rewrite (65) as

$$\begin{aligned} I(k^* + \omega^* + u) &= I(k^* + u) \\ \Leftrightarrow \rho^{-1}(k^* + \omega^* + u) &= \rho^{-1}(k^* + u) + \frac{\omega^*}{q'}. \end{aligned} \quad (67)$$

We show that, if $p > q'$ and coprime, there exists at least one pair (k^*, ω^*) of the column index k^* and shifting index ω^* that satisfy (67) for every row index $u \in \mathcal{Q}$. Before proceeding, we check a computation of $\rho^{-1}(k + q' + u)$ for every $u \in \mathcal{Q}$. By the definition, it follows

$$\begin{aligned} \rho^{-1}(k + q' + u) &= \left((k + q' + u)(q')^{-1} - R_1 \right) \bmod p \\ &= \left(\left((k + u)(q')^{-1} - R_1 \right) \bmod p + 1 \right) \bmod p \\ &= \left(\rho^{-1}(k + u) + 1 \right) \bmod p. \end{aligned} \quad (68)$$

Note that (68) indicates when ω^* is chosen to q' , it satisfies (67), for $k^* \in \mathbb{Z}$ such that $\rho^{-1}(k^* + u) < p - 1$.

What task remains is to show the existence k^* satisfies $\rho^{-1}(k^* + u) < p - 1$ for every row index $u \in \mathcal{Q}$, which implies the existence of identical columns in \mathbf{D} and completes the proof. To this end, we find a set of $\bar{k}(\bmod p)$ such that $\rho^{-1}(\bar{k} + u) = p - 1$. From the definition, we have

$$\begin{aligned} \rho^{-1}(\bar{k} + u) &\equiv p - 1 \pmod{p} \\ (\bar{k} + u)(q')^{-1} - R_1 &\equiv p - 1 \pmod{p} \\ \bar{k} &\equiv (p - 1 + R_1)q' - u \pmod{p} \\ \bar{k} &\equiv (R_1 - 1)q' - u \pmod{p}. \end{aligned} \quad (69)$$

Note that $(R_1 - 1)q'$ is a constant. Since the right-hand side of (69) varies by $u \in \mathcal{Q}$, the cardinality of set of $\bar{k}(\bmod p)$ such that $\rho^{-1}(\bar{k} + u) = p - 1$ is $|\mathcal{Q}| = q'$. Since $p > q'$, this implies there exists $k^*(\bmod p) \in \{0, 1, \dots, p - 1\}$ such that $\rho^{-1}(k^* + u) < p - 1$, and $k^* \in \mathbb{Z}$ such that $\rho^{-1}(k^* + u) < p - 1$ exists as well.

Consequently, if coprime $p > q'$, there must exist at least one pair of identical columns in \mathbf{D} . The existence of identical columns in \mathbf{D} implies $\text{spark}(\mathbf{D}) = 2$. Theorem 2 in [11] states that there exist the unique solution of a linear equation $\mathbf{Z} = \mathbf{D}\mathbf{X}$ for every K -sparse solution \mathbf{X} only if

$$K < \frac{\text{spark}(\mathbf{D}) - 1 + \text{rank}(\mathbf{X})}{2}, \quad (70)$$

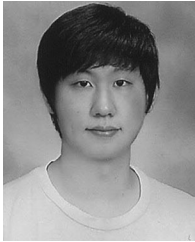
where spark is the minimum number of linearly dependent columns in \mathbf{D} . If $\text{spark}(\mathbf{D}) = 2$, for signals \mathbf{X} with $\text{rank}(\mathbf{X}) \leq 2K - 1$, the condition $p > q'$ violates (70). ■

REFERENCES

- [1] H. J. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Math.*, vol. 117, no. 1, pp. 37–52, 1967.
- [2] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.
- [3] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 375–391, Apr. 2010.
- [4] Y. Chen, H. Chi, T. Jin, S. Zheng, X. Jin, and X. Zhang, "Sub-Nyquist sampled analog-to-digital conversion based on photonic time stretch and compressive sensing with optical random mixing," *J. Lightw. Technol.*, vol. 31, no. 21, pp. 3395–3401, Nov. 2013.
- [5] Y. Zhao, Y. H. Hu, and J. Liu, "Random triggering-based sub-Nyquist sampling system for sparse multiband signal," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1789–1797, Jul. 2017.
- [6] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.
- [7] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [8] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [9] S. Park, N. Y. Yu, and H.-N. Lee, "An information-theoretic study for joint sparsity pattern recovery with different sensing matrices," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5559–5571, Sep. 2017.
- [10] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.
- [11] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.
- [12] S. A. Varma and K. M. M. Prabhu, "A new approach to near-theoretical sampling rate for modulated wideband converter," in *Proc. Int. Conf. Signal Process. Commun.*, 2014, pp. 1–5.
- [13] M. Sakare, "A power and area efficient architecture of a PRBS generator with multiple outputs," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 64, no. 8, pp. 927–931, Aug. 2017.
- [14] L. Vera and J. R. Long, "A 40-Gb/s 2^{11} -1 PRBS with distributed clocking and a trigger countdown output," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 63, no. 8, pp. 758–762, Aug. 2016.
- [15] T. Chen, M. Guo, Z. Yang, and W. Zhang, "A novel and efficient compressive multiplexer for multi-channel compressive sensing based on modulated wideband converter," in *Proc. IEEE Inf. Technol. Netw. Electron. Autom. Control Conf.*, May 2016, pp. 347–351.
- [16] T. Haque, R. T. Yazicigil, K. J.-L. Pan, J. Wright, and P. R. Kinget, "Theory and design of a quadrature analog-to-information converter for energy-efficient wideband spectrum sensing," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 2, pp. 527–535, Feb. 2015.
- [17] Y. Chen, A. J. Goldsmith, and Y. C. Eldar, "On the minimax capacity loss under sub-Nyquist universal sampling," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3348–3367, Jun. 2017.
- [18] Y. Chen, Y. C. Eldar, and A. J. Goldsmith, "Shannon meets Nyquist: Capacity of sampled Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4889–4914, Aug. 2013.
- [19] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.
- [20] Y. Jin and B. D. Rao, "Support recovery of sparse signals in the presence of multiple measurement vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3139–3157, May 2013.
- [21] *IEEE Standard for Letter Designations for Radar-Frequency Bands*, IEEE Standard 521-2002, Jan. 2003.
- [22] F. E. Nathanson, P. J. O'Reilly, and M. N. Cohen, *Radar Design Principles: Signal Processing and the Environment*. Raleigh, NC, USA: Sci Tech Publishing Inc., 2004.
- [23] T. O. Dickson *et al.*, "An 80-Gb/s $2/\text{sup } 31/1$ pseudorandom binary sequence generator in SiGe BiCMOS technology," *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2735–2745, Dec. 2005.
- [24] A. Gharib, A. Talai, R. Weigel, and D. Kissinger, "A 1.16 pJ/bit 80 Gb/s 2^{11} -1 PRBS generator in SiGe bipolar technology," in *Proc. 9th Eur. Microw. Integr. Circuit Conf.*, 2014, pp. 277–280.
- [25] Hittite Microwave, "Hittite's 18 GHz ultra wideband track-and-hold amplifier enhances high speed ADC performance," 2011. [Online]. Available: http://www.analog.com/media/en/technical-documentation/technical-articles/track-n-hold_0411.pdf
- [26] R. Singleton, "Maximum distance Q-nary codes," *IEEE Trans. Inf. Theory*, vol. 10-IT, no. 2, pp. 116–118, Apr. 1964.
- [27] Y. Chen, M. Mishali, Y. C. Eldar, and A. O. Hero III, "Modulated wideband converter with non-ideal lowpass filters," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Jun. 2010, pp. 3630–3633.
- [28] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," arXiv09013403, 2009.
- [29] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1872–1882, Apr. 2009.



Jehyuk Jang received the B.S degree in electronic engineering from the Kumoh National Institute of Technology, Gumi, Korea, in 2014 and M.S degree in information and communication engineering from GIST, Gwangju, Korea. He is currently working toward the Ph.D. degree at the School of Electrical Engineering and Computer Science, GIST, Korea. His research interests include sub-Nyquist sampling and compressed sensing.



Sanghun Im received the B.S. degree in electronics engineering from Soongsil University, Seoul, Korea, in 2009, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2011 and 2016, respectively. He is currently working with Hanwha Systems, Seongnam, Korea. His current research interests include communication theories and signal processing for wireless communications and physical layer security.



Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. From 1999 to 2002, he was a Research Staff member with HRL Laboratories, LLC, Malibu, CA, USA. From 2002 to 2008, he was an Assistant Professor with the University of Pittsburgh, PA, USA. In 2009, he was with the School of Electrical Engineering and Computer Science, GIST, Korea, where he is currently affiliated.

His research interests include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He was the recipient of several prestigious National Awards, including the Top 100 National Research and Development Award in 2012, the Top 50 Achievements of Fundamental Researches Award in 2013, and the Science/Engineer of the Month (January 2014).

Article

Speckle Reduction on Ultrasound Liver Images Based on a Sparse Representation over a Learned Dictionary

Mohamed Yaseen Jabarulla and Heung-No Lee *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea; yaseen@gist.ac.kr

* Correspondence: heungno@gist.ac.kr

Received: 20 April 2018; Accepted: 28 May 2018; Published: 31 May 2018



Abstract: Ultrasound images are corrupted with multiplicative noise known as speckle, which reduces the effectiveness of image processing and hampers interpretation. This paper proposes a multiplicative speckle suppression technique for ultrasound liver images, based on a new signal reconstruction model known as sparse representation (SR) over dictionary learning. In the proposed technique, the non-uniform multiplicative signal is first converted into additive noise using an enhanced homomorphic filter. This is followed by pixel-based total variation (TV) regularization and patch-based SR over a dictionary trained using K-singular value decomposition (KSVD). Finally, the split Bregman algorithm is used to solve the optimization problem and estimate the de-speckled image. The simulations performed on both synthetic and clinical ultrasound images for speckle reduction, the proposed technique achieved peak signal-to-noise ratios of 35.537 dB for the dictionary trained on noisy image patches and 35.033 dB for the dictionary trained using a set of reference ultrasound image patches. Further, the evaluation results show that the proposed method performs better than other state-of-the-art denoising algorithms in terms of both peak signal-to-noise ratio and subjective visual quality assessment.

Keywords: ultrasound; speckle reduction; medical image processing; sparse representation; K-singular value decomposition; dictionary learning; B-mode imaging

1. Introduction

In the last 20 years, there has been growing interest in the use of ultrasound imaging for a variety of applications, such as observing the blood flow through an organ or other structures; determining bone density; imaging the heart, a fetus, or ocular structures; or diagnosing cancers [1,2]. Ultrasound imaging has been widely applied owing to its ability to produce real-time images and videos. Ultrasound images are captured in real-time by transmitting high frequency sound waves through body tissue. It comprises an array of transducer elements that sequentially echo the signal for each spatial direction to generate a raw line signal. The scan is converted to construct a Cartesian image from the processed raw line signal [2].

In recent years, many researchers have attempted to develop computer-aided diagnostic (CAD) systems for diagnosing liver and breast cancers [3–6] based on ultrasound imaging. The aim of these systems is to differentiate benign and malignant lesion tissues as well as cysts [7]. A CAD system carries out the diagnosis in four stages: data preprocessing, image segmentation, feature extraction, and classification [4]. Data preprocessing is the first and most vital step in the CAD system process because it reconstructs an image without eliminating the important features by reducing signal-dependent multiplicative noise called speckle [8].

The development of a precise speckle reduction model is an important step to achieve efficient denoising filter design. Recent review articles [4,9], reported that speckle reduction filters are

categorized into two broad approaches: spatial filtering and multiscale methods. Techniques under spatial domain filtering include enhanced Frost filtering [10], Lee filtering [11], mean filtering [12], Wiener filtering [13], Kuan filtering [14], and median filtering [15]. Spatial filters utilize local statistical properties to reduce speckle noise. However, small details may not be preserved [9]. Several methods [16–19] use multiscale filtering, which uses the wavelet transform to preserve the image signal regardless of its frequency content. Donoho et al. [20] proposed reducing noise in the wavelet domain by soft thresholding. However, their approach lacked translation invariance when using the discrete wavelet transform. This is resolved by eliminating up and down samplers in the wavelet transform by using a stationary wavelet transform [21], which is a redundant technique because the number of input and output samples at each level is the same. A multiresolution technique called translation invariant image enhancement was proposed in [22]. The proposed technique incorporates noise reduction and directional filtering. Directional filtering is executed using eigenvalues by analyzing the structure of each pixel's neighborhood. Rudin et al. [23,24] and Perona et al. [25] proposed successful image denoising techniques called total variation (TV) and anisotropic smoothing, respectively. These models were improved and extended upon in later works [26,27]. However, all these methods are computationally expensive. In recent years, more efficient denoising techniques such as sparse representation (SR) have been proposed [28–31]. In digital image processing, many signals are sparse; i.e., they contain many coefficients either equal to or close to zero in a specific domain. The objective of SR is to efficiently reconstruct the signal with a linear combination of a few dictionary atoms from the transformed signal domain [32].

This study was conducted with the objective of developing filtering algorithm that can reduce noise without losing significant features or eliminating edges. To this end, this paper proposes, a technique that reduces the speckle noise in ultrasound imaging systems by applying a relatively new signal reconstruction model known as SR [32] to deal with complicated noise properties. Sparse representation provides superior estimation even in an ill-conditioned system [33], and has been found to be very useful in medical imaging applications. However, one challenge of designing this system is the presence of a multiplicative speckle signal because dictionary learning methods are not effective on multiplicative and correlated noise. We overcome this by using two different methods. Firstly, the speckle noise is transformed into additive noise using an enhanced homomorphic filter that can also capture high and low frequency signal of the image. Secondly, we introduced TV regularization of the image and sparse prior over learned dictionaries. Total variation regularization is efficient for noisy image, while the patch-based dictionaries are well adapted to texture features [34], and reduces the artifacts in smooth pixel regions [35]. The advantage of the sparse prior is that it utilizes fewer dictionary columns to reconstruct a noiseless ultrasound image without losing many important features of the signal. Therefore, in our proposed model we combined the two approaches, the patch-based SR over learned dictionaries and the pixel-based TV regularization method, for efficient speckle reduction. The K-singular value decomposition (KSVD) algorithm [36] is used to learn two modified dictionaries from reference ultrasound image datasets and the corrupted images; these are referred to as dictionaries 1 and 2, respectively. The results are evaluated on both dictionaries and compared with conventional algorithms to show that the speckle noise is suppressed effectively in the ultrasound image using SR.

The rest of the paper is organized as follows. Noise model and related works are described in Section 2. The proposed SR framework for speckle reduction in ultrasound imaging is presented in Section 3. In Section 4, the experiments and results obtained are discussed. The paper is concluded in Section 5.

2. Background

2.1. Ultrasound Noise Model

Ultrasound imaging system are often affected by multiplicative speckle [37]. Scattering time differences lead to constructive and destructive interference of the ultrasound pulses that are reflected

from biological tissues. Speckle patterns can be classified depending on the spatial distribution, number of scatters per resolution cell, and properties of the imaging system [9]. Speckle noise affects the detectability of the target and reduces the contrast and resolution of the images, making it difficult for a clinician to provide a diagnosis.

In ultrasound, the multiplicative noise models are based on the product of the original signal and noise. Thus, the intensity of a noisy signal depends on the original image intensity. The mathematical expression for a multiplicative speckle model is given by

$$y(i, j) = x(i, j)h(i, j), \quad (1)$$

where $y(i, j)$ is the speckled image, $x(i, j)$ is the original image, and $h(i, j)$ is the speckle noise. The spatial location of an image is represented using indexes i and j , where index i ranges from 1 to N , and index j from 1 to M .

2.2. Related Work on Multiplicative Noise Reduction

Several algorithms have been proposed to deal with more complex multiplicative and additive speckle noise models [38]. For instance, the Kuan, Frost, Lee filters, and speckle reducing anisotropic diffusion (SRAD) filter [39] are effective on the multiplicative noise model. Other filters, specifically the median, Wiener, and wavelet filters [40], are designed for the additive noise model [4]. However, each filter has certain advantages and limitations [38]. In a few filter models, the quality of the processed image is affected by the window size: large window sizes cause image blurring, degrading the fine details of an image. Conversely, small window sizes do not denoise the image sufficiently. Other widely used multiplicative noise reduction algorithms are based on the TV regularization term [23,41], nonlocal methods [42,43], and wavelet-based approaches [16]. Total variation-based methods effectively remove flat-region-based noise and preserve the edges of images. However, fine details are lost because of over-smoothed textures. Nonlocal algorithms depend on similarities of image patches. Their performance is limited by dissimilar image patches. However, wavelet-based approaches preserve texture information better than TV-based methods. This approach assumes that images in the SRs are based on a fixed dictionary [29,36]. However, certain characteristics of the processed image might not be captured because the dictionary does not contain any similar image content.

To overcome the above disadvantages, over the past few years, researchers have sought to develop an algorithm based on SR in the field of image and signal processing [32]. This is because the pattern similarities of image signals such as textures and flat regions, mean that the signal can be efficiently approximated as a linear combination using a dictionary of only a few functions called atoms [29,34,36]. Elad and Aharon [36] proposed an image denoising algorithm using an adaptive dictionary called KSVD that is based on sparse and redundant representations. It includes sparse coding and dictionary atoms that are updated to better fit the data. The advantage of KSVD compared to fixed dictionaries is that it is effective at removing additive Gaussian noise using the linear combinations of a few atoms, by learning a dictionary from noisy image patches and then reconstructing each patch.

A dictionary $A \in \mathbb{R}^{N_r \times N_c}$, composed of N_c columns of N_r elements, is called a sparse-land model [36]. K-singular value decomposition seeks the best signal representation of image signal y from the sparsest representation α :

$$\hat{\alpha} = \operatorname{argmin} \|\alpha\|_0 \text{ subject to } \|y - A\alpha\|_2 \leq \varepsilon,$$

where the vectorization of $y(i, j)$ is denoted by vector $y \in \mathbb{R}^{M \times 1}$ and ε is the few number of non-zero entries in α . K-singular value decomposition replaces the dictionary update and sparse coding stages with a simple singular value decomposition. The orthogonal matching pursuit (OMP) method [44] is an effective method to find the sparse approximation. In the OMP, if the noise level is below the approximation, the image patches are rejected. The singular value decomposition constructs better

atoms by combining patches to reduce noise for ultrasound speckle reduction. K-singular value decomposition has also proved to effectively reduce the speckle produced by additive white Gaussian noise on corrupted images [29,36].

The filtering algorithm comprises two steps. First, the dictionary is trained from a set of image data patches or from noisy image patches based on KSVD. The next step uses $\hat{\alpha}$ to compute SR using dictionary A and denoises the image [29].

The method proposed in [45] also uses a dictionary learning approach for denoising ultrasound images. A homomorphic filter is used to convert multiplicative noise into additive white Gaussian noise and then the noiseless signal is reconstructed over image patches (atoms) to create the SR from a learned dictionary. However, noise in flat regions still exists and poor edges make the reconstructed images difficult to analyze. In [34], the authors proposed an image denoising technique that operates directly on multiplicative noise and is based on three terms: SR over an adaptive dictionary, a TV regularization term, and a data-fidelity term. However, the proposed model is nonconvex because of the product between the unknown dictionary and sparse coefficients and the data-fidelity term is a log function. Therefore, solving the squared l_2 norm is difficult. This optimization problem is overcome by the split Bregman technique. However, these methods do not contain high- and low-frequency components of the image. We obtain this information using an enhanced homomorphic filter designed to improve the final image. Furthermore, we utilize the advantages of combining a TV regularization term and SR learned over two modified dictionaries.

3. Sparse Representation Framework for Speckle Reduction

As discussed above, we define our proposed scheme for ultrasound speckle reduction by considering the multiplicative noise model [37] obtained by an ultrasound transducer. Equation (1) can thus be rewritten as

$$y_{\partial}(i, j) = x_{\mathbb{R}}(i, j)n_{\sigma}(i, j), \tag{2}$$

where $y_{\partial}(i, j)$ is the degraded B-mode image signal [46], $x_{\mathbb{R}}(i, j)$ represents the ideal image that must be recovered, and $n_{\sigma}(i, j)$ represents the speckle noise, generally modelled as a Rayleigh probability density function with random variables [11,47]. Each term includes coordinates (i, j) defined according to the acquisition geometry.

In general, a homomorphic filter [48] is a well-proven technique for converting multiplicative noise. In this study, we modified it by taking the log of the multiplicative noisy signal and filtering the image using a Butterworth high-pass (BW-HP) filter to attenuate low frequencies in the transmitted signal while preserving the high frequencies in the reflected component. The equation of the BW-HP filter is

$$H_B(u, v) = \frac{1}{1 + [D_0/\sqrt{u^2 + v^2}]^{2f}}, \tag{3}$$

where, D_0 is the cut-off frequency and f is the order of the filter. We varied the frequency values u and v of the i and j spatial coordinates. We used the BW-HP filter because it generates fewer ringing artifacts on the image signal.

We also used a Gaussian low pass (GLP) filter to smooth the low-frequency signal component in the log domain. The equation of the GLP filter is

$$H_G(u, v) = e^{-D^2(u,v)/2D_0^2}, \tag{4}$$

where $D(u,v)$ is the distance from the origin in the frequency plane. Finally, the additive noise signals were estimated by applying inverse transform.

Figure 1 shows the steps used to convert an original noisy image into an image with additive noise using the enhanced homomorphic transform. This technique consists of five steps. We first take the log on both sides of Equation (2) and use a two-dimensional fast Fourier transform (FFT)

to represent the image in the frequency domain. Then, the Fourier image is filtered with two filter functions, those are the BW-HP and GLP filters [12]. The BW-HP filter increases the contrast of the image signal corresponding to the high-frequency component. The GLP filter smooths the noise signal without eliminating the entire low-frequency component. Both filtered signals are applied to the two-dimensional inverse fast Fourier transform (IFFT). Finally, taking the exponent of the image, we obtain the transformed image. This process is discussed in detail below.

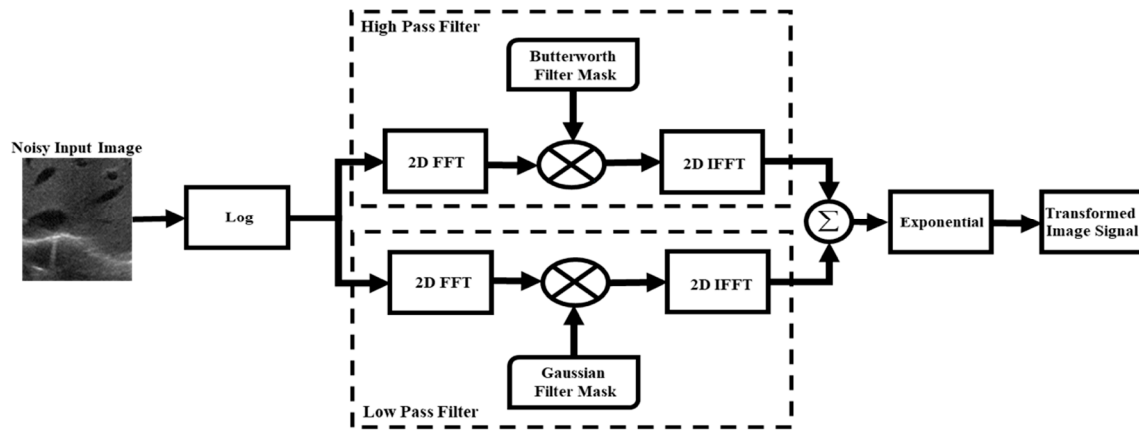


Figure 1. Flow diagram of the enhanced homomorphic filter. FFT: fast Fourier transform; IFFT inverse fast Fourier transform.

Step 1: Take the log on both sides of the $x_{\Re}(i, j)$ and the $n_{\sigma}(i, j)$ signal; now the multiplicative noise can be written as

$$\log(y_{\partial}(i, j)) = \log(x_{\Re}(i, j)) + \log(n_{\sigma}(i, j)), \tag{5}$$

After being transformed logarithmically, the signal now contains Gaussian additive noise [49]. We remove $\log(x_{\Re}(i, j))$ from the speckled ultrasound image by applying an additive noise suppression algorithm. Thus, the problem is now to estimate $\log(x_{\Re}(i, j))$ from noisy data.

Step 2: Apply FFT to convert the image into the frequency domain. Equation (5), thus becomes,

$$y_{\partial}(u, v) = F_{x_{\Re}}(u, v) + F_{n_{\sigma}}(u, v), \tag{6}$$

where, $F_{x_{\Re}}(u, v)$ and $F_{n_{\sigma}}(u, v)$ are the FFT of $\log(x_{\Re}(i, j))$ and $\log(n_{\sigma}(i, j))$, respectively.

Step 3: Apply BW-HP and GLP to the $y_{\partial}(u, v)$ by means of two filter functions $H_B(u, v)$ and $H_G(u, v)$ from Equations (3) and (4) respectively in the frequency domain. The filtered version of $S(u, v)$ is written as

$$S(u, v) = H_B(u, v)y_{\partial}(u, v) + H_G(u, v)y_{\partial}(u, v). \tag{7}$$

Step 4: Take the inverse Fourier transform of Equation (7) to get the converted signal in the spatial domain

$$\bar{S}(i, j) = \mathcal{F}^{-1}\{S(u, v)\}.$$

Step 5: Finally, we obtain the transformed image $t(i, j)$ by taking the exponent of the image using the following equation

$$t(i, j) = \exp\{\bar{S}(i, j)\}.$$

In this paper, we model the transformed image as additive noise degradation $W(i, j)$ of the original image $x_{\Re}(i, j)$, i.e.,

$$t(i, j)x_{\Re}(i, j) + W(i, j). \tag{8}$$

This completes how we have used the homomorphic filter to transform the speckle noise into additive noise. The two filter functions are utilized to improve edge information by enhancing contrast and smooths the additive noise of the transformed image.

Figure 2 shows the output of the enhanced homomorphic filter at the BW-HP and GLP filter stages. It is clear that the image in Figure 2b has an increased intensity because the low frequency signal is attenuated and the image in Figure 2c is smoothed by the GLP filter. The sum of these two signals is the final transformed noisy image.

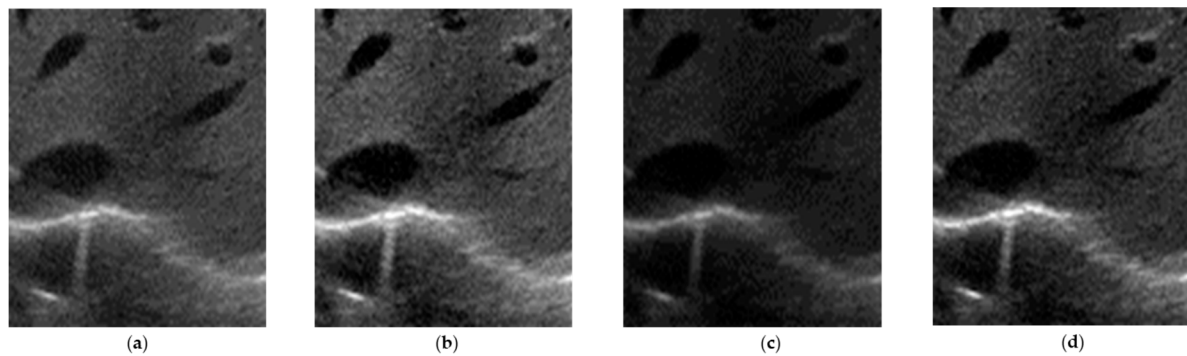


Figure 2. (a) Noisy ultrasound image; (b) Butterworth high-pass (BW-HP) filtered image; (c) Gaussian low pass (GLP) filtered image; and (d) transformed output of ultrasound noisy image.

An ultrasound image $x_{\mathbb{R}}(i, j)$ can be represented as sparse in the gradient domain. We thus define here a difference signal. A pixel-based TV regularization can be performed on the transformed image for more effective denoising. The horizontal and vertical difference matrices are defined below [50].

$$V_i x_{\mathbb{R}}(i, j) = \begin{cases} x_{\mathbb{R}}(i + 1, j) - x_{\mathbb{R}}(i, j), & \text{if } i < n \\ 0, & \text{if } i = n \end{cases}$$

$$V_j x_{\mathbb{R}}(i, j) = \begin{cases} x_{\mathbb{R}}(i, j + 1) - x_{\mathbb{R}}(i, j), & \text{if } j < m \\ 0, & \text{if } j = m \end{cases}$$

Further, the difference signal of $x_{\mathbb{R}}(i, j)$ is defined as

$$V_{i,j} x_{\mathbb{R}}(i, j) = \begin{pmatrix} V_i x_{\mathbb{R}}(i, j) \\ V_j x_{\mathbb{R}}(i, j) \end{pmatrix}.$$

We can show that there exists a dictionary $A \in \mathbb{R}^{N_r \times N_c}$ with which the original image can be sparsely represented as

$$x_{\mathbb{R}} = A\alpha,$$

where $x_{\mathbb{R}}$ is the vectorization of the recovered signal $x_{\mathbb{R}}(i, j)$ such that $x_{\mathbb{R}} \in \mathbb{R}^{N_r}$. If a signal $x_{\mathbb{R}}$ is K -sparse in the dictionary $A \in \mathbb{R}^{N_r \times N_c}$ for $N_c > N_r$, we imply that the signal can be represented with K columns of the dictionary. The column vector $\alpha \in \mathbb{R}^{N_c \times 1}$ is the vector of the coefficients. Then, by optimizing the following convex problem, the signal $x_{\mathbb{R}}$ can be recovered:

$$\begin{aligned} & \min \|\alpha\|_0, \\ & \text{subject to } \|t - A\alpha\|_2^2 \leq \varepsilon. \end{aligned} \tag{9}$$

In Equation (9), a $NM \times 1$ column vector t is the vectorization of the transformed image $t(i, j)$, note that $NM = N_r$. Also note that ε is a utility parameter selectable according to the noise strength.

This convex constrained problem can be transformed into an unconstrained optimization problem using the Lagrange multiplier method [51]:

$$\min \|t - A\alpha\|_2^2 + \tau \|\alpha\|_0. \quad (10)$$

Using the unconstrained problem, we are able to combine a regularization term, which is weighted by parameter $\tau > 0$ and a quadratic data-fidelity term. Equation (10) is not ready for use yet since we do not know the sparsity dictionary A . Therefore, we use the following approach where the dictionary, the sparse representation coefficient vector α , and the image vector $x_{\mathcal{R}}$ are estimated altogether. The overall optimized discrete sparse model proposed in this paper, for denoising the ultrasound image, can be written as

$$\left\{ \hat{x}_{\mathcal{R}}, \hat{\alpha}_{ij}, \hat{A} \right\} = \min_{x_{\mathcal{R}}, \alpha_{ij}, A} \lambda \|Vx_{\mathcal{R}}\|_1 + \tau \sum_{ij} \|R_{ij}t - A\alpha_{ij}\|_2^2 + \tau \sum_{ij} \|\alpha_{ij}\|_0, \quad (11)$$

where R_{ij} is an operation that extracts a square image patch from the transformed image t located at the i, j pixels of the image. The notation $\|\cdot\|_1$ is used to imply the l_1 norm, which is the sum of the absolute values of the argument signal, which in this case is the difference signal $Vx_{\mathcal{R}}$. There are two positive parameters λ and τ used to balance the contribution of different terms. In Equation (11), the first and second terms are the TV regularization norm and the sparse representation prior. Optimization in Equation (11) seeks to find a solution with which each patch of the recovered image can be represented by a dictionary matrix with sparse coefficient α in the sense of a bounded error. The l_0 norm gives the sparsity constraint which controls the sparsity coefficients of any small image patch.

As mentioned in Related Work Section 2.2, there is a sparse coding stage that utilizes the KSVD iterative process. In the first stage, sparse coding is performed assuming fixed $x_{\mathcal{R}}$ and A . In the second stage, dictionary A is updated to minimize using known sparse coefficients α and $x_{\mathcal{R}}$. The sparse coefficients $\hat{\alpha}_{ij}$ are computed using the OMP method [52] because of its efficiency and simplicity. Elad et al. [29] showed that learning a dictionary trained from good quality image patches and noisy images results in better performance.

In this paper, we use two approaches to train the dictionary. The first approach is to use a group of image patches taken from many ultrasound reference images. We call the dictionary obtained from this approach Dictionary 1. The second approach is to use the corrupted images and call them Dictionary 2. We aim to compare the performance difference based on these two approaches. The comparison is made in the Results section.

It should be noted that Equation (11) is non-convex because of the non-differentiable TV regularization term and the product of the unknowns A and α_{ij} . We overcome this by using the split Bregman iterative approach [53].

Overall, the proposed algorithm can be summarized as follows:

1. Convert the multiplicative noise into additive noise using an enhanced homomorphic filter and capture the high- and low-frequency components to retain detailed information.
2. Apply pixel-based TV regularization to smooth the filtered image signal.
3. Apply patch-based sparse representation over a dictionary trained using the KSVD algorithm. We employed two modified dictionaries—one trained with a set of reference ultrasound image patches and another trained using the speckled image patches.
4. Iterate between the TV regularization and sparse representation procedure to improve the reconstructed image.

Figure 3 summarizes the proposed algorithm.

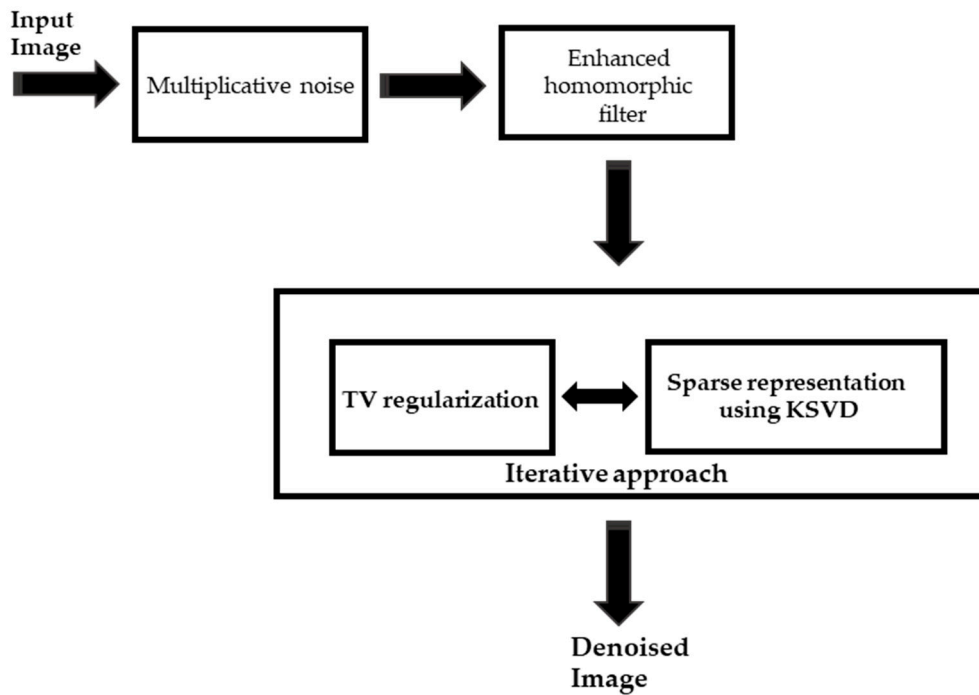


Figure 3. Proposed despeckle model for an ultrasound image. KSVD: K-singular value decomposition.

3.1. Performance Estimation

The reconstructed denoised image using the proposed algorithm were compared with the original image. Two image quality metrics were used for quantitative performance measurements: peak signal-to-noise ratio (PSNR) and mean structural similarity (MSSIM) [54]. Peak signal-to-noise ratio is defined as:

$$PSNR = 10 \log_{10} \frac{N_{\max}}{\frac{1}{MN} \sum_{n=1}^N \sum_{m=1}^M |x(n, m) - \hat{x}(n, m)|^2}, \quad (12)$$

where N_{\max} represents the maximum fluctuations in the input image. Here, $N_{\max} = (2^n - 1)$, $N_{\max} = 255$, when the components of a pixel are encoded using eight bits. N denotes the number of pixels processed, $x(n, m)$ is the original signal, and $\hat{x}(n, m)$ is the recovered image signal. In MSSIM, the structures of the two images are compared after normalizing the variance and subtracting the luminance as follows:

$$MSSIM = \frac{1}{N} \sum_{i=1}^N [l(\hat{x}, x)]^\alpha \cdot [c(\hat{x}, x)]^\beta \cdot [s(\hat{x}, x)]^\gamma, \quad (13)$$

where $l(\hat{x}, x)$ denotes luminance, $c(\hat{x}, x)$ denotes contrast, and $s(\hat{x}, x)$ denotes structure comparison functions. Further, α , β , and γ are weighted parameters that are used to adjust the relative importance of the three components.

4. Experimental Results and Discussion

4.1. Simulations on Synthetic Images

In this section, we analyze the performance of the proposed approach on the synthetic Shepp–Logan phantom test image [55] (Figure 4a) with a speckle noise variance of $\sigma = 10$ (Figure 4b) of a 256×256 pixel size. This result helps us to understand the effectiveness of the simulated image, clearly determine the distinctive features of the image, and optimize the algorithm before testing on the clinical datasets. We compared the proposed algorithm with some standard speckle reduction

filters for ultrasound liver images [4]. The compared algorithms were local statistical filters such as the Frost filter [10], Lee filter [11], 3×3 Wiener filter [13], Kuan filter [14], 3×3 median filter [15], and speckle reducing anisotropic diffusion (SRAD) filter [39]. In addition, multiscale filters such as wavelets [40] were evaluated. The despeckled images in Figure 4e–g show that the Frost, wavelet, and Kuan filters do not effectively reduce noise. In contrast, Figure 4h–j show that the median, Wiener, and SRAD filters, reduce most noise; however, the edges are not preserved and artificial noises can be introduced to a certain extent. This result verifies that the proposed SR technique reduces noise and preserves the edges better than the conventional methods on synthetic images. Table 1 shows the PSNR value and MSSIM value. The proposed algorithm reconstructs the original image with a PSNR value of 36.86 dB with Dictionary 1 and 37.04 dB with Dictionary 2.

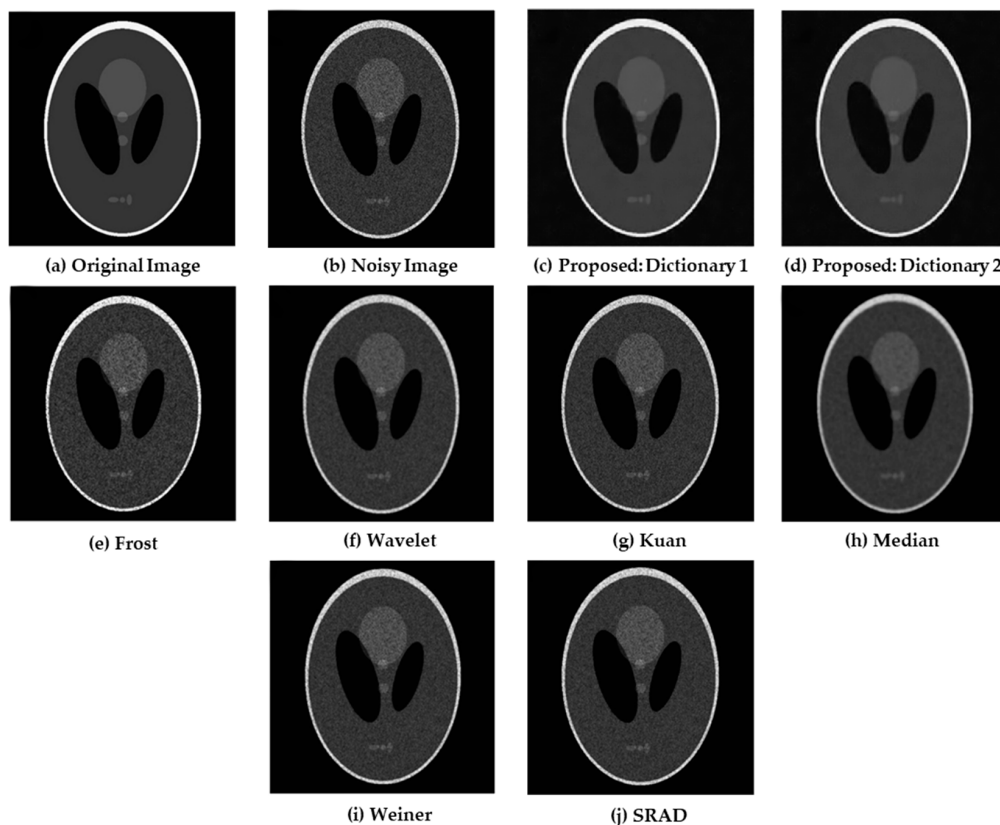


Figure 4. (a) Original image; (b) noisy image. Results of the proposed method with (c) Dictionary 1 and (d) Dictionary 2; Results of the (e) Frost; (f) wavelet; (g) Kuan; (h) median; (i) Wiener; and (j) speckle reducing anisotropic diffusion (SRAD) filters.

Table 1. Peak signal-to-noise ratio (PSNR) and mean structural similarity (MSSIM) for the synthetic images for $\sigma = 10$.

Models	PSNR (dB)	MSSIM
Noise image	32.113	0.727
Frost	32.466	0.768
Wavelet	33.214	0.801
Kuan	32.895	0.794
Median	34.597	0.839
SRAD	33.434	0.827
Weiner	33.782	0.834
Proposed: Dictionary 1	36.862	0.953
Proposed: Dictionary 2	37.044	0.967

4.2. Clinical Liver Ultrasound Images

The proposed algorithm efficiency was estimated using a set of B-mode greyscale ultrasound liver images. The images were obtained using the ECUBE 12R ultrasound research system from Alpinion medical systems, Seoul, Korea. The components used to generate the ultrasound images include a 128-element linear transducer at a center frequency of 5 MHz, a lateral beam width of 1.5 mm, and a pulse length of 1 mm. In our experiment, sparse coding was performed using two dictionaries with a 64×256 size, designed to handle patches of 8×8 size pixels ($N = 64$ and $K = 256$)—one trained from a noisy image and the other trained from a set of reference images.

The training data were constructed from a dataset comprising 3245 reference ultrasound images. The random collection of 16×16 dictionary atoms ($K = 256$) is presented in Figure 5a and the dictionary trained on the noisy image itself by overlapping patches is represented in Figure 5b. Where, every dictionary atom occupies a cell of 8×8 pixel ($N = 64$). We performed the tests on the three ultrasound reference images shown in Figures 6a, 7a and 9a. The KSVD algorithm was initialized with a trained dictionary and executed 180 iterations, as recommended in [29].

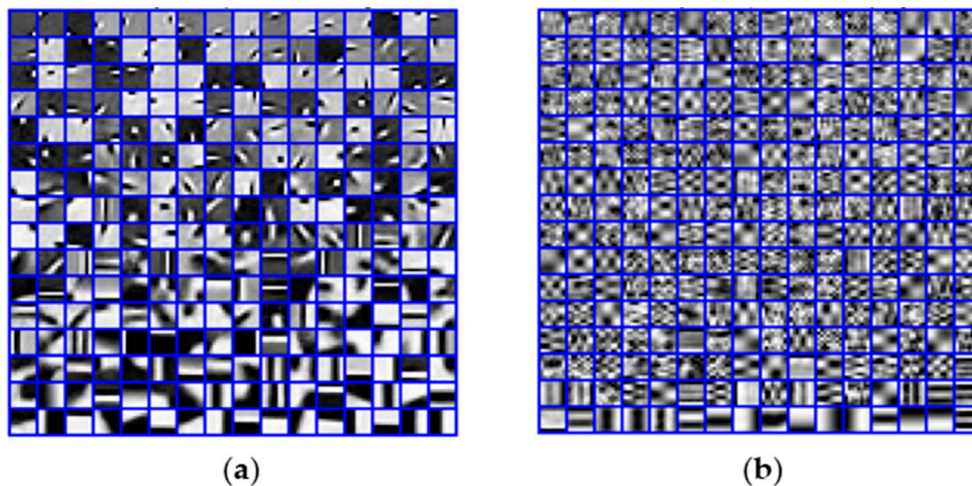


Figure 5. The random collections of 16×16 atoms ($K = 256$) of trained dictionary from (a) a reference set of 3245 ultrasound images and (b) a noisy image.

The numerical evaluation was performed using PSNR and MSSIM (as discussed in Section 3.1) on the proposed algorithm and compared with the denoising methods Frost filter [10], Lee filter [11], 3×3 Wiener filter [13], Kuan filter [14], 3×3 median filter [15], SRAD filter [39], and wavelet filter [40].

Figure 6a, shows a right lobe liver image with size 256×256 pixels, where the lateral size is given by the x -axis, and the axial size is given by the y -axis. In this original image, we included a speckle noise parameter $\sigma = 10$ and the PSNR was calculated using Equation (12). It is clear that detailed information of the image is highly distorted, as shown in Figure 6b with a PSNR value of 28.148 dB. Figure 6c,d show the denoising results obtained by the proposed method using Dictionary 1 with a PSNR value of 35.033 dB and Dictionary 2 with a PSNR value of 35.537 dB. It is clear that the SR over learned dictionaries improves both edges and smooth features by eliminating the noise and reconstructs the image as much closer to the original image, as shown in Figure 6a.

Figure 7 shows the comparative experimental results obtained on real-time ultrasound images. For this experiment, we obtained a 256×256 -pixel liver image of a healthy person with a PSNR value of 24.6271 dB. The radio frequency (RF) frames were obtained using a linear transducer with a frequency range of 8 MHz. This frequency range was selected because of its suitability for liver imaging, and we considered natural speckle noise for these experiments. The original speckled image was then denoised using the proposed algorithm with both dictionaries and also using conventional algorithms. To assess the speckle reduction, we selected two regions in of the speckled image. The two

regions in the case of Figure 7a are displayed as a red square and a green square. The red one indicates the diaphragm of a liver and the green square shows the presence of an excessive noisy region observed from deeper tissue. The differences can be noticed from the filtered images in dashed red and the green square. Figure 7d–f show that detailed information lost by the blurring effect on the results obtained with Frost filter, median filter, and Kuan filter. In particular, the wavelet filter, Weiner filter, and the SRAD filter are not very effective in reducing speckle and perform poorly in retrieving sharp edge information, as can be seen in Figure 7g–i. Figure 7b shows the results for the proposed method using Dictionary 1 (PSNR = 30.3345 dB) and Figure 7c shows the results for the proposed method using Dictionary 2 (PSNR = 30.8073). It is clear that the image denoised using the proposed SR method reconstructed image very close to the original image. It can also be seen that the dictionary trained on the noisy image gives better results than using a set of multiple references images. The results of this comparative experiment show that the proposed algorithm not only reduces the speckle noise but also preserves the edge information. Table 2 shows the PSNR and MSSIM values to quantify the results numerically for noise parameter $\sigma = 15$.

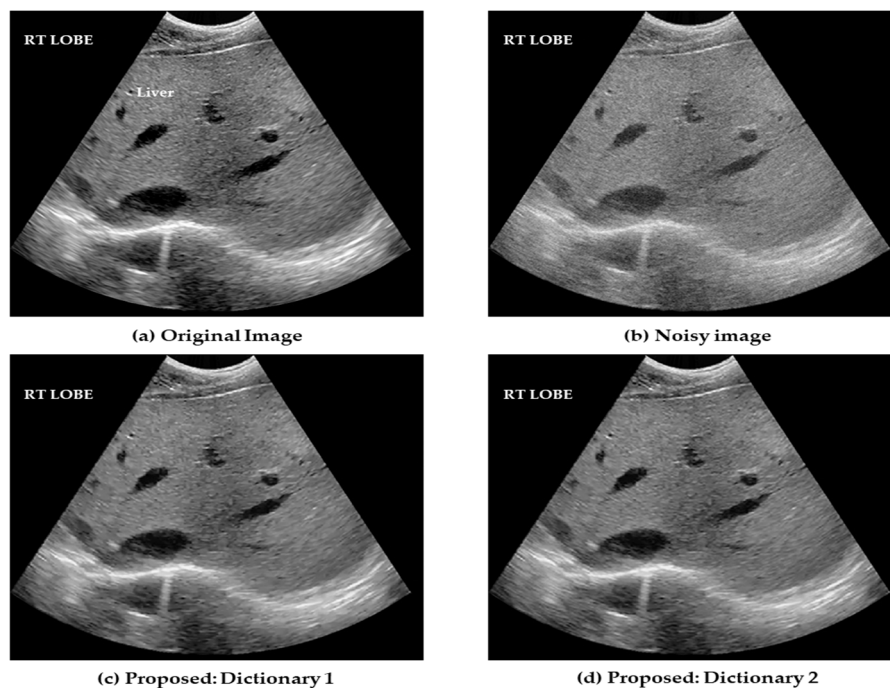


Figure 6. Reconstruction of liver right lobe images. (a) Original ultrasound image; (b) Speckled ultrasound image (PSNR = 28.148 dB); Images reconstructed using (c) Dictionary 1 (PSNR = 35.033 dB) and (d) Dictionary 2 (PSNR = 35.537 dB).

Table 2. PSNR and MSSIM for the ultrasound liver image for $\sigma = 15$.

Models	PSNR (dB)	MSSIM
Frost	28.966	0.822
Median	25.497	0.659
Wavelet	27.772	0.782
SRAD	28.766	0.813
Kuan	28.279	0.801
Weiner	29.218	0.834
Proposed: Dictionary 1	30.334	0.901
Proposed: Dictionary 2	30.807	0.926

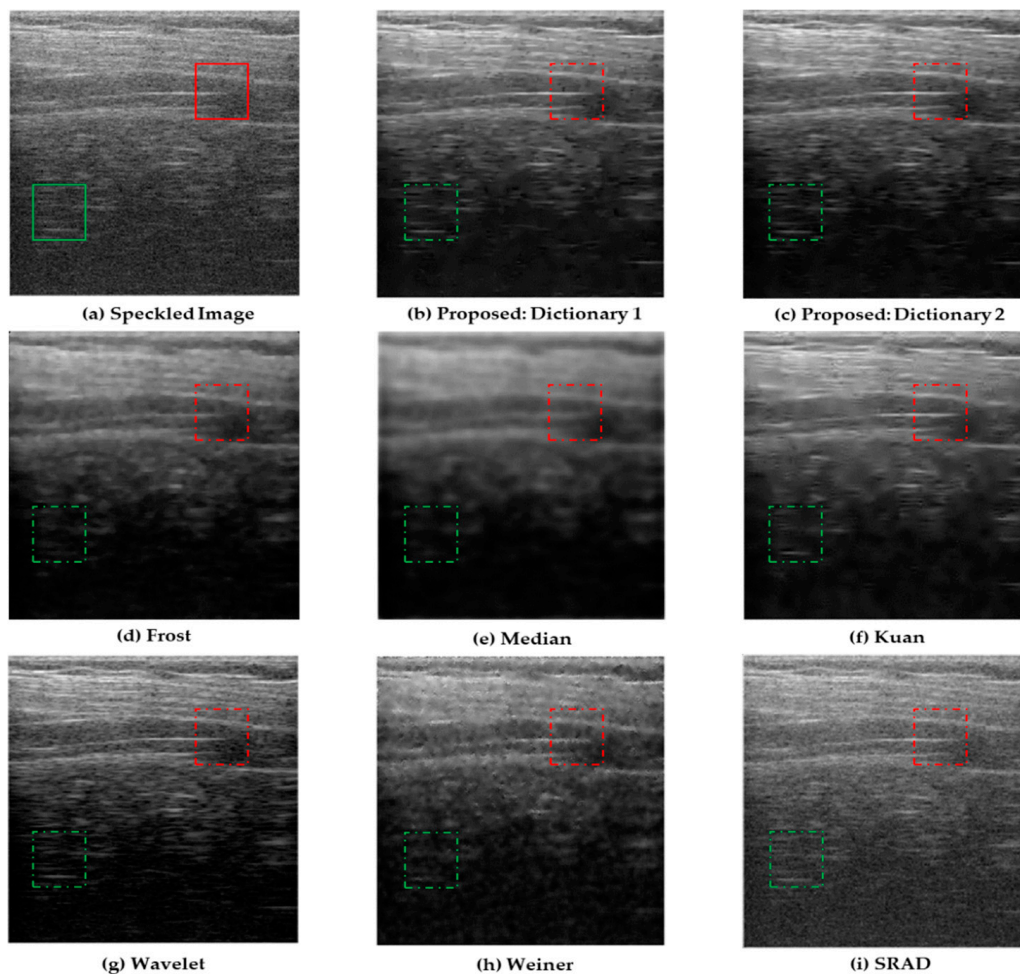


Figure 7. Despeckled results obtained for the ultrasound liver dataset using a linear transducer with a frequency of 8 MHz. The red and the green boxes highlight the differences observed from the noisy and filtered images. (a) Speckled image and results yielded by the proposed method using (b) Dictionary 1 and (c) Dictionary 2 as well as results using the (d) Frost; (e) median; (f) Kuan; (g) wavelet; (h) Weiner; and (i) SRAD filters.

Speckle is an arbitrary granular texture noise that degrades ultrasound image quality. This experiment was performed to evaluate different noise variances by comparing the PSNR obtained using the proposed algorithm and other despeckling algorithms. The simulated result using the noise levels 10, 15, 20, 25, and 30 are illustrated in Figure 8. The results clearly depict that, for different noise variances, the proposed algorithm gives the best PSNR value of all the algorithms on speckle reduction.

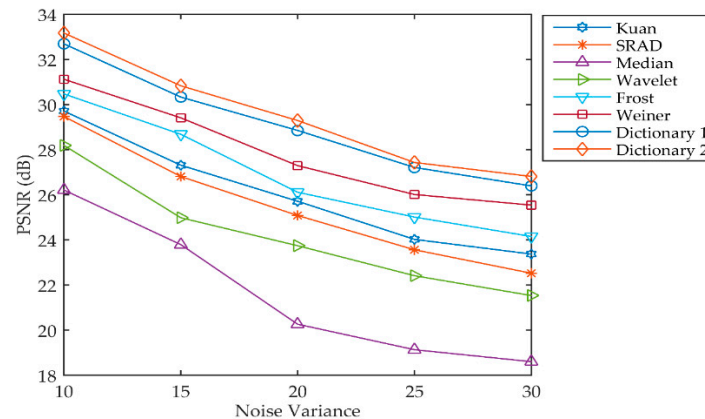


Figure 8. Comparison of PSNRs obtained by different methods. SRAD: speckle reducing anisotropic diffusion.

The experiments presented above were performed on ultrasound liver images, and the performance compared with conventional methods. However, our algorithm can also be utilized for a wide range of ultrasound images. To prove this, we conducted experiments on a real thrombus (blood clot) image with a left ventricular mass [56]. The visual assessment was performed using the proposed technique and the results compared to those obtained by various other algorithms. The reference image size was 256×256 pixels in order to fit our proposed model. The data were obtained from an open medical imaging dataset on GitHub [57]. The ultrasound image along with a marked note are shown in Figure 9a. The dashed white box in Figure 9b–j indicate regions of the ventricular mass. The thrombus data-set results presented in Figure 9h–j show that the wavelet, Weiner, and SRAD filters performed very poorly in noise reduction. The difference can be seen from the white note marked on the right atrium of the reference ultrasound image in Figure 9a. Figure 9e–g shows that Frost, median, and Kuan reduces speckle but tends to over-smooth the image, which leads to the loss of a distinctive feature of the unclear mass. Among all the methods, Figure 9c,d show good results for the SR-based on learned dictionaries 1 and 2. Several details are well preserved and the speckle noise is reduced efficiently. Figure 10 shows the zoomed sub-images of Figure 9 to observe a clear visualization of the despeckled images. The red box highlights the texture details in the noisy image and the filtered image for a comparative visual assessment. It can be noted that from the Frost, Median, and Kuan filtered data displayed in Figure 10d–f, an unclear mass (blood clot) and texture feature are blurred and over smoothed. Figure 10h,i show that the Weiner and SRAD filters are not much more effective on speckle reduction. These filters also greatly reduce the contrast, making images more indistinguishable from the background. This effect is especially noticeable in the case of the Wavelet filter as shown in Figure 10g. It was found that the anatomical structure was more clearly visible in Figure 10b,c obtained using the SR framework, where the speckle is reduced around the unclear mass without removing its features such as edges and texture. These results were comparatively better than those of Figure 10d–i of the standard despeckling methods. Thus, the proposed algorithm has various advantages for use in CAD systems based on image analysis, such as segmentation and edge detection. Future work will include extensive laboratory and clinical testing on diseased and healthy subjects for a more rigorous validation of the system. In conclusion, our approach reconstructs the detailed information in real ultrasound images, not only by preserving edge information but also by eliminating artifacts and reducing speckle noise.

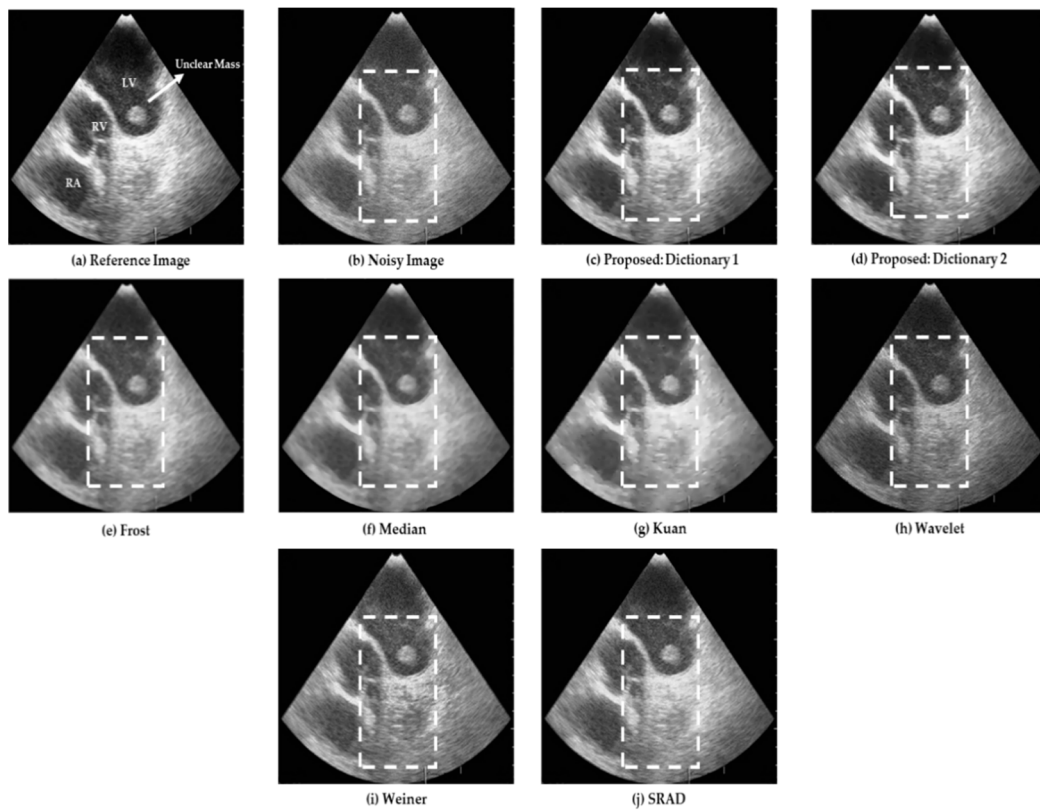


Figure 9. (a) Ultrasound image of the thrombus in the left ventricle. LV: left ventricle, RA: right atrium and RV: right ventricle and (b) noisy image. Despeckled ultrasound images of proposed method using (c) Dictionary 1 and (d) Dictionary 2. Results using the (e) Frost, (f) median, (g) Kuan, (h) wavelet, (i) Weiner, and (j) SRAD filters. The dashed white box indicates the region of image showing visual enhancement owing to despeckling.

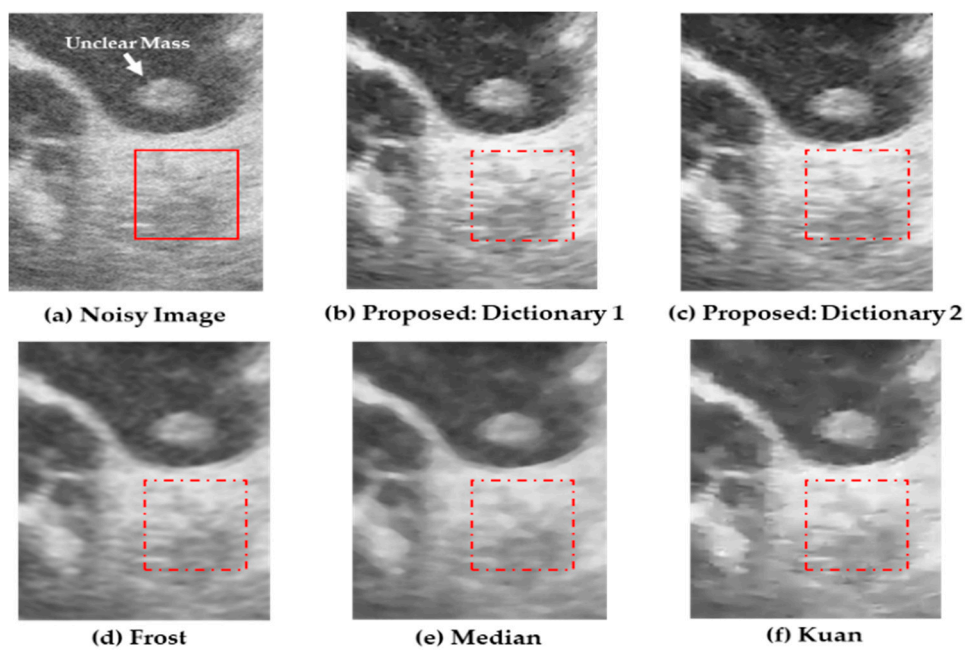


Figure 10. Cont.

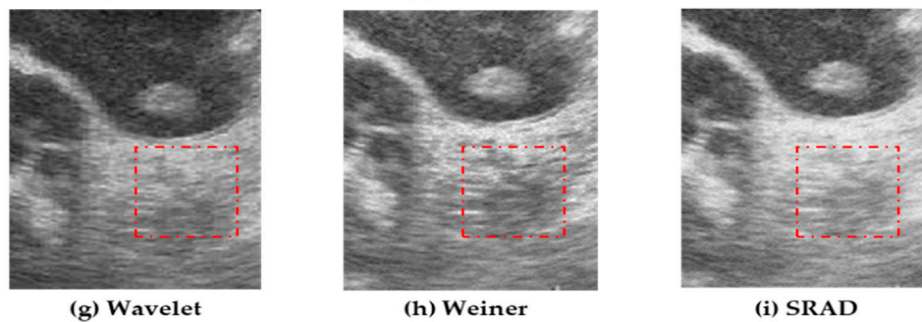


Figure 10. (a) Zoomed sub-image of noisy thrombus ultrasound images. The red boxes highlight texture details of images for visual assessment. Results of proposed method using (b) Dictionary 1 and (c) Dictionary 2. Results using the (d) Frost; (e) median; (f) Kuan; (g) wavelet; (h) Wiener; and (i) SRAD filters.

5. Conclusions

In this paper, we presented a method that reconstructed ultrasound images by suppressing multiplicative speckle noise using the SR framework. The proposed method utilizes an enhanced homomorphic filter, TV regularization, and sparse prior over two learned dictionaries. In addition, the KSVD algorithm is used to train the two dictionaries—one trained with a set of reference ultrasound image patches and another trained with the speckled image patches. Both training options were tested with the synthetic images and various clinical ultrasound images. The experimental results obtained for different noise levels proved superior to those of other standard denoising methods. The results also show that the two modified dictionaries performed well with sparse and TV regularization terms. Overall, the proposed SR framework reconstructs the image signals by removing speckle noise while preserving the texture and yielding a smoother image than conventional methods without eliminating edges.

Author Contributions: M.Y.J. and H.N.L. formulated and designed the experiments; M.Y.J. performed the experiments and analyzed the data; M.Y.J. wrote the manuscript; H.N.L. made revisions to the manuscript and supervised the research work.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant provided by the Korean government (MSIP) [NRF-2018R1A2A1A19018665].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Szabo, T.L. *Diagnostic Ultrasound Imaging: Inside Out*; Academic Press Series in Biomedical Engineering; Elsevier Academic Press: New York, NY, USA, 2004; p. 549. ISBN 0-12-680145-2.
2. Martial, B.; Cachar, D. Acquire real-time RF digital ultrasound data from a commercial scanner. *Electron. J. Tech. Acoust.* **2007**, *3*, 16.
3. Lee, H.; Chen, Y.P.P. Image based computer aided diagnosis system for cancer detection. *Expert Syst. Appl.* **2015**, *42*, 5356–5365. [[CrossRef](#)]
4. Jabarulla, M.Y.; Lee, H.N. Computer aided diagnostic system for ultrasound liver images: A systematic review. *Optik* **2017**, *140*, 1114–1126. [[CrossRef](#)]
5. Zanotel, M.; Bednarova, I.; Londero, V.; Linda, A.; Lorenzon, M.; Girometti, R.; Zuiani, C. Automated breast ultrasound: Basic principles and emerging clinical applications. *Radiol. Med.* **2018**, *123*, 1–12. [[CrossRef](#)] [[PubMed](#)]
6. Acharya, U.R.; Koh, J.E.W.; Hagiwara, Y.; Tan, J.H.; Gertych, A.; Vijayanathan, A.; Yaakup, N.A.; Abdullah, B.J.J.; Fabell, M.K.B.M.; Yeong, C.H. Automated diagnosis of focal liver lesions using bidirectional empirical mode decomposition features. *Comput. Biol. Med.* **2017**, *94*, 11–18. [[CrossRef](#)] [[PubMed](#)]

7. Grazioli, L.; Ambrosini, R.; Frittoli, B.; Grazioli, M.; Morone, M. Primary benign liver lesions: Benign focal liver lesions can origin from all kind of liver cells: Hepatocytes, mesenchymal and cholangiocellular line. *Eur. J. Radiol.* **2017**, *26*, 378–398. [[CrossRef](#)] [[PubMed](#)]
8. Burckhart, C.B. Speckle in ultrasound B-mode scans. *IEEE Trans. Sonics Ultrason.* **1978**, *25*, 1–6. [[CrossRef](#)]
9. Narayanan, S. A view on despeckling in ultrasound imaging. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2009**, *2*, 85–98.
10. Lopes, A.; Touzi, R.; Nezry, E. Adaptive Speckle Filters and Scene Heterogeneity. *IEEE Trans. Geosci. Remote Sens.* **1990**, *28*, 992–1000. [[CrossRef](#)]
11. Lee, J.S. Digital image enhancement and noise filtering by use of local statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* **1980**, *2*, 165–168. [[CrossRef](#)] [[PubMed](#)]
12. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 3rd ed.; Pearson Education, Inc.: London, UK, 2008; ISBN 0-13-168728-x/978-0-13-168728-8.
13. Goldstein, J.S.; Reed, I.S.; Scharf, L.L. A multistage representation of the Wiener filter based on orthogonal projections. *IEEE Trans. Inf. Theory* **1998**, *44*, 2943–2959. [[CrossRef](#)]
14. Kuan, D.T.; Sawchuk, A.A.; Strand, T.C.; Chavel, P. Adaptive noise smoothing filter for images with signal-dependent noise. *IEEE Trans. Pattern Anal. Mach. Intell.* **1985**, *7*, 165–177. [[CrossRef](#)] [[PubMed](#)]
15. Simon, P.; Patrick, H. Median Filtering in Constant Time. *IEEE Trans. Image Process.* **2007**, *16*, 2389–2394.
16. Achim, A.; Bezerianos, A.; Tsakalides, P. Novel Bayesian multiscale method for speckle removal in medical ultrasound images. *IEEE Trans. Med. Imaging* **2001**, *20*, 772–783. [[CrossRef](#)] [[PubMed](#)]
17. Chen, Z.J.; Chen, C.H.Y. Efficient statistical modeling of wavelet coefficients for image denoising. *Int. J. Wavelets Multiresolut. Inf. Process.* **2009**, *7*, 629–641. [[CrossRef](#)]
18. Vishwa, A.; Sharma, S. Modified method for denoising the ultrasound images by wavelet thresholding. *Int. J. Intell. Syst. Appl.* **2012**, *4*, 25. [[CrossRef](#)]
19. Shen, Y.; Liu, Q.; Lou, S.; Hou, Y.L. Wavelet-Based Total Variation and Nonlocal Similarity Model for Image Denoising. *IEEE Signal Process. Lett.* **2017**, *24*, 877–881. [[CrossRef](#)]
20. Donoho, D.L. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **1995**, *41*, 613–627. [[CrossRef](#)]
21. Matsuyama, E.; Tsai, D.-Y.; Lee, Y.; Tsurumaki, M.; Takahashi, N.; Watanabe, H.; Chen, H.-M. A modified undecimated discrete wavelet transform based approach to mammographic image denoising. *J. Digit. Imaging* **2013**, *26*, 748–758. [[CrossRef](#)] [[PubMed](#)]
22. Kim, Y.S. Improvement of ultrasound image based on wavelet transform: Speckle reduction and edge enhancement. *SPIE Med. Imaging* **2005**, 5747, 1085–1092.
23. Chambolle, A. An algorithm for total variation minimizations and applications. *J. Math. Imaging Vis.* **2004**, *10*, 89–97.
24. Rudin, L.I.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.* **1992**, *60*, 259–268. [[CrossRef](#)]
25. Perona, P.; Malik, J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 629–639. [[CrossRef](#)]
26. Chao, S.M.; Tsai, D.M. An improved anisotropic diffusion model for detail and edge-preserving smoothing. *Pattern Recognit. Lett.* **2010**, *31*, 2012–2023. [[CrossRef](#)]
27. Tschumperle, D.; Deriche, R. Vector-valued image regularization with PDEs: A common framework for different applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 506–517. [[CrossRef](#)] [[PubMed](#)]
28. Zhao, Y.; Yang, J. Hyperspectral image denoising via sparse representation and low-rank constraint. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 296–308. [[CrossRef](#)]
29. Elad, M.; Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries in wavelet domain. *IEEE Trans. Image Process.* **2006**, *15*, 754–758. [[CrossRef](#)]
30. Deka, B.; Bora, P.K. Removal of correlated speckle noise using sparse and overcomplete representations. *Biomed. Signal Process. Control* **2013**, *8*, 520–533. [[CrossRef](#)]
31. Fan, J.; Wu, Y.; Li, M.; Liang, W.; Zhang, Q. SAR Image Registration Using Multiscale Image Patch Features with Sparse Representation. *Biomed. Signal Process. Control* **2017**, *10*, 1483–1493. [[CrossRef](#)]
32. Wright, J.; Ma, Y.; Mairal, J.; Sapiro, G.; Huang, T.S.; Yan, S. Sparse representation for computer vision and pattern recognition. *Proc. IEEE* **2010**, *98*, 1031–1044. [[CrossRef](#)]
33. Bruckstein, M.E.A.M.; Donoho, D.L.; Elad, M. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.* **2009**, *51*, 34–81. [[CrossRef](#)]

34. Li, S.; Wang, G.; Zhao, X. Multiplicative noise removal via adaptive learned dictionaries and TV regularization. *Digit. Signal Process.* **2016**, *50*, 218–228.
35. Liu, K.; Tan, J.; Su, B. An Adaptive Image Denoising Model Based on Tikhonov and TV Regularizations. *Adv. Multimed.* **2014**, *2014*, 934834. [[CrossRef](#)]
36. Aharon, M.; Elad, M.; Bruckstein, A.M. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322. [[CrossRef](#)]
37. Tay, P.C.; Garson, C.D.; Acton, S.T.; Hossack, J.A. Ultrasound despeckling for contrast enhancement. *IEEE Trans. Image Process.* **2010**, *19*, 1847–1860. [[CrossRef](#)] [[PubMed](#)]
38. Joel, T.; Sivakumar, R. An extensive review on Despeckling of medical ultrasound images using various transformation techniques. *Appl. Acoust.* **2018**, *138*, 18–27. [[CrossRef](#)]
39. Youngjian, Y.; Acton, S.T. Speckle reducing anisotropic diffusion. *IEEE Trans. Image Process.* **2002**, *11*, 1260–1270. [[CrossRef](#)] [[PubMed](#)]
40. Hussain, S.A.; Gorashi, S.M. Image Denoising based on Spatial/Wavelet Filter using Hybrid Thresholding Function. *Int. J. Comput. Appl.* **2012**, *42*, 5–13.
41. Aubert, G.; Aujol, J.-F. A variational approach to removing multiplicative noise. *SIAM J. Appl. Math.* **2008**, *68*, 925–946. [[CrossRef](#)]
42. Buades, A.; Coll, B.; Morel, J.M. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **2005**, *4*, 490–530. [[CrossRef](#)]
43. Gilboa, S.O.G. Nonlocal operators with applications to image processing. *SIAM J. Multiscale Model. Simul.* **2008**, *7*, 1005–1028. [[CrossRef](#)]
44. Cai, T.T.; Wang, L. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inf. Theory* **2011**, *57*, 4680–4688. [[CrossRef](#)]
45. Deka, B.; Bora, P.K. Despeckling of medical ultrasound images using sparse representation. In Proceedings of the 2010 International Conference Signal Processing and Communications (SPCOM), Bangalore, India, 18–21 July 2010. ISSN 2165-0608.
46. Cobbold, R.S.C. *Foundations of Biomedical Ultrasound*; Oxford University Press: Oxford, UK, 2007.
47. Yahya, N.; Kamel, N.S.; Malik, A.S. Subspace-based technique for speckle noise reduction in ultrasound images. *Biomed. Eng. Online* **2014**, *13*, 154. [[CrossRef](#)] [[PubMed](#)]
48. Arsenault, H.H.; Levesque, M. Combined homomorphic and local-statistics processing for restoration of images degraded by signal-dependent noise. *Appl. Opt.* **1984**, *23*, 845–850. [[CrossRef](#)] [[PubMed](#)]
49. Xie, H.; Pierce, L.E.; Ulaby, F.T. Statistical properties of logarithmically transformed speckle. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 721–727. [[CrossRef](#)]
50. Candes, E.; Candes, E.; Romberg, J.; Romberg, J. *11-Magic: Recovery of Sparse Signals via Convex Programming*; Caltech: Pasadena, CA, USA, 2005; pp. 1–19.
51. Afonso, M.V.; Bioucas-Dias, J.M.; Figueiredo, M.A.T. An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE Trans. Image Process.* **2011**, *20*, 681–695. [[CrossRef](#)] [[PubMed](#)]
52. Davis, G.; Mallat, S.G.; Avellaneda, M. Adaptive greedy approximations. *Constr. Approx.* **1997**, *13*, 57–98. [[CrossRef](#)]
53. Xiang, F.; Wang, Z. Split Bregman iteration solution for sparse optimization in image restoration. *Optik* **2014**, *125*, 5635–5640. [[CrossRef](#)]
54. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *4*, 600–612. [[CrossRef](#)]
55. Shepp, L.; Logan, F. The Fourier reconstruction of a head section. *IEEE Trans. Nucl. Sci.* **1974**, *21*, 21–43. [[CrossRef](#)]
56. Llach, F. Hypercoagulability, renal vein thrombosis, and other thrombotic complications of nephrotic syndrome. *Kidney Int.* **1985**, *3*, 429–439. [[CrossRef](#)]
57. GitHub. Available online: <https://github.com/sifikas/medical-imaging-datasets> (accessed on 1 May 2018).





Depth-estimation-enabled compound eyes

Woong-Bi Lee, Heung-No Lee *

Gwangju Institute of Science and Technology (GIST), School of Electrical Engineering and Computer Science, Gwangju, 61005, Republic of Korea



ARTICLE INFO

Keywords:

Computational imaging
Three-dimensional imaging
Compound eyes

ABSTRACT

Most animals that have compound eyes determine object distances by using monocular cues, especially motion parallax. In artificial compound eye imaging systems inspired by natural compound eyes, object depths are typically estimated by measuring optic flow; however, this requires mechanical movement of the compound eyes or additional acquisition time. In this paper, we propose a method for estimating object depths in a monocular compound eye imaging system based on the computational compound eye (COMPU-EYE) framework. In the COMPU-EYE system, acceptance angles are considerably larger than interommatidial angles, causing overlap between the ommatidial receptive fields. In the proposed depth estimation technique, the disparities between these receptive fields are used to determine object distances. We demonstrate that the proposed depth estimation technique can estimate the distances of multiple objects.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Compound eyes, such as those of arthropods, have attracted widespread research interest owing to their unique features – such as wide fields of view (FOVs), excellent motion detection capability, and sensitivity to light intensity – that indicate their great potential for use in numerous applications, including unmanned aerial vehicles and endoscopic medical tools [1–4]. Recently, cameras inspired by compound eyes found in nature have been developed using curved optics and electronics [5,6] and discrete component integration at macroscopic levels [7].

Visual methods for depth estimation can be grouped into two main categories based on whether they use binocular or monocular cues [8]. Binocular cues are obtained from the minor disparities between the views of two eyes when the eyes are located close to one another and have overlapping views. These slightly different images of the same scene are sent to the brain and integrated into a single image containing depth information [9]. By contrast, monocular cues are obtained from two-dimensional images captured by a single eye; these cues include interposition, motion parallax, relative size and clarity, texture gradient, linear perspective, and light and shadow [8].

Some insects, such as praying mantids, that have binocular vision systems in the fronts of their heads use binocular cues to estimate target distances [9,10]. However, unlike humans' camera-like eyes that can focus on objects by changing the shapes or positions of their lenses,

insects' compound eyes are inherently immobile and unable to focus owing to their structural limitations [8]. Thus, the binocular cues used for depth estimation in compound eyes are much less efficient, yielding images with low spatial resolutions and limited effective depth estimation ranges [11,12].

Insects can also estimate object distances using monocular cues. The motion parallax of objects in a visual scene that is caused by the relative motion between the observer and the objects yields information about object distances [8,13]. Specifically, nearby objects produce more apparent motion than distant ones. Insects' visual systems can easily detect the depths of objects that move independently of their surroundings by using motion parallax. For example, grasshoppers judge depths accurately by using the motion parallax generated by peering movements, that is, by moving their head from side to side [9], and bees measure distances by monitoring the apparent motion of an object relative to its surroundings [14].

Recently, artificial compound eyes that mimic natural compound eyes have been proposed. In these eyes, each ommatidium (individual imaging unit) has a limited acceptance angle, thus avoiding optical crosstalk among neighboring ommatidia [5–7,13]. In [6,13], object depths were estimated using monocular cues from optic flows (i.e., pattern of apparent motion) based on the phenomenon in which a closer object appears to move faster than a farther one. However, this method requires rotation or movement of the compound eye.

* Correspondence to: #C317, School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Republic of Korea.

E-mail addresses: wblee@gist.ac.kr (W.-B. Lee), heungno@gist.ac.kr (H.-N. Lee).

<https://doi.org/10.1016/j.optcom.2017.12.009>

Received 22 September 2017; Received in revised form 19 November 2017; Accepted 3 December 2017
0030-4018/© 2017 Elsevier B.V. All rights reserved.

In this paper, we propose a method for estimating object depths in a monocular compound eye imaging system based on the computational compound eye (COMPU-EYE) framework described in [15]. In COMPU-EYE, each ommatidium has a larger acceptance angle than its interommatidial angle, causing the ommatidial receptive fields to overlap significantly. As in binocular depth estimation methods, depth estimation in COMPU-EYE involves processing the multiple and slightly differing views received by the ommatidia by using a proposed digital signal processing (DSP) technique. Depth information can be estimated by using the dependences of the disparities between the ommatidial observations on object distance. We perform a numerical experiment to verify the effectiveness of the proposed method. In our experiment, we demonstrate that the proposed depth estimation technique can not only estimate the distances of multiple objects but also reconstruct object images with high resolution.

Depth estimation using the disparities between multiple subimages has been studied in multicamera systems such as integral imaging [16]. Integral imaging is a three-dimensional imaging and sensing system that uses an array of optical units. Each optical unit consists of a microlens and an array of photosensors, and it produces an elemental image. From multiple elemental images, a three-dimensional image is reconstructed optically or computationally [17]. In [18], an iterative reconstruction algorithm was proposed for improving image quality given distance information. A stereo matching method that used the spatial variations of parallax shifts in elemental images was proposed for depth estimation [2,19]. We note that multicamera setups are essentially different from our work. First, our structure can be considered a degraded integral imaging system with a single photosensor in each elemental image; this imitates the structure of apposition compound eyes found in nature. The number of sensors is thus reduced dramatically, and the sensors can be implemented in a fully hemispherical structure that provides a large FOV [5]. Some studies on integral imaging considered curved surfaces for realizing a large FOV [20]. However, with planar sensors, they require additional optical components like random phase masks; otherwise, mismatch occurs [20]. Second, three-dimensional information is highly compressed using a single photosensor per lens. Thus, more sophisticated reconstruction algorithms are required for imaging and depth estimation.

In Section 2, we describe the COMPU-EYE system model and the principle of depth estimation. In Section 3, we propose our depth estimation method, and in Section 4, we discuss the results. Finally, we present our conclusions in Section 5.

2. COMPU-EYE system model and depth estimation

2.1. COMPU-EYE system model

We consider an apposition compound eye imaging system in which a hemispherical eye observes a planar object. The hemispherical compound eye can be implemented by reformulating a stretchable set of a microlens and photodetector array [5]. As a result, this compound eye has a large FOV. This compound eye consists of a two-dimensional array of M ommatidia that are uniformly spaced with an interommatidial angle of $\Delta\phi$. As illustrated in Fig. 1(a), each ommatidium receives incident light within its acceptance angle $\Delta\phi$. Based on the object's location, each observation at each ommatidium can be specified by a transfer function that describes the fraction of the input light that each ommatidium observes. We assume that the object is located a distance d (measured in millimeters) away from the compound eye and that the image to be reconstructed consists of N pixels that form an $N \times 1$ input vector $\mathbf{x} = [x_1, \dots, x_N]^T$ in lexicographic order. Let y_i denote the output sample obtained by a photodetector at the i th ommatidium for $i \in \{1, 2, \dots, M\}$. Through ray tracing analysis, y_i can be obtained using the linear equation $y_i = a_{i,d}x$, where $a_{i,d}$ is a $1 \times N$ vector whose elements represent the visibility of the i th ommatidium at each of the N pixels of the object located at a distance of d [15]. Given

the structure of the compound eye, specifically, the acceptance angles, interommatidial angles, and sizes of the compound eye and ommatidia, the receptive fields of ommatidia at a distance of d are determined. Each element of $a_{i,d}$ is obtained by calculating the intersection area of the receptive field of the i th ommatidium and the j th pixel in the object for $j \in \{1, 2, \dots, N\}$. The data acquisition model for M ommatidial observations can be expressed as a system of linear equations as follows:

$$\mathbf{y} = \mathbf{A}_d \mathbf{x} + \mathbf{n}, \quad (1)$$

where $\mathbf{y} = [y_1, \dots, y_M]^T$ is a set of M output samples, $\mathbf{A}_d \in \mathbb{R}^{M \times N}$ denotes a measurement matrix whose i th row is $a_{i,d}$, and \mathbf{n} is an $M \times 1$ noise vector.

A signal is typically considered sparse if it can be represented with few nonzero elements. We note that any natural image can be represented as a sparse signal in a certain domain, such as by applying a wavelet, discrete cosine, or discrete Fourier transform [21]. That is, $\mathbf{x} = \mathbf{w}^T \mathbf{s}$ and $\mathbf{w}\mathbf{x} = \mathbf{s}$, where \mathbf{s} is a sparse $N \times 1$ vector and \mathbf{w} is an $N \times N$ sparsifying matrix. By exploiting the sparse representation of \mathbf{x} , Eq. (1) can be expressed as

$$\mathbf{y} = \mathbf{A}_d \mathbf{w}^T \mathbf{s} + \mathbf{n}. \quad (2)$$

To obtain sufficiently high resolution, the number of pixels to be reconstructed is set to be larger than the number of ommatidia, that is, $N > M$. Then, Eq. (2) becomes an underdetermined system of linear equations. Given \mathbf{A}_d and \mathbf{y} , \mathbf{s} can be obtained by solving the following convex optimization problem [22]:

$$\hat{\mathbf{s}} = \min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{A}_d \mathbf{w}^T \mathbf{s}\|_2 < \epsilon, \quad (3)$$

where ϵ is a small constant. From $\hat{\mathbf{s}}$, the object image can be reconstructed by solving $\hat{\mathbf{x}} = \mathbf{w}^T \hat{\mathbf{s}}$.

2.2. Distance and measurement matrix

The COMPU-EYE imaging system proposed in [15] yields resolution improvements beyond the number of ommatidia owing to its use of large ommatidial acceptance angles in combination with a DSP technique. The large acceptance angles enable each pixel to be observed multiple times by multiple ommatidia with different perspectives. However, these ommatidial observations are severely distorted owing to the overlap in the ommatidial receptive fields. Given a measurement matrix, DSP can be used to reconstruct high-resolution images from the distorted observations by solving the underdetermined linear system in Eq. (1). The measurement matrix strongly depends on the object's properties, such as its distance. In [15], the object distance was assumed to be fixed and known, and the measurement matrix corresponding to this distance was given to the DSP system. However, assuming prior knowledge about object distances is impractical in reality. The reconstruction process works well only if the measurement matrix is correct; if an inappropriate measurement matrix is used, then the reconstructed image is severely distorted.

In the framework of COMPU-EYE imaging, we propose a new depth estimation method. In conventional compound eyes, $\Delta\phi$ is designed to be smaller than or equal to $\Delta\phi$ to avoid aliasing [5,6,23]. As shown in Fig. 1(a), each ommatidium observes an independent section within $\Delta\phi$. Consider two objects, P_1 and P_2 , that are located at different distances from a compound eye. If the objects are observed by a single ommatidium in Fig. 1(a), their distances cannot be inferred. In contrast, the COMPU-EYE system has enlarged, overlapping ommatidial receptive fields, because $\Delta\phi$ is much larger than $\Delta\phi$, as seen in Fig. 1(b). We note that a large acceptance angle can be realized by increasing the diameter of the photodetector, decreasing the focal length of the microlens, or using a material of higher refractive index for the microlens [15]. This configuration is shown in Fig. 1(b), in which object P_2 is observed by two ommatidia; thus, the compound eye can deduce that object P_2 is farther away than object P_1 . When many ommatidia are present, the number of

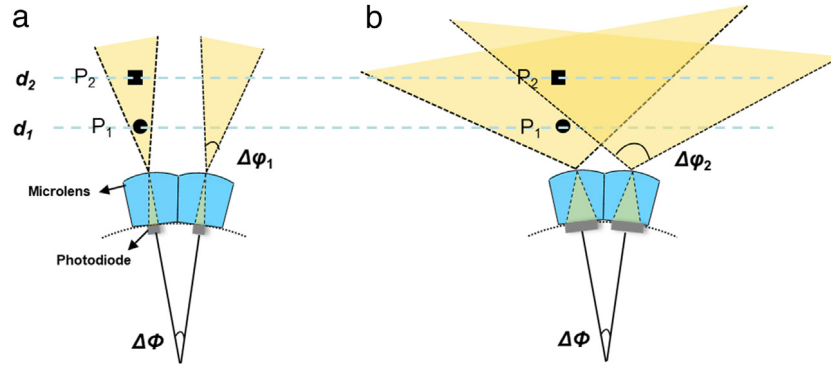


Fig. 1. Structures and fields of view of (a) Conventional compound eye with $\Delta\phi_1 \leq \Delta\phi$ and (b) Proposed COMPU-EYE system with $\Delta\phi_2 \gg \Delta\phi$.

ommatidia viewing the object and the area of the object that is visible by the ommatidia depend upon the object distance. The variation of these quantities with object distance is used for depth estimation in the proposed method.

Here, we give an example of the variation for different object distances. In Fig. 2, the measurement matrices and corresponding number of nonzero elements per column are shown, in which a compound eye consists of 5×5 ommatidia with a radius of 6.92 mm, focal length of micro lens of 1.35 mm, $\Delta\phi = 12^\circ$, and $\Delta\phi = 30^\circ$. Three objects are located at $d_1 = 2$ mm, $d_2 = 20$ mm, and $d_3 = 40$ mm from the compound eye. The object plane is composed of 12×12 pixels with a uniform distribution. As the object moves away from the compound eye, the areas of the ommatidial receptive fields and the overlap between them both increase. Accordingly, as shown in Fig. 2(a), the number of nonzero elements in the measurement matrix increases with object distance. In Fig. 2(b), the number of nonzero elements per column in the measurement matrices varies with respect to the object distances, implying that each pixel is uniquely observed by a different set of ommatidia with different perspectives. Thus, a unique measurement matrix is generated with respect to object distance. By using the relationship between the unique measurement matrix and the object distance, we propose the following method for estimating object distances.

2.3. System model for depth estimation

First, we set the range of interest $R = [d_{\min}, d_{\max}]$, where d_{\min} and d_{\max} are the minimum and maximum distances, respectively. The range of interest can be application-specific; for example, it can be 10–25 mm for endoscopic applications [24]. For DSP, we assume that the object distance can be sampled as a set of discrete distances $\mathbf{d} = \{d_1, d_2, \dots, d_L\}$ within the range of interest, where L is the number of distance elements. In this paper, we consider uniform discrete distances within the range of interest. The depth resolution $\Delta d = (d_{\max} - d_{\min})/L$ depends on the number of distance elements and depth range of interest. According to the predetermined \mathbf{d} , a measurement matrix \mathbf{A}_{d_l} for $l \in \{1, 2, \dots, L\}$ can be obtained from the structure of the compound eye and the object located a distance d_l away from the compound eye. By concatenating L measurement matrices, a dictionary matrix $\mathbf{A} \in \mathbb{R}^{M \times (L \cdot N)}$ can be formed as $\mathbf{A} = [\mathbf{A}_{d_1} \mathbf{A}_{d_2} \dots \mathbf{A}_{d_L}]$. Then, the linear representation of \mathbf{y} in Eq. (1) can be rewritten in terms of all possible measurement matrices as

$$\mathbf{y} = \sum_{i=1}^L \mathbf{A}_{d_i} \mathbf{x}_i = \mathbf{A}\mathbf{X}, \quad (4)$$

where $\mathbf{X} = [\mathbf{x}_1^T, \dots, \mathbf{x}_L^T]^T = [x_{1,1}, \dots, x_{1,N}, \dots, x_{L,1}, \dots, x_{L,N}]^T \in \mathbb{R}^{(L \cdot N) \times 1}$. When an object is located at a certain distance in the set \mathbf{d} , a valid observation \mathbf{y} can be sufficiently represented by a linear combination of the columns from the corresponding measurement matrix. For

example, when the object distance matches the l th measurement matrix, the linear equation becomes

$$\mathbf{y} = \mathbf{A}\mathbf{X}_0, \quad (5)$$

where $\mathbf{X}_0 = [0, \dots, 0, x_{l,1}, \dots, x_{l,N}, 0, \dots, 0]^T$ is a sparse coefficient vector whose entries are zero except for those associated with the l th measurement matrix. \mathbf{x} can be sparsely represented as $\mathbf{x} = \mathbf{w}^T \mathbf{s}$. Similarly, \mathbf{X} can be sparsely represented as $\mathbf{X} = \mathbf{W}^T \mathbf{S}$. Here, \mathbf{S} is an $L \cdot N \times 1$ sparse vector and \mathbf{W} is a block diagonal matrix containing L instances of \mathbf{w} , that is, $\mathbf{W} = \text{diag}(\underbrace{\mathbf{w}, \dots, \mathbf{w}}_L) \in \mathbb{R}^{(L \cdot N) \times (L \cdot N)}$, where $\text{diag}(\cdot)$ represents a diagonal matrix. By using \mathbf{S} , Eq. (4) becomes

$$\mathbf{y} = \mathbf{A}\mathbf{W}^T \mathbf{S} = \mathbf{B}\mathbf{S}, \quad (6)$$

where $\mathbf{S} = [s_1^T, \dots, s_L^T]^T = [s_{1,1}, \dots, s_{1,N}, \dots, s_{L,1}, \dots, s_{L,N}]^T \in \mathbb{R}^{(L \cdot N) \times 1}$ and $\mathbf{B} = \mathbf{A}\mathbf{W}^T$. As does Eq. (3), l_1 norm minimization provides a sparse vector $\hat{\mathbf{S}}$:

$$\hat{\mathbf{S}} = \arg \min_{\mathbf{S}} \|\mathbf{S}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{A}\mathbf{W}^T \mathbf{S}\| \leq \eta, \quad (7)$$

where η is a small constant.

3. Depth estimation method

After $\hat{\mathbf{S}}$ has been obtained from Eq. (7), the problem of estimating object distances can be reformulated as a classification problem whose objective is to find the distances at which the object has the highest probability of being located. Because the compound eye imaging system can be sparsely represented in Eq. (6) and the measurement matrices are uniquely generated with respect to object distances, sparse-representation-based classification (SRC) can be used to estimate object locations. SRC has been widely studied, and its accuracy has been demonstrated in many applications including face recognition [25] and brain computer interface systems [26]. SRC usually finds the most compact representation of a test sample, where the representation is expressed as a linear combination of columns in an overcomplete dictionary matrix, and then, it determines a class that contributes most to represent the test sample [27]. In this paper, we use SRC to estimate the depths of multiple objects. Unlike conventional SRC, the observed signal in this depth estimation framework is superposed with respect to the number of objects. Therefore, our problem is defined as a multiclass classification problem. We first describe an SRC-based depth estimation algorithm in the compound eye imaging system; we then propose an iterative depth estimation method that updates dictionaries in a coarse-to-fine manner.

We first specify a classification rule by using sparse signal reconstruction. As l_1 norm minimization provides a sparse solution for Eq. (7), most of the nonzero components in $\hat{\mathbf{S}}$ reside in the class in which the object exists with high probability. One of the classification rules is

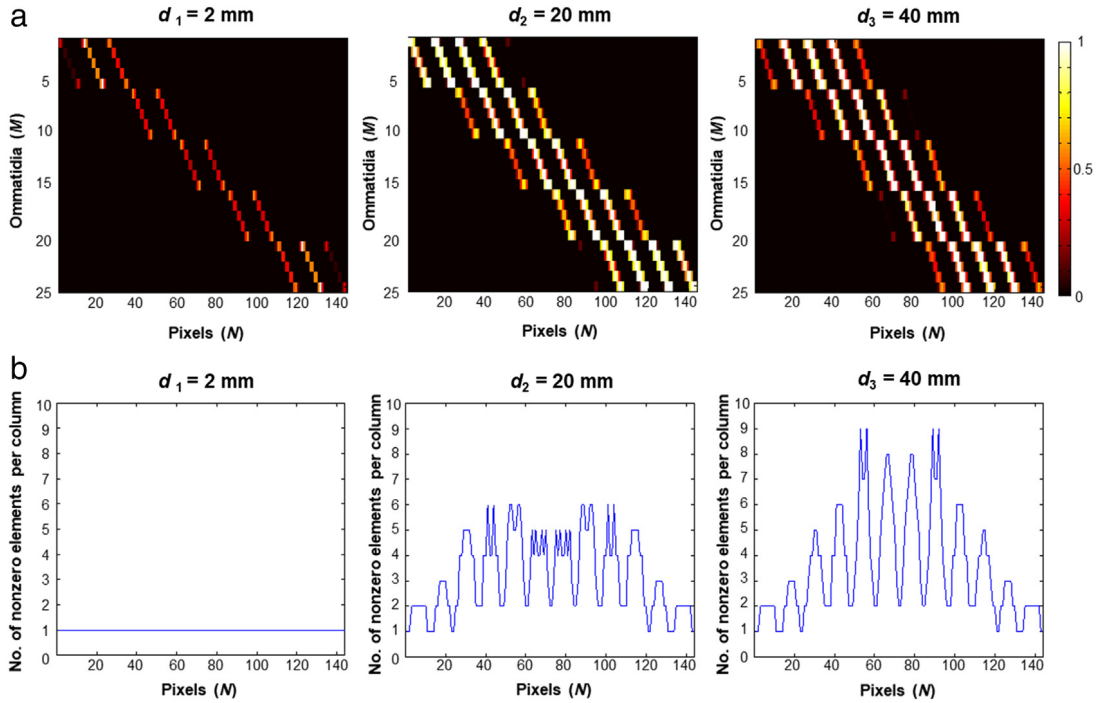


Fig. 2. (a) Measurement matrices and (b) their number of nonzero elements per column for $d_1 = 2$ mm, $d_2 = 20$ mm, and $d_3 = 40$ mm, where $M = 5 \times 5$, $N = 12 \times 12$, $\Delta\phi = 12^\circ$, and $\Delta\varphi = 30^\circ$.

Table 1

SRC-based depth estimation algorithm.

Initial parameters: $\mathbf{y}, \mathbf{d} = \{d_1, d_2, \dots, d_L\}, \mathbf{w}, \eta, \alpha$
Step 1: Set $\mathbf{A} = [\mathbf{A}_{d_1}, \mathbf{A}_{d_2}, \dots, \mathbf{A}_{d_L}]$ and $\mathbf{W} = \text{diag}(\underbrace{\mathbf{w}, \dots, \mathbf{w}}_L)$.
Step 2: Solve Eq. (7) from \mathbf{y} given \mathbf{A} and \mathbf{W} , and obtain $\hat{\mathbf{S}}$.
Step 3: Calculate the regularized residuals:
$r_l := \frac{\ \mathbf{y} - \mathbf{B}\delta_l(\hat{\mathbf{S}})\ ^2}{\ \delta_l(\hat{\mathbf{S}})\ ^2} \text{ for } l = 1, \dots, L.$
Step 4: Obtain the class of existence $I_e = \{l r_l < \alpha\}$ and the estimated distance of the object $\hat{\mathbf{d}} = \{d_l l \in I_e\}$.

to use the residuals [22]. For each class, we define its characteristic function $\delta_l : \mathbb{R}^{L \cdot N} \rightarrow \mathbb{R}^{L \cdot N}$ that selects the coefficients of $\hat{\mathbf{S}}$ associated with the l th class while nullifying the coefficients of other classes. Thus, for $\hat{\mathbf{S}} \in \mathbb{R}^{L \cdot N}$, $\delta_l(\hat{\mathbf{S}}) \in \mathbb{R}^{L \cdot N}$ is obtained by including the elements corresponding to the l th class and nulling all elements of $\hat{\mathbf{S}}$ from other classes. By using the characteristic function, we denote the regularized residuals as

$$r_l := \frac{\|\mathbf{y} - \mathbf{B}\delta_l(\hat{\mathbf{S}})\|^2}{\|\delta_l(\hat{\mathbf{S}})\|^2}. \quad (8)$$

If the object is located at d_l , the r_l value is smaller than those at other distances. We denote I_e as a set of the indices of estimated distances at which the objects are expected to be located. With r_l for $l = 1, \dots, L$, the classification rule is given by

$$I_e := \{l | r_l < \alpha\}, \quad (9)$$

where α is an arbitrary constant. A set of distances $\hat{\mathbf{d}}$ where the object is expected to be located can be determined by

$$\hat{\mathbf{d}} = \{d_l | l \in I_e\}. \quad (10)$$

Then, the images that only correspond to the estimated distances are reconstructed by solving $\hat{\mathbf{x}}_l = \mathbf{w}^T \hat{\mathbf{S}}_l$ for $l \in I_e$. The SRC-based depth estimation algorithm is summarized in Table 1.

Thus far, depth estimates have been obtained by finding locations in a dictionary, where the signals have small residuals. To improve the

depth accuracy, the number of distance elements L must be increased in the form of the dictionary. However, the dictionary cannot include infinitely many possible distances owing to computational complexity and memory storage. To solve Eq. (7), $O(M \cdot N \cdot L)$ computations for every iteration and $O(M \cdot N \cdot L)$ storage are required; these are proportional to the number of distance elements [28]. We note that the l_1 norm minimization in Eq. (7) finds a sparse solution whose nonzero elements are most closely associated with the most correlated measurement matrix. By using the fact that the measurement matrices of neighboring distances are relatively more correlated than those of farther distances in dictionary matrix \mathbf{A} , we propose an iterative depth estimation method that is more efficient in terms of computational complexity and memory storage. Instead of universally searching for the object distances at once, we iteratively refine the set of distances in a coarse-to-fine manner [29]. The distances are investigated in detail only around regions where objects are expected to be present.

For iteration index i , we first choose a set of coarse distances within the range of interest $R^{(i)}$ as $\mathbf{d}_l^{(i)}$ for $l = 1, 2, \dots, L_i$, at which the objects can potentially be located. The depth interval is $\Delta d^{(i)} = d_{l+1}^{(i)} - d_l^{(i)}$. Accordingly, $\mathbf{A}^{(i)}$ and $\mathbf{W}^{(i)}$ can be generated from the structure of the compound eye imaging system. The sparse signal $\hat{\mathbf{S}}$ is reconstructed by solving Eq. (7), and the estimate of distances $\hat{\mathbf{d}}^{(i)}$ can be obtained by solving Eq. (8)–(10). Then, the set of distances is updated by refining the range of interest and the depth interval. The range of interest is refined around the estimated distances, that is, $R^{(i+1)} = [\hat{d}_j^{(i)} - \Delta d^{(i)}/2, \hat{d}_j^{(i)} + \Delta d^{(i)}/2]$ for $j = 1, \dots, |\hat{\mathbf{d}}^{(i)}|$, where $|\cdot|$ represents the cardinality. The depth interval is refined as $\Delta d^{(i+1)} = \Delta d^{(i)}/K$ for a positive real number $K > 1$. Then, the updated set of finer distances is

$$\mathbf{d}^{(i+1)} = \{\hat{d}_{j,k}^{(i+1)}\} \quad (11)$$

where $\hat{d}_{j,k}^{(i+1)} = (\hat{d}_j^{(i)} - \Delta d^{(i)}/2) + (k-1)\Delta d^{(i+1)}$ for $j = 1, \dots, |\hat{\mathbf{d}}^{(i)}|$ and $k = 1, \dots, [K]$. We repeat this process until the depth interval is sufficiently fine. The iterative coarse-to-fine depth estimation algorithm is summarized in Table 2.

Table 2

Iterative depth estimation algorithm.

Initial parameters: \mathbf{y} , $R^{(1)}$, $\mathbf{d}^{(1)} = \{d_1^{(1)}, d_2^{(1)}, \dots, d_{L_1}^{(1)}\}$, \mathbf{w} , η , α , $i = 1$

Step 1: Set $\mathbf{A}^{(i)} = [\mathbf{A}_{d_1^{(i)}} \dots \mathbf{A}_{d_{L_i}^{(i)}}]$ and $\mathbf{W}^{(i)} = \text{diag}(\mathbf{w}, \dots, \mathbf{w})$.

Step 2: Solve Eq. (7) from \mathbf{y} given $\mathbf{A}^{(i)}$ and \mathbf{W} , and obtain $\hat{\mathbf{S}}$.

Step 3: Calculate the regularized residuals:

$$r_l := \|\mathbf{y} - \mathbf{B}\delta_l(\hat{\mathbf{S}})\|^2 / \|\delta_l(\hat{\mathbf{S}})\|^2 \text{ for } l = 1, \dots, L.$$

Step 4: Obtain the set of indices of estimated distances $I_e^{(i)} = \{l | r_l < \alpha_i\}$.

Step 5: Update

$$R^{(i+1)} = [\hat{d}_j^{(i)} - \Delta d^{(i)}/2, \hat{d}_j^{(i)} + \Delta d^{(i)}/2],$$

$$\Delta d^{(i+1)} = \Delta d^{(i)}/K \text{ for } K > 1, \mathbf{d}^{(i+1)} = \{\hat{d}_{j,k}^{(i+1)}\} \text{ and } L_{i+1} = |\mathbf{d}^{(i+1)}|,$$

where $\hat{d}_{j,k}^{(i+1)} = (\hat{d}_j^{(i)} - \Delta d^{(i)}/2) + (k-1)\Delta d^{(i+1)}$

for $j = 1, \dots, |\hat{\mathbf{d}}^{(i)}|$ and $k = 1, \dots, \lceil K \rceil$.

Step 6: Set $i = i + 1$ and repeat from Step 1 until the depth resolution is sufficiently fine.

4. Results

To evaluate the performance of our depth estimation technique, we consider a hemispherical compound eye with a radius of 6.92 mm, where each ommatidium has a height of 1.35 mm [5]. The compound eye consists of an $M = 80 \times 80$ array of uniformly spaced ommatidia with $\Delta\phi = 1.8^\circ$ and $\Delta\varphi = 45^\circ$, such that $\Delta\varphi \gg \Delta\phi$. The 200×200 mm object consists of $N = 100 \times 100$ pixels. Thus, each measurement matrix has dimensions of 6400×10000 . For the sparsifying basis \mathbf{w} , we use a db2 wavelet transform and a level of two. To solve Eq. (7), we use the fast and efficient alternating direction method [28].

First, we determine the depth estimation accuracy for the proposed compound eye. Because the measurement matrices corresponding to neighboring distances are more correlated with each other, we set a distance of 108 mm from the compound eye as the reference distance and compare with other distances by increasing the depth intervals. To evaluate the depth estimation accuracy with respect to the depth interval, we consider a sparse signal as an input, that is, $\mathbf{w} = \mathbf{I}$, where \mathbf{I} represents an identity matrix. In each assessment, a sparse signal dimension of 10000×1 with 5%, 7.5%, and 10% of randomly located nonzero elements is used. The distance of the input signal is randomly chosen between the reference distance and the comparison distance. The tolerance α in Eq. (9) is set to be 0.1. This assessment is repeated 100 times. As seen in Fig. 3, as the object distances increase, the accuracy of the proposed depth estimation increases. For signals with 5% sparsity, if the depth intervals are larger than 0.3 mm, the proposed depth estimation works with more than 97% accuracy. For the l_1 norm minimization in Eq. (7), the reconstruction performance depends on the sparsity of the input signal, that is, low accuracy for the input signal with large sparsity. Thus, as the sparsity increases, the performance of the proposed depth estimation deteriorates as well.

The proposed COMPU-EYE imaging system used for evaluating the image reconstruction is shown in Fig. 4. The hemispherical compound eye observes an object consisting of four characters: G, i, S, and T. The characters are located at three different distances from the compound eye. G is 108 mm away from the compound eye, i and S are 109 mm away, and T is 112 mm away, as shown in Fig. 4(b). As shown in Fig. 4(c), the characters overlap one another, preventing the distance information from being inferred. The DSP technique introduced in Section 3 can be used to decompose each letter given its distance.

We demonstrate the performance of the proposed depth estimation method when the object distances are included in the set of potential distances in the dictionary. We assume that the depth range of interest is from 108 mm to 112 mm and the target depth resolution is 1 mm. Within the range of interest, the distances are uniformly sampled with 1-mm resolution, that is, $\mathbf{d} = \{108, 109, 110, 111, 112\}$. For depth estimation and object reconstruction, we construct a dictionary matrix $\mathbf{A} = [\mathbf{A}_{108} \mathbf{A}_{109} \mathbf{A}_{110} \mathbf{A}_{111} \mathbf{A}_{112}]$ in accordance with the potential distances.

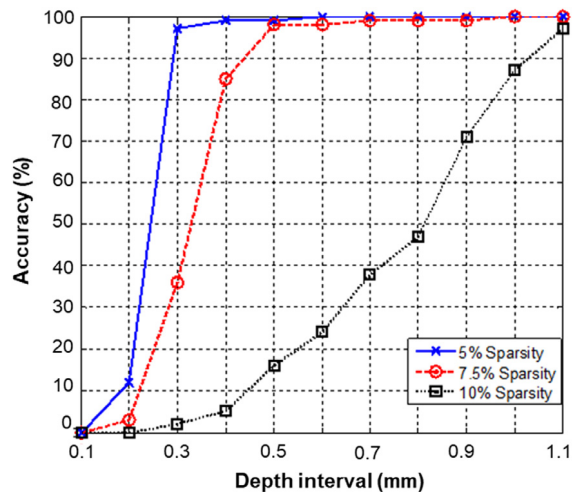


Fig. 3. Depth estimation accuracy (%) with respect to depth interval.

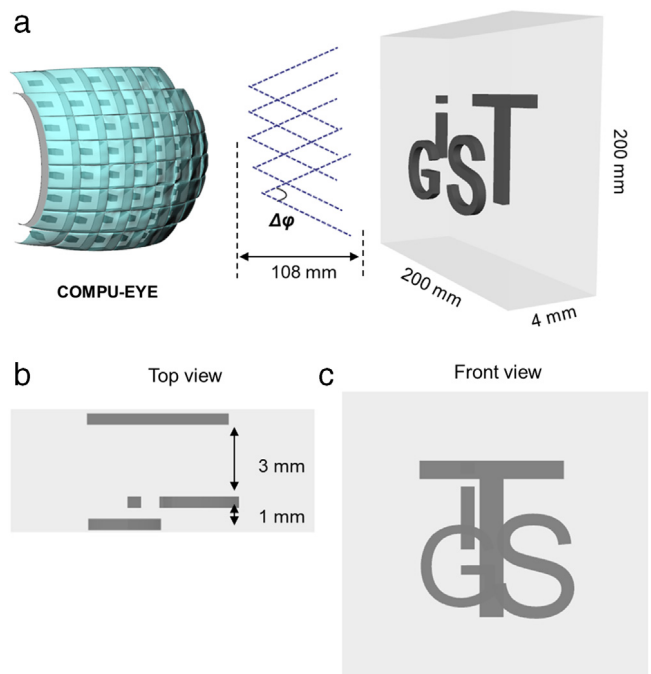


Fig. 4. Proposed COMPU-EYE imaging system: (a) Three-dimensional, (b) Top, and (c) Front views.

Given \mathbf{A} , we can solve Eq. (7) to obtain $\hat{\mathbf{S}}$ from \mathbf{y} . Then, $\hat{\mathbf{X}}$ can be obtained by calculating $\hat{\mathbf{X}} = \mathbf{W}^T \hat{\mathbf{S}}$. The reconstructed $\hat{\mathbf{S}}$ and $\hat{\mathbf{X}}$ are shown in Fig. 5(a) and (b), respectively. Owing to the sparse signal reconstruction, most of the nonzero signals in Fig. 5(a) are concentrated in the set of indices corresponding to distances of 108 mm, 109 mm, and 112 mm. We note that the reconstruction errors in Fig. 5(a) and (b) for 110 mm and 111 mm are caused by coherence among neighboring measurement matrices. As indicated in Fig. 5(c), the regularized residuals of the set of indices corresponding to distances of 108 mm, 109 mm, and 112 mm are smaller than those corresponding to the other distances. As a result, the index set of the estimated distances and the estimated distances of the objects are determined as $I_e = \{1, 2, 5\}$ and $\hat{\mathbf{d}} = \{108, 109, 112\}$, respectively. In Fig. 5(d), the reconstructed signals $\hat{\mathbf{x}}_i = \mathbf{w}^T \hat{\mathbf{s}}_i$ for $i \in \{1, 2, 5\}$ are represented as images. Note that the observation is highly distorted owing to the overlap among the ommatidial receptive fields.

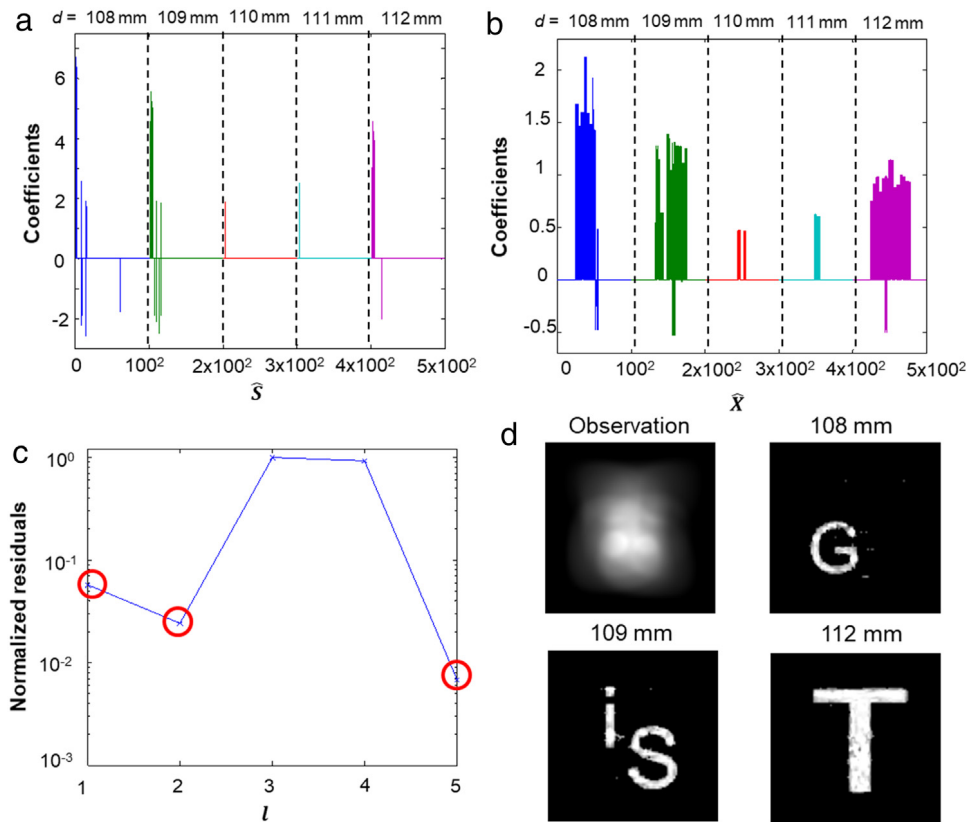


Fig. 5. (a) Reconstruction of \hat{S} , (b) Reconstruction of \hat{X} , (c) Normalized regularized residuals, and (d) Ommatidial observations and reconstructed images with respect to estimated distances.

The reconstructed characters at 108, 109, and 112 mm are clearly visible. This result indicates that COMPU-EYE achieves 1-mm depth resolution. We note that the reconstruction resolution is also improved by 1.56 times because 100×100 -pixel images are reconstructed from 80×80 -pixel ommatidial observations.

We now investigate the performance of the iterative depth estimation for the object shown in Fig. 4. We assume that the potential object locations are unknown and that the range of interest is from 100 mm to 120 mm, that is, $R^{(1)} = [100 \text{ mm}, 120 \text{ mm}]$. For the SRC-based depth estimation method in Table 1 to achieve a depth resolution of 1 mm, the dictionary requires 21 concatenated measurement matrices with dimensions of 6400×210000 . The computational complexity of this task necessitates the use of the iterative depth estimation method described in Table 2. We first formulate a set of coarse distances $\mathbf{d}^{(1)} = \{100, 110, 120\}$ and $\mathbf{A} = [\mathbf{A}_{100} \ \mathbf{A}_{110} \ \mathbf{A}_{120}]$ correspondingly. The result of iterative depth estimation is shown in Fig. 6. At the 1st iteration, because the objects are located at around 110 mm, the residual value at 110 mm is the smallest. Thus, the object distance is estimated as 110 mm for $\alpha = 0.3$ at the 1st iteration. For detailed depth estimation, we further set a dictionary with finer distances at around 110 mm. The range of interest is refined as $R^{(2)} = [105, 115]$ and the depth interval, as $\Delta d^{(2)} = 3$ for $K = 3.33$. Then, the set of distances is updated as $\mathbf{d}^{(2)} = \{105, 108, 111, 114\}$. At the 2nd iteration, the residual values at distances of 108 mm and 111 mm are smaller than those at other distances. Thus, we estimate that the objects are located at around 108 mm and 111 mm for $\alpha = 0.2$. The range of interest is refined as $R^{(3)} = [107, 109] \cup [110, 112]$ and the depth interval as $\Delta d^{(3)} = 1$ for $K = 3$. Then, the set of distances is updated as $\mathbf{d}^{(3)} = \{107, 108, 109, 110, 111, 112\}$. At the 3rd iteration, the object distances are estimated as 108 mm, 109 mm, and 112 mm from the compound eye for $\alpha = 0.1$. As a result, the objects are efficiently reconstructed by using the proposed iterative depth estimation method.

Now, we aim to demonstrate depth estimation for an object with continuous depths. As a target, we consider a plane object that is slanted at 23° toward the compound eye and located 108 mm away from the compound eye. When the range of interest is from 108 mm to 111 mm, the object distance can be uniformly sampled as $\mathbf{d} = \{108, 109, 110, 111\}$ in Fig. 7(b). The proposed depth estimation method provides a depth map of the object with 1-mm depth resolution, as shown in Fig. 7(d). Consequently, an object with continuous depths can be well reconstructed by using the estimated distances, as seen in Fig. 7(c). In this manner, continuous depths can be estimated. We note that if we densely sample the range of distance, the depth map will be more accurate; however, there is a limit to the depth resolution, as seen in Fig. 3.

5. Conclusion

We have proposed a depth estimation method based on the COMPU-EYE imaging system, in which the ommatidial acceptance angle is much larger than the interommatidial angle. The ommatidial receptive fields overlap, and the disparities between ommatidial observations vary with object distance. As a result, the uniqueness of the generated measurement matrix depends upon the object distance. In the proposed technique, the dependences of the disparities between the ommatidial observations and the measurement matrix uniqueness on object distance are used to estimate the depth. This work helps not only to estimate object distances but also to reconstruct objects with high resolution, and it is therefore essential for future development of the COMPU-EYE system.

Generally, disparity-based depth estimation methods have limitations for very distant objects because the disparities decrease [30]. By varying the acceptance angles of the ommatidia or arranging the ommatidia irregularly, the range of depth estimation can be extended

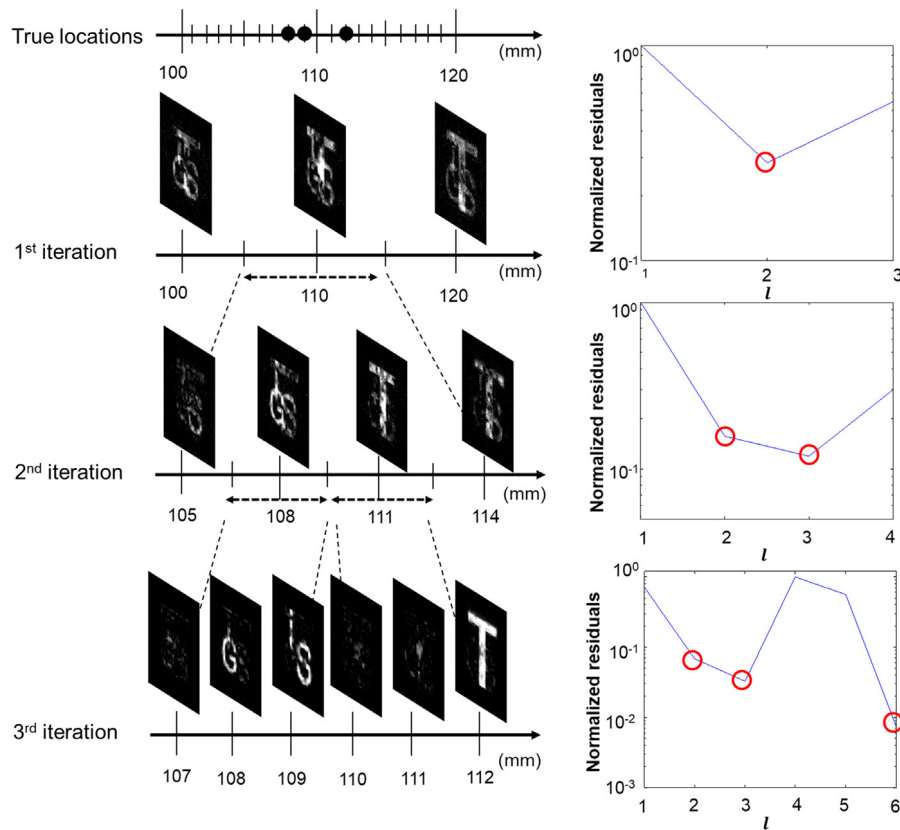


Fig. 6. An example of the iterative depth estimation method.

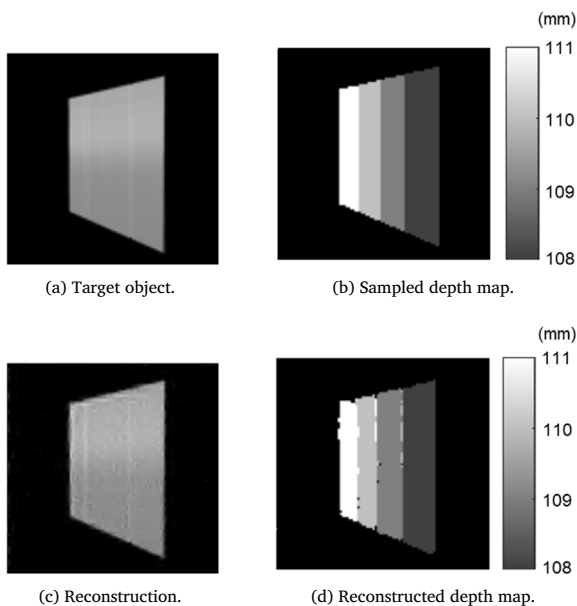


Fig. 7. Depth estimation and object reconstruction for a slanted object.

adaptively, that is, large acceptance angles for small distances and small acceptance angles for large distances [31]. Our future work will focus on improving the depth resolution by designing COMPU-EYE to have high incoherence among measurement matrices with respect to object distances. Furthermore, we will improve the depth estimation performance by applying the l_0 -norm based minimization to solve Eq. (6).

This has been shown to provide better reconstruction performance than the l_1 -norm minimization [32].

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) [NRF-2015R1A2A1A05001826].

References

- [1] A. Borst, J. Plett, Optical devices: Seeing the world through an insect's eyes, *Nature* 497 (2013) 47–48.
- [2] K. Kagawa, K. Yamada, E. Tanaka, J. Tanida, A three-dimensional multifunctional compound-eye endoscopic system with extended depth of field, *Electron. Comm. Jpn.* (2012) 14–27.
- [3] F. Expert, F. Ruffier, Flying over Uneven Moving Terrain Based on Optic-Flow Cues Without Any Need for Reference Frames or Accelerometers, *Bioinspir. Biomim.*, IOP Publishing, 2015, pp. 1–25.
- [4] A. Hassanfiroozi, Y.-P. Huang, B. Javidi, H.-P.D. Shieh, Hexagonal liquid crystal lens array for 3D endoscopy, *Opt. Express* 23 (2015) 971–981.
- [5] Y.M. Song, Y. Xie, V. Malyarchuk, J. Xiao, I. Jung, K.J. Choi, Z. Liu, H. Park, C. Lu, R.H. Kim, R. Li, K.B. Crozier, Y. Huang, J.A. Rogers, Digital cameras with designs inspired by the arthropod eye, *Nature* 497 (2013) 95–99.
- [6] D. Floreano, R. Pericet-Camara, S. Violette, F. Ruffier, A. Brückner, R. Leitel, W. Buss, M. Menouni, F. Expert, R. Juston, M.K. Dobrzynski, G. L'Epplattenier, F. Reckenwald, H.A. Mallot, N. Franceschini, Miniature curved artificial compound eyes, *Proc. Natl. Acad. Sci.* 110 (2013) 9267–9272.
- [7] O. Cogal, Y. Leblebici, An insect eye inspired miniaturized multi-camera system for endoscopic imaging, *IEEE Trans. Biomed. Circuits Syst.* 11 (2017) 212–224.
- [8] M. Sarkar, *Bioinspired Optical Imaging, Biologically Inspired Computer Vision*, Wiley-VCH Verlag GmbH & Co., 2015, pp. 109–142. KGaA.
- [9] M.F. Land, L. Chittka, R.F. Chapman, *Vision, Structure and Function*, Cambridge University Press, 2012, pp. 708–737.
- [10] S. Rossel, Binocular vision in insects: How mantids solve the correspondence problem, *Proc. Natl. Acad. Sci. U.S.A.* (1996) 13229–13232.
- [11] M.F. Land, Visual acuity in insects, *Annu. Rev. Entomol.* 42 (1997) 147–177.

- [12] K. Kral, Behavioural-analytical studies of the role of head movements in depth perception in insects, birds and mammals, *Behav. Process.* (2003) 1–12.
- [13] F. van Breugel, K. Morgansen, M.H. Dickinson, Monocular distance estimation from optical flow during active landing maneuvers, *Bioinspir. Biomim.* (2014) 025002–025010.
- [14] H.E. Esch, J.E. Burns, Honeybees use optic flow to measure the distance of a food source, *Naturwissenschaften* (1995) 38–40.
- [15] W.B. Lee, H. Jang, S. Park, Y.M. Song, H.N. Lee, COMPU-EYE: A high resolution computational compound eye, *Opt. Express* 24 (2016) 2013–2026.
- [16] X. Xiao, B. Javidi, M. Martinez-Corral, A. Stern, Advances in three-dimensional integral imaging: Sensing, display, and applications [Invited], *Appl. Opt.* 52 (2013) 546–560.
- [17] A. Stern, B. Javidi, Three-dimensional image sensing, visualization, and processing using integral imaging, *Proc. IEEE* 94 (2006) 591–607.
- [18] L. Cao, J. Peter, Iterative reconstruction of projection images from a microlens-based optical detector, *Opt. Express* 19 (2011) 11932–11943.
- [19] Y. Gao, W. Liu, P. Yang, B. Xu, Depth estimation based on adaptive support weight and SIFT for multi-lenslet cameras, 6th International Symposium on Advanced Optical Manufacturing and Testing Technologies (AOMATT 2012), SPIE2012, 2012, pp. 84190C–84194.
- [20] Y. Yuan, X. Wang, X. Wu, J. Zhang, Y. Zhang, Improved resolution integral imaging using random aperture coding based on compressive sensing, *Optik - Int. J. Light Electron Opt.* 130 (2017) 413–421.
- [21] E.J. Candes, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory* 52 (2006) 5406–5425.
- [22] D.L. Donoho, M. Elad, V.N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Trans. Inform. Theory* (1995) 6–18.
- [23] O. Cogal, Y. Leblebici, An insect eye inspired miniaturized multi-camera system for endoscopic imaging, *IEEE Trans. Biomed. Circuits Syst.* (2016) 1–13.
- [24] K.W. Seo, D.-w. Lee, B.R. Min, A 3-D information acquisition algorithm for close range endoscopy, in: R. Magjarevic, J.H. Nagel (Eds.), *World Congress on Medical Physics and Biomedical Engineering 2006: August 27–September 1, 2006 COEX Seoul, Korea Imaging the Future Medicine*, Springer, Berlin, Heidelberg, 2007, pp. 2612–2615.
- [25] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2009) 210–227.
- [26] S. Younghak, L. Seungchan, L. Junho, L. Heung-No, Sparse representation-based classification scheme for motor imagery-based brain–computer interface systems, *J. Neural Eng.* 9 (2012) 056002.
- [27] X. Song, Z. Liu, X. Yang, S. Gao, A new sparse representation-based classification algorithm using iterative class elimination, *Neural Comput. Appl.* 24 (2014) 1627–1637.
- [28] J.F. Yang, Y. Zhang, Alternating direction algorithms for 1 problems in compressive sensing, *SIAM J. Sci. Comput.* (2011).
- [29] D. Malioutov, M. Cetin, A.S. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays, *IEEE Trans. Signal Process.* 53 (2005) 3010–3022.
- [30] R. Horisaki, S. Irie, Y. Ogura, J. Tanida, Three-dimensional information acquisition using a compound imaging system, *Opt. Rev.* 14 (2007) 347–350.
- [31] H. Ryoichi, K. Keiichiro, N. Yoshizumi, T. Takashi, M. Yasuo, T. Jun, Irregular lens arrangement design to improve imaging performance of compound-eye imaging systems, *Appl. Phys. Express* 3 (2010) 022501.
- [32] Y. Esmaeili Salehani, S. Gazor, I.-M. Kim, S. Yousefi, ℓ_0 -norm sparse hyperspectral unmixing using arctan smoothing, *Remote Sens.* 8 (2016) 187.

A Sub-Nyquist Radar Electronic Surveillance System

JEONG PARK¹, JEHYUK JANG¹, SANGHUN IM²,
AND HEUNG-NO LEE¹, (Senior Member, IEEE)

¹School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

²Hanwha Systems, Seongnam 13524, South Korea

Corresponding author: Heung-No Lee (heungno@gist.ac.kr)

This work was supported by a grant-in-aid of Hanwha Systems.

ABSTRACT The modulated wideband converter (MWC) is well-known for a sub-Nyquist wideband sampling capability based on compressed sensing (CS) theory. In this paper, our goal is to use the MWC as a base to design a sub-Nyquist radar electronic surveillance (ES) system. Our focus is then to extend the capabilities of the previous MWC system in order to meet the challenges, i.e., a very long acquisition time, a much larger simultaneous monitoring bandwidth, and a faster digital signal processing receiver. To this end, we present a new performance analysis framework and then a new digital domain receiver. The proposed performance analysis framework will be useful in comparing signal-acquisition performance of the proposed ES system with those of other sub-Nyquist receivers, including those of the classical Nyquist rate receivers, without resorting to extensive simulations. This framework can also be used to study the complex interplays of important system parameters of MWC, such as the sampling rate, the number of parallel channels, the period of Pseudo random sequence, and thus guides us in selecting the right system dimensions and parameters for desired performance. Radar surveillance application has its inherent needs for very long acquisition time and simultaneous monitoring of very large frequency range. To meet this challenge, a fast signal recovery system needs to be developed, so that radar signal logistics can be retained and recovered from compressed samples. We have proposed a split and synthesis process in which the radar signal recovery problem over a long signal acquisition time can be divided into many small CS signal recovery problems, and the solutions for small pieces are put together later on at the end. In addition, a sub-sampling method is proposed to have the multiple measurement vector problem complete signal recovery faster without noticeable performance loss.

INDEX TERMS Compressed sensing, electronic surveillance, modulated wideband converter, multiple measurement vectors, radar signal.

I. INTRODUCTION

Electronic surveillance (ES) systems monitor radar signals emitted from opponent radar systems, which detect subjects by transmitting radar signals and receiving them when reflected back from the subjects. Radar ES systems are useful for recognizing the intent of a threat in advance. Opponent radar signals are spectrally sparse, spread around a wide frequency band, and are unknown in advance, presenting unique challenges for signal processing.

Radar ES systems can utilize a Nyquist-rate receiver, such as the rapidly swept superheterodyne receiver (RSSR) [1], [2]. RSSR chronologically samples the sub-bands of a wideband region. However, RSSR inevitably

misses some of the signals. For wideband signals, the sweeping period must be relatively long, though opponent radar signals are brief in duration. As shown in Fig. 1, although a signal may appear in some frequency bands for a while, the RSSR's sweep samples empty sub-bands and fails to catch the signal. The inevitable failure of catching some signals from the RSSR sweep can become a critical problem depending on its applications such as detecting missiles and monitoring hostile aircraft.

To take samples of spectrally sparse signals at a rate far below the Nyquist rate without information losses, the modulated wideband converter (MWC) [3], [4] and random modulation pre-integration (RMPI) [5], [6] have been proposed.

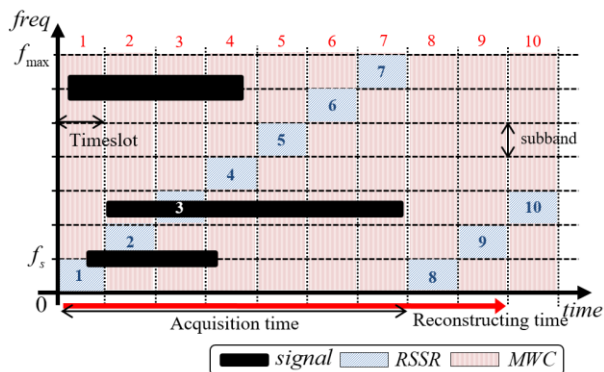


FIGURE 1. Signal model and signal acquisition schemes of the RSSR and MWC denoted as blue and red colored integers, respectively.

Based on the compressive sensing (CS) theory [7], [8], both systems compress the received spectrum by mixing it with rapidly alternating pseudo-random sequences, takes the low-rate samples over a certain acquisition time, and finally reconstructs the original spectrum from the collected samples in the digital domain. We aim to design an ES system based on MWC.

Notably, in terms of the probability of successful signal acquisition, MWC has not been compared with other sub-Nyquist receivers or conventional Nyquist receivers with similar hardware costs. To this end, one can construct prototypes and repeat hundreds of signal-acquisition tests. However, that is inefficient and expensive. Thus, a performance analysis model which predicts the signal acquisition performance of systems with similar hardware costs will be of highly valuable.

In the ES scenario, owing to tactical purposes such as the avoidance of reverse tracing by the enemy, modern radar systems frequently switch their signal characteristics. The longer radar samples are acquired, the more information of foe’s intention can be scrutinized. In the ES system exploiting MWC, since CS reconstruction algorithms deal with only a finite length of compressed sample, the reconstructed samples possibly contain only a portion of a radar signal. Hence, for MWC to retain sufficient amount of information for the intention analysis, a sufficiently long acquisition time is needed. A CS signal reconstruction algorithm covering the entire acquisition time would give the best performance, but such an algorithm would require a very high computational complexity. To compound the matter, the signal bandwidth we aim to study is very large as well. As radar systems cover very wide frequency regions including C-band (4-8GHz), X-band (8-12GHz) and Ku-band (12-18GHz) [9], the bandwidth of interest for simultaneous monitoring purpose needs to be wider than the 2GHz bandwidth of the previously studied MWC [3]. As the result, the radar ES system we aim to study in this paper requires very large system dimension for any CS signal reconstruction algorithm to work. Large system dimension entails high computational complexity. Our goal therefore is to focus on how to divide the observation time into small segments of time to reduce complexity of signal

reconstruction, and how to put the segmented signals of interest together without losing quality.

Our contributions in this paper are two fold, one is a novel signal-acquisition probability analysis and a low complexity radar ES system design for very wide bandwidth monitoring applications. First, we present a new performance evaluation framework which allows analytic comparison of the signal-acquisition performances of several wideband signal receivers. This allows us to compare receiver architectures while avoiding realization of all the receivers and exhaustive testing in simulation. Our analysis demonstrates the specific benefits of MWC over conventional RSSR. In addition, the analytic method can be applied to other sub-Nyquist receivers based on CS. For example, analysis applied on RMPI is included. This framework can also return design parameters for the radar ES system.

Second, the design of a low complexity and a wideband monitoring ES receiver using MWC is presented in this paper. We show how a long acquisition time is divided into continuously disjoint timeslots, how a CS reconstruction algorithm works for each segment in a single time slot, and how all of the reconstructed segments are synthesized. We call this split and synthesis process and show this effort reduces the total computational complexity required for reconstruction of radar signals for a long acquisition time at the cost of slight degradation in reconstruction performance. In addition, a sub-sampling method is presented aiming to further reduce the computational complexity of a CS reconstruction algorithm working within a time slot. Namely, the subsamples are selected based on the principle components of the received signal.

In Section II, we briefly introduce wideband signal receivers, including RSSR, RMPI, and MWC, and formulate problems for the radar ES system. Section III details the analysis of signal-acquisition probabilities. Section IV presents our sub-Nyquist radar ES system design, including the split-process and synthesis process. A pre-processing method for the CS reconstruction algorithm is detailed in Section V. Sections VI and VII present the results of our simulations and our conclusions, respectively.

II. PROBLEM FORMULATION AND BACKGROUND

The input $x(t)$ is modeled as an aggregation of radar signals generated from a range of radar systems. In particular, the input is defined by the following equation:

$$x(t) = \sum_{i=1}^N r_i(t), \quad 0 \leq t < \infty, \quad (1)$$

where $r_i(t)$ is a radar signal from the i -th radar system and is located widely within $\mathcal{F}_{NYQ} = [-f_{max}, f_{max}]$, where f_{max} can be of the order of GHz. Including the carrier frequencies, the pulse description words (PDWs) such as pulse repetition intervals (PRIs), time-of-arrivals (TOAs), time-of-departures (TODs), pulse widths, and duty cycles are unknown *a priori*. For each $r_i(t)$, we model the carrier frequency ranges within $[0, f_{max})$, and the bandwidth B_i is

truncated to $B_{\min} \leq B_i \leq B_{\max}$. For separable $r_i(t)$, we assume that the spectra of $r_i(t)$ are disjointed. In addition, the aggregation $x(t)$ is sparse in the frequency domain, i.e., $2NB_{\max} \ll f_{\max}$. Briefly, we acquire a successively incoming signal $x(t)$, which is regarded as a spectrally sparse multiband signal [3] with unknown parameters.

A. RAPIDLY SWEPT SUPERHETERODYNE RECEIVER

The RSSR [1] is a representative Nyquist receiver designed to cover wideband regions with a low rate of analog-to-digital convertors (ADC). RSSR receives a multiband signal and divides the entire spectrum into multiple subbands by exploiting a bank of bandpass filters. The subbands are then sampled by an ADC. With time-division multiplexing [2], RSSR chronologically takes time-domain samples of the subbands in sequence. The blue boxes in Fig. 1 depict the time-division multiplexing, where the numbered regions in time-frequency domain are the signal acquiring subbands of RSSR.

Despite the simple system structure of RSSR, in Fig. 1, it inevitably fails to acquire some of radar signals outside the current acquiring subband. One can reduce the failures of signal acquisitions by using faster ADC in the state of the art [10], but such ADC have many implementation problems such as prohibitive cost, high energy consumption, low memory, and low ADC resolution [6], [11]. The implementation limits also restricts to wider the bandwidth f_{\max} of signal acquisition. Intuitively, the probability that RSSR fails to acquire the whole radar signals would increase as the number of signals and/or the range of the input spectrum increases. In Section III, we compute the probability of successful signal acquisition by RSSR.

B. RANDOM MODULATION PRE-INTEGRATION

RMPI [5] is a channelized sub-Nyquist receiver that acquires a multiband signal at one acquisition time T_{acq} . For each channel, the multiband signal is mixed with a pseudorandom (PR) sequence and the result is then integrated. An ADC samples the mixed result at the sub-Nyquist rate f_s after the integrating module. If m is the number of channels, the channel-end sampling rate [6] f_{bs} is defined as follows:

$$f_{bs} \triangleq mf_s. \tag{2}$$

With the system matrix from the analog architecture, RMPI reconstructs multiband signal with a CS algorithm. However, because the matrix is block diagonal in form, the system matrix is considerably large to compute expeditiously. For one block, the number of rows and columns corresponds to the number of channels and f_{nyq}/f_s respectively, and each block is repeated f_{nyq}/f_{bs} times. This large block-diagonal matrix causes high computational complexity and long reconstruction times, which renders RMPI impractical for the ES applications.

In addition, the range of the sample sequence containing signal information corresponds to Nyquist frequency and is digitized at an interval of $1/T_{acq}$. Hence, for N given signals

and a minimum bandwidth B_{\min} , the number of nonzero entries in a sampled sequence is more than $2NB_{\min}T_{acq}$. The large number of nonzero entries impedes RMPI's signal-acquisition performance.

C. MODULATED WIDEBAND CONVERTER

To resolve missing signals in the RSSR and avoid the computational limits of RMPI, we examined the MWC [3] for ES, which comprises analog and digital modules. In the analog module, the MWC takes samples containing compressed information of $x(t)$ at a rate below the Nyquist rate. In the digital module, the post-digital signal process (DSP) and a CS recovery algorithm reconstruct the compressed samples into the Nyquist-rate sample of $x(t)$.

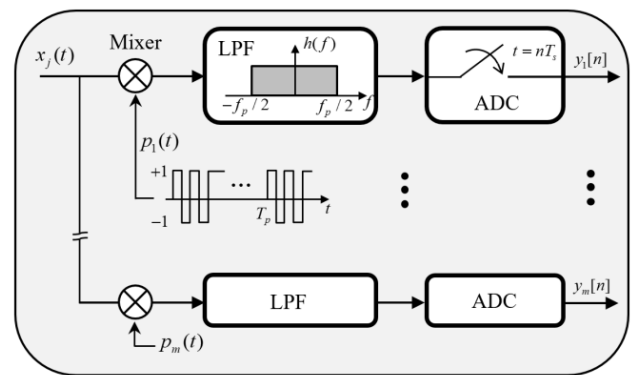


FIGURE 2. Analog module of the MWC.

The analog module of the MWC comprises m channels, including a series of mixers, low-pass filters (LPFs), and ADCs, as shown in Fig. 2. For each channel, the multiband signal is mixed with a T_p -periodic PR sequence, $p_i(t)$. The spectrum of the sequence has $M = 2M_0 + 1$ weighted impulses at intervals of $f_p = T_p^{-1}$. The mixed signal passes through an anti-aliasing LPF whose cutoff frequency is defined as $f_s/2 = qf_p/2$, where $q = 2q_0 + 1 > 0$ is an odd integer. As a result, the mixer and LPF divide the input frequency range $[-f_{\max} - q_0f_p, f_{\max} + q_0f_p]$ into $L = 2L_0 + 1$ sub-bands at intervals of f_p , as presented in [3]. The sub-bands are then compressed by multiplying them with the Fourier coefficients $c_{i,l}$ of the PR sequence and projecting into $[-f_s/2, f_s/2)$. Next, the ADC samples the compressed sub-bands at the rate of f_s . For the i -th channel, the discrete-time Fourier transform (DTFT) of the output of the ADC can be expressed by

$$\check{y}_i \left(e^{j2\pi f T_s} \right) = \sum_{l=-L_0}^{L_0} c_{i,l} \mathbf{X}(f - lf_p) \tag{3}$$

for $-f_s/2 \leq f < f_s/2$ [3]. From the projection into $[-f_s/2, f_s/2)$, the information from $q = f_s/f_p$ subbands are piled in a single row of the \mathbf{X} matrix. Note that $M \leq L$ Fourier coefficients of $p_i(t)$ are unique and $q - 1 = L - M$ coefficients are repetitions. The MWC then reconstructs the Nyquist sample of $x(t)$ from the compressed samples in the digital module of the MWC.

In the DSP, the channel expansion method [3] is applied to extend the number of the equation (3) by disjointing the correlations of the $q-1$ repeated Fourier coefficients $c_{i,l}$. The channel expansion method is represented by the following equation:

$$\begin{aligned} \check{y}_{i,k}[\tilde{n}] &= (y_i[n]e^{-j2\pi kf_p n T_s}) * h_D[n] \Big|_{n=\tilde{n}q} \\ &= (y_i[n]e^{-j2\pi kn/q}) * h_D[n] \Big|_{n=\tilde{n}q}, \end{aligned} \quad (4)$$

where $k \in \{-q_0, \dots, q_0\}$. As shown in (4), for each k , $y_i[n]$ is modulated with a different frequency kf_p and convoluted with q LPFs $h_D[n]$, whose cutoff frequencies are π/q . The sequence is then decimated by q . As a result, the outcome of channel expansion is

$$\check{Y}_i(e^{j2\pi f T_p}) = \sum_{l=-M_0}^{M_0} c_{i,(l+k)} \mathbf{X}(f - lf_p) \quad (5)$$

for $-f_p/2 \leq f \leq f_p/2$. Consequently, we can obtain mq equations from m analog channels. With the continuous-to-finite (CTF) block in [3], the DTFT $\check{Y}_i(e^{j2\pi f T_p})$ becomes a finite sequence. For m channels, (5) can be expressed as follows:

$$\check{Y}[n] = \check{\mathbf{C}}\check{Z}[n], \quad (6)$$

where the measurement matrix $\check{Y} \in \mathbb{R}^{mq \times v}$ corresponds to the output from the MWC, $\check{\mathbf{C}} \in \mathbb{C}^{mq \times M}$ is the sensing matrix, $\check{Z} \in \mathbb{R}^{M \times v}$ contains the signal information, v is the length of column in \check{Y} yielded from the CTF block, and $M = 2M_0 + 1$. The matrix equation (6) is exploited to reconstruct the multi-band signal through a CS recovery algorithm, at which point the MWC has successfully acquired signals.

The main difference between the MWC and RMPI is how to sparsify the original continuous spectrum. First, the CS model of RMPI discretizes the continuous spectrum at the Nyquist rate. Since the received spectrum in the ES scenario usually consists of disjoint continuous narrow bands, the discretization of such a spectrum yields not only a huge size of CS model but also high sparsity. CS theory states the sparse reconstruction of the original spectrum is successful only for a low sparsity. While, the CS model of MWC divides the continuous spectrum into disjoint subbands $X(f - lf_p)$ for $l \in \{-M_0, \dots, M_0\}$ at intervals f_p . The sparsity of MWC is counted as the number of nonzero subbands, where the spectra of radar signals $r_i(t)$ are contained. With the low sparsity and the small size of CS model, we design an ES system based on MWC.

D. PROBLEM FORMULATION

Although MWC was designed to acquire a multiband signal, this architecture is difficult to directly implement in a radar ES system. Radar ES faces a tradeoff between acquisition time and post-processing time. When the acquisition time is shorter than the post-processing time, the ES system faces a bottleneck in outputting the acquired signal. As a result, continuous incoming signals stack up and are not included in the output. Meanwhile, the reconstruction process includes

complex computations in both the DSP and CS algorithms. In the DSP, for a given compressed sample length v , the computational complexity of (4) is a function of $O(v^2)$, which grows as $v \triangleq T_{acq}/T_s$ increases. One option to reduce computational complexity would be to increase the sampling period T_s of the ADC and/or reduce the acquisition time T_{acq} to reduce the length v . However, increasing the sampling period is impractical because it reduces the channel expansion factor q in (4). With the reduced q , the channel expansion method is no longer useful. Moreover, the acquisition time cannot be reduced. Because opponent radar systems frequently change their radar signal characteristics, a radar ES system is needed to acquire signals over a long acquisition time to efficiently detect an enemy radar system. Another option would involve dividing a long acquisition time into several timeslots to reduce computational complexity. However, this division scheme raises the problem of time-aliasing [12]. When segments of a long signal are processed individually and concatenated into the original signal, time-aliasing degrades the reconstructed signal at the borders of the segments. If we resolve the time-aliasing problem, we can greatly reduce computational complexity with sufficiently accurate reconstruction.

Processing a large number of samples acquired over a long acquisition time entails high computational complexity in the CS recovery algorithm. The problem of (6) in the MWC is referred to as the multiple measurement vector (MMV) problem, and the computational complexity of the algorithm is greatly influenced by the size of the measurement matrix. By reducing the measurement matrix without missing signals, computational complexity can be reduced for the radar ES system.

III. SIGNAL-ACQUISITION PROBABILITY ANALYSIS

This section compares the signal acquisition performances of the MWC, RMPI, and RSSR using a novel probability analysis. These receivers aim to watch out a very wide bandwidth of frequencies where unknown radar signals of interest may exist. Signals of no interest can be removed from consideration easily for the receivers, such as radar signals from friendly forces and commercial signals. In this analysis, therefore, for the frequency bands of interest, we assume uniform distribution of the occurrences of unknown carrier frequencies. We also assume that receivers aim to receive multiple radar signals each having bandwidth of B Hz over a very wide range of frequencies up to f_{nyq} . From the analytic results, we observe that the analysis guides the design of system parameters such as the number of channels m , the ADC sampling rate f_s , and the cycle of PR sequence f_p^{-1} . Our analytic method allows for instant comparisons without building and simulating each receiver.

From the perspective of sampling theory, the success of lossless sub-Nyquist sampling by the MWC at a given sampling rate depends proportionally on the sum of the bandwidths of the occupied sub-bands. Hence, once we learn the number of occupied sub-bands, we can expect successful

lossless acquisition for a given number of input signals. To calculate the probability of successful signal acquisition, we generate random variables representing the numbers of input signals, split spectra, and totally occupied sub-bands and derive their distributions.

First, we derive a lower bound of probability for signal acquisition via the MWC. Let X denote the number of received signals in a timeslot and Y denote split signals. Then, the conditional probability mass function (PMF) of Y given X can be defined as follows:

$$P_{Y|X}(y|x) = {}_x C_y (p_{s,MWC})^y (1 - p_{s,MWC})^{x-y}, \quad (7)$$

where p_s is the probability that a signal is split by the grid of sub-bands. By assuming that the carrier frequency of the signal can be uniformly drawn, we calculate p_s using the equation of:

$$p_{s,MWC} = 1 - \sum_{i=-M_0}^{M_0} \int_{if_p - \frac{f_p}{2} + \frac{B}{2}}^{if_p + \frac{f_p}{2} - \frac{B}{2}} \frac{1}{f_{NYQ}} dx = \frac{B}{f_p}, \quad (8)$$

where $B < f_p$ is the bandwidth of each signal. Focusing on the positive sub-bands corresponding to real signals, the number of split and un-split spectra occupants K can be defined as $K = (X - Y) + 2Y = X + Y$. Note that the occupants do not overlap. The conditional PMF of occupant K is expressed as

$$P_{K|X}(k|x) = P_{X+Y|X}(x+y|x) = P_{Y|X}(y|x) = P_{Y|X}(k-x|x). \quad (9)$$

The moment generating function of the occupant is derived as follows:

$$M_{K|X}(\mu_x) = \sum_{k=x}^{2x} e^{\mu_x k} \cdot P_{K|X}(k|x) = \sum_{k=x}^{2x} e^{\mu_x k} \cdot P_{Y|X}(k-x|x) = \sum_{k=x}^{2x} e^{\mu_x k} \cdot {}_x C_{k-x} (p_{s,MWC})^{k-x} (1 - p_{s,MWC})^{x-(k-x)} \quad (10)$$

Assuming that the MWC achieves lossless sub-Nyquist sampling if and only if $K \leq \kappa_{MWC}$, using the Chernoff bound, the lower bound of successful sampling is obtained by the following equation:

$$P_{MWC}(\text{Successful sampling}) = P_{K|X}(k \leq \kappa_{MWC} | x) > 1 - \min_{\mu_x \geq 0} e^{-\mu_x \kappa_{MWC}} M_{K|X}(\mu_x), \quad (11)$$

where κ_{MWC} is the maximum sparsity that allows the CS problem to be exactly solved. κ_{MWC} can be determined by

the equation in [3], which can be written as follows:

$$mq \approx 2\kappa_{MWC} \log(M/\kappa_{MWC}). \quad (12)$$

As presented in [13], the parameters μ_x can be obtained by solving the following equation:

$$\mu_x = \arg \min_{\mu_x \geq 0} (\ln(E[e^{\mu_x x}]) - \mu_x \kappa_{MWC}). \quad (13)$$

By substituting κ_{MWC} and the last expressions of (8) and (10) into (11), the lower bound of successful sampling can be expressed in terms of system parameters:

$$P_{MWC}(\text{Successful sampling}) > 1 - \min_{\mu_x \geq 0} e^{-\mu_x \kappa_{MWC}} \left(\frac{f_p - B}{\sqrt{B}f_p}\right)^{2x} \sum_{k=x}^{2x} {}_x C_{k-x} \left(\frac{e^{\mu_x B}}{f_p - B}\right)^k \quad (14)$$

Note that $f_p = f_{nyq}/M$.

Second, we derive the probability of signal acquisition for the RMPI. Owing to the architecture of the RMPI, the occupants correspond to digitized signal bands B/T_{acq}^{-1} among the digitized Nyquist range f_{nyq}/T_{acq}^{-1} . For N received signals, the signals occupy at least $N \lfloor BT_{acq} \rfloor$ slots. Because the band is not always exactly fit to the digitized graduation, the probability of extra occupants exceeds a portion of B compared to the width of the minimum occupants $T_o^{-1} \lfloor BT_{acq} \rfloor$ at one bin T_{acq}^{-1} , i.e.,

$$p_{s,RMPI} = (B - T_{acq}^{-1} \lfloor BT_{acq} \rfloor) / T_{acq}^{-1} = BT_{acq} - \lfloor BT_{acq} \rfloor. \quad (15)$$

We then express the maximum recoverable sparsity of RMPI κ_{RMPI} as $m_{RMPI}/2 = f_{nyq}/2f_s$. For the l_0 minimization problem, which is the optimal but mathematically intractable solver, the algorithm estimates the sparsity as half of the number of equations [14]. If the RMPI fails to acquire a signal with higher κ_{RMPI} than κ_{MWC} , MWC is superior. As found in (11), the RMPI fails at sampling when the minimum occupants $N \lfloor BT_{acq} \rfloor > \kappa_{RMPI}$ and succeeds when the maximum occupants $N (\lfloor BT_{acq} \rfloor + 1) < \kappa_{RMPI}$. In the remaining cases, the probability of successful sampling is the complement of the sampling failure, which is the number of occupants over κ_{RMPI} with $p_{s,RMPI}$. Consequently, the probability of successful sampling for RMPI can be expressed as follows:

$$P_{RMPI}(\text{Successful sampling}) = \begin{cases} 0, & \lfloor BT_{acq} \rfloor > \kappa_{RMPI} \\ 1, & N(\lfloor BT_{acq} \rfloor + 1) < \kappa_{RMPI} \\ 1 - \sum_{i=\lceil \kappa_{RMPI} - N \lfloor BT_{acq} \rfloor \rceil}^N P_{s,RMPI}^i, & o.w. \end{cases} \quad (16)$$

Lastly, we calculate the signal-acquisition probability for the RSSR. Because only the information of signals whose spectrum is fully located in a band currently being acquired

by an activated filter bank is preserved in the output samples from the RSSR, the probability of successful acquisition under the assumption of uniform distribution of the signals in the frequency domain can be described by the following equation:

$$P_{RSSR} (\text{Successful Sampling}) = \left(\frac{W_{BPF} - B}{f_{nyq}/2 - B} \right)^x, \quad (17)$$

where W_{BPF} is the bandwidth of the filter banks.

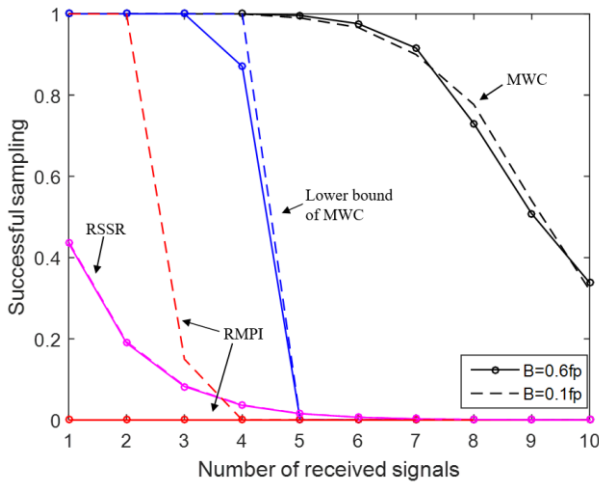


FIGURE 3. The probability of signal acquisition among the MWC, RMPI, and RSSR.

Fig. 3 shows that the MWC has the highest rate of signal acquisition. For a fair comparison, it was necessary to assign the same number of channels and sampling rates to all the receivers, including the RSSR. MWC and RMPI require a sampling rate of at least $f_s = qf_p$ for m channels. In other words, a necessary requirement for the total sampling rate is mqf_p . Because RSSR uses a single ADC with a sampling rate of W_{BPF} , we can set the ADC sampling rate to $W_{BPF} = mqf_p$. For the simulation, we set the conditions as $m = 4$, $f_{nyq} = 4\text{GHz}$, $f_s = 220\text{MHz}$, $f_p = 31.5\text{MHz}$, $T_{acq} = 1.11\mu\text{s}$, and $B = 0.1f_p$ or $0.6f_p$. We acquired empirical simulation results from the MWC by the simultaneous orthogonal matching pursuit (SOMP) algorithm [15]. The simulation found that the probability that many samples will be split signals was lower than it would have been in theory because superposition is avoided when the test signal is generated. As shown in Fig. 3, compared to the MWC’s performance at $\kappa_{MWC} = 4$, the RMPI could not acquire the signals with $B = 0.6f_p$ even though the more successful sampling criteria with $\kappa_{RMPI} = 9$ was applied. For that signal bandwidth, high sparsity occurred, which cannot be recovered by the SOMP algorithm. This advantage of the MWC is one reason that we adopted it for our radar ES system design. In addition, use of (13) helped facilitate the design process by allowing easy prediction of the system’s signal-acquisition performance in terms of system parameters.

IV. SPLIT-SYNTHESIS METHOD

The high computational complexity makes the signal reconstruction consumes longer time and may result in failure in continuous signal acquisition. A failure occurs when the signal reconstruction time for the acquired sample obtained over the preceding acquisition time exceeds the acquisition time for the next sample acquisition. Even for the SOMP algorithm [15] which is one of the simplest signal reconstruction algorithms, the reconstruction time can easily exceed the acquisition time. Thus, to reduce the rate of failure of continuous signal reconstruction over a long acquisition time, careful new design on reducing the computational complexity is needed. Here we propose a split and synthesis method. Given a long signal acquisition time, the split-synthesis method trades-offs the computational complexity with the performance of signal reconstruction. Fig. 4 depicts the split-synthesis method.

A. SPLIT-PROCESS

After the MWC samples an aggregated radar signal over a long acquisition time T_{acq} , for example 0.13msec [9], the radar ES system imposes a uniform grid on the acquisition time at intervals of time slot T_{slot} . The aggregated radar signals in (1) are then reformulated as follows:

$$x(t) = \sum_{j=1}^G x_j(t - (j - 1)T_{slot}) \quad (18)$$

for $0 \leq t < T_{slot}$, where $x_j(t)$ corresponds to the slice of aggregated radar signal in the j -th time slot. The number of time slots G is chosen to reduce computational complexity, and this choice is discussed later in this section. For the j -th time slot, the MWC output (3) can be expressed as follows:

$$\mathbf{Y}_j[n'] = \mathbf{C}\mathbf{Z}_j[n'], \quad (19)$$

where the measurement matrix is $\mathbf{Y}_j \in \mathbb{R}^{m \times \tilde{l}_d}$, $\mathbf{C} \in \mathbb{C}^{m \times L}$ is the sensing matrix, $\mathbf{Z}_j \in \mathbb{R}^{L \times \tilde{l}_d}$ contains information of the aggregated radar signal, and

$$\tilde{l}_d = \frac{T_{slot}}{T_s} = q \cdot \frac{T_{slot}}{T_p} = q \cdot l_d. \quad (20)$$

The relation (19) can be represented in matrix form as follows:

$$\begin{pmatrix} y_{j,1} [1] & \cdots & y_{j,1} [\tilde{l}_d] \\ \vdots & & \vdots \\ y_{j,m} [1] & \cdots & y_{j,m} [\tilde{l}_d] \end{pmatrix} = \begin{bmatrix} c_{1,-L_0} & \cdots & c_{1,0} & \cdots & c_{1,L_0} \\ \vdots & & \vdots & & \vdots \\ c_{m,-L_0} & \cdots & c_{m,0} & \cdots & c_{m,L_0} \end{bmatrix}$$

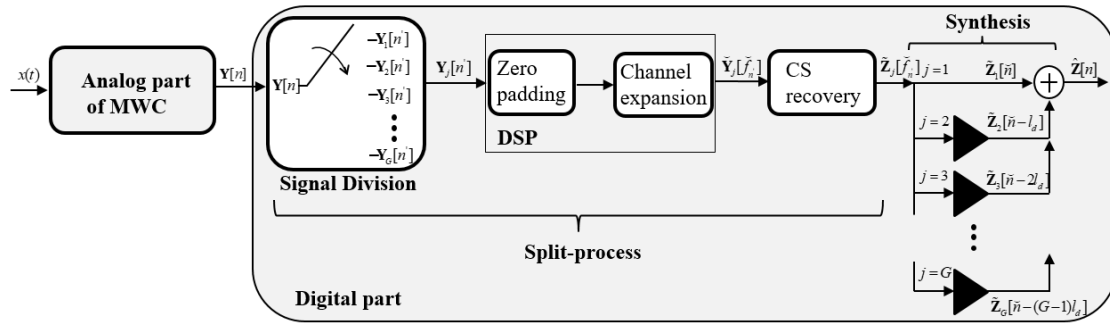


FIGURE 4. Block diagram of the proposed radar electronic surveillance system.

$$\times \begin{bmatrix} z_{j,-L_0}[1] & \cdots & z_{j,-L_0}[\tilde{l}_d] \\ \vdots & & \vdots \\ z_{j,0}[1] & \cdots & z_{j,0}[\tilde{l}_d] \\ \vdots & & \vdots \\ z_{j,L_0}[1] & \cdots & z_{j,L_0}[\tilde{l}_d] \end{bmatrix} \quad (21)$$

The next step is to extend the number m of channels in (19). We present a straightforward channel expansion method to enlarge the rows of the measurement matrix \mathbf{Y}_j and the sensing matrix \mathbf{C} in (19); we then adopt the zero-padding technique that alleviates time aliasing discussed Section II-D. Before expanding the channels, qk zeros are added to the right side of each row of \mathbf{Y}_j in (21), where $k < l_d$. Thereafter, as presented at (3), we exploit the way in which the $q = f_s/f_p$ sub-bands are piled in a single row of the \mathbf{Z}_j matrix. By disassembling the q piled sub-bands, the rows of \mathbf{Y}_j and \mathbf{C} can be expanded via a fast Fourier transform (FFT) and simple matrix reorganization. If the zero-paddings before the FFT are not included, severe time aliasing occurs because the results from FFT beyond \tilde{l}_d are lost, whereas they should be retained for the next timeslot. Thus, we can begin with padding zeros and performing the FFT on the right side of (21) to change the column indices of \mathbf{Y}_j as the frequency axis. In other words, $y_{j,i}[n']$ becomes $y_{j,i}[f_{n'}]$ for $f_{n'} = 1, 2, \dots, \tilde{l}_d + qk$.

By disjointing q piled sub-bands for the i -th row, we reorganize \mathbf{Y}_j to be a $qm \times (l_d + k)$ matrix such that

$$\check{y}_{j,i+v}[f_{n'}] = \sum_{v=0}^{q-1} y_{j,i}[v \cdot (l_d + k) + f_{n'}] \quad (22)$$

for $f_{n'} = 1, 2, \dots, l_d + k$. In the sensing matrix \mathbf{C} , by disjointing $q-1$ repeated $c_{i,l}$ times, such that $c_{i,M_0+s} = c_{i,-M_0+(s-1)}$ for $s \in [-q_0 + 1, q_0]$, \mathbf{C} is also expanded as follows:

$$\check{c}_{i+v,l} = \sum_{v=0}^{q-1} c_{i,-L_0+v+l} \quad (23)$$

for $l = 0, 1, \dots, M$. Consequently, with the channel expansion step, the relationship at (19) is transformed as follows:

$$\check{\mathbf{Y}}_j[f_{n'}] = \check{\mathbf{C}}\check{\mathbf{Z}}_j[f_{n'}], \quad (24)$$

where $\check{\mathbf{Y}}_j \in \mathbb{R}^{mq \times (l_d+k)}$, $\check{\mathbf{C}} \in \mathbb{C}^{mq \times M}$, and $\check{\mathbf{Z}}_j \in \mathbb{R}^{M \times (l_d+k)}$. The equal effect of qm enlarged equations helps to recover the input signals [14].

Compared to the conventional convolution method of (4), computational complexity is reduced by FFT expansion. For one channel, the computational complexity of the convolution method is $O(\tilde{l}_d^2)$, whereas the FFT method reduces it to $O(\tilde{l}_d \log \tilde{l}_d)$. When the FFT is performed in tandem with the split process, the computational complexity is reduced to $O((\tilde{l}_d + Gk) \log(\tilde{l}_d/G + k))$. Because the ES system samples radar signals over a long acquisition time, the effect of k additional zeros is negligible, i.e., $k \ll \tilde{l}_d$. To reduce the computational complexity, the number of timeslots G can be calculated as follows:

$$G = \arg \min_{G \in \mathbb{N}} (\tilde{l}_d + Gk) \log(\tilde{l}_d/G + k). \quad (25)$$

For example, when $l_d = 256$, $k = 2$, and $G = 76$, the computational complexities of the convolution expansion, FFT alone, and FFT with signal division are 65536, 2048, and 685, respectively. The reduced computational complexity of the split process shortens the system's computation time.

B. SYNTHESIS PROCESS

A CS algorithm recovers the signal information as $\tilde{\mathbf{Z}}_j[f_{n'}] \in \mathbb{R}^{4N \cdot (l_d+k)}$ by solving the MMV problem of (24), while $\tilde{\mathbf{Z}}_j[\tilde{n}]$ is generated by applying an inverse of FFT. The procedures from the split process through generation of $\tilde{\mathbf{Z}}_j[\tilde{n}]$ are repeated for each of the G timeslots, and the results from all the timeslots are synthesized as follows:

$$\hat{\mathbf{Z}}[\tilde{n}] = \sum_{j=1}^{G-1} \tilde{\mathbf{Z}}_j[\tilde{n} - (j-1)l_d] + \tilde{\mathbf{Z}}_G[\tilde{n} - (G-1)l_d], \quad (26)$$

where the matrix $\tilde{\mathbf{Z}}_G$ of the last timeslot decimates the columns in $[l_d + 1, l_d + k]$. As depicted in Fig. 4, each $\tilde{\mathbf{Z}}_j$ is delayed for l_d using buffers and then synthesized such that the $[l_d + 1, l_d + k]$ columns of $\tilde{\mathbf{Z}}_{j-1}$ are added to the k front-columns of $\tilde{\mathbf{Z}}_j$ to alleviate the time-aliasing. With (22), (23) and the synthesis process, we address the time-aliasing, as verified in Section VI.

Time-wise information, including PRI, TOA, and TOD, can be found from the reconstructed radar waveform or the estimation method presented in [16]. The carrier frequencies can be estimated with pulse spectrum density estimation [17].

Direction-of-arrival (DOA) can also be estimated using a crossed-loop/monopole antenna and a multiple signal classification (MUSIC) algorithm [18].

TABLE 1. Parameters of the radar ES system.

Parameters (Symbol)	Relationship	Value
Acquisition time (T_{acq})	$T_{acq} = G \cdot T_{slot}$	0.1302 ms
The number of time slots (G)	-	100
Nyquist frequency (f_{nyq})	$f_{nyq} \geq f_{max}$	4 GHz
Bandwidth (B)	$B_{min} \leq B \leq B_{max}$	≤ 31.49 MHz
◇ Channel expansion factor (q)	> 0	7
◇ Period of PR sequences (T_p)	$T_p = 1 / f_p$	0.03176 ns
◇ ADC sampling rate (f_s)	$f_s = q \cdot f_p$	220 MHz
◇ Number of PR pattern (M)	$M = L - q + 1$	127
◇ Physical channels (m)	-	4
◇ Virtual channels (\tilde{m})	$\tilde{m} = mq$	28

The dependent parameters are denoted as ◇ for a given hardware.

The parameters of the proposed system are listed in Table 1. For a given analog system, the PR sequence parameters, ADC sampling rate, and number of channels are the dependent parameters. The long acquisition time T_{acq} used to capture reliable radar signals rapidly increases the computational complexity in the digital signal reconstruction. Note that the long acquisition time also improves frequency resolution. However, the split-synthesis process can reduce the reconstructing time. In addition to the split-synthesis process over the long acquisition time, we provide a sub-sampling method to reduce the computational complexity of MMV algorithm for every timeslot.

V. MMV ALGORITHM PRE-PROCESSING

This section proposes a sub-sampling method which uses pre-processing to proportionally reduce the computational complexity of the MMV recovery algorithm in each timeslot. As discussed in the previous section, the radar ES system could greatly reduce the computational complexity of the channel expansion by splitting a long period of acquisition time into discrete timeslots. However, because there are still unnecessary measurement vectors in (24) and the total computational complexity of MMV recovery for G timeslots is still high owing to the long acquisition time, the reconstruction time can still exceed the acquisition time and cause bottlenecks. Because the MMV recovery algorithm involves matrix multiplication and/or inversion, the computational complexity of the algorithm rapidly increases with the number of measurement vectors. In addition, [19] shows that the recovery performance of the MMV algorithm is saturated with the number of measurements. Thus, we propose a sub-sampling method as a preliminary step to reduce the computational complexity of the following MMV algorithm. This method strategically selects a subset of measurement vectors without missing the support set which indicate the indices

of nonzero sub-bands. The sub-sampling method is detailed in Section V-A and the benefits in terms of computational complexity and support recovery performance are presented in Section V-B.

A. SUB-SAMPLING METHOD

By the linearity of (24), selecting columns of $\check{\mathbf{Y}}_j$ is equivalent to selecting columns of the signal matrix $\check{\mathbf{Z}}_j$. We therefore select the columns of $\check{\mathbf{Y}}_j$ based on the structure of the signal matrix $\check{\mathbf{Z}}_j$. The rows of $\check{\mathbf{Z}}_j$ contain spectrally orthogonal sub-bands of the discrete spectrum of $x(t)$ at intervals of f_p . From the discrete Fourier transform, the column indices represent the frequency grid in intervals of $1/T_{slot}$. Each narrow-band spectrum of $x_j(t)$ is contained within the rows. Some of the narrow-band spectra may be split by the borderline of the sub-bands based on their center frequencies.

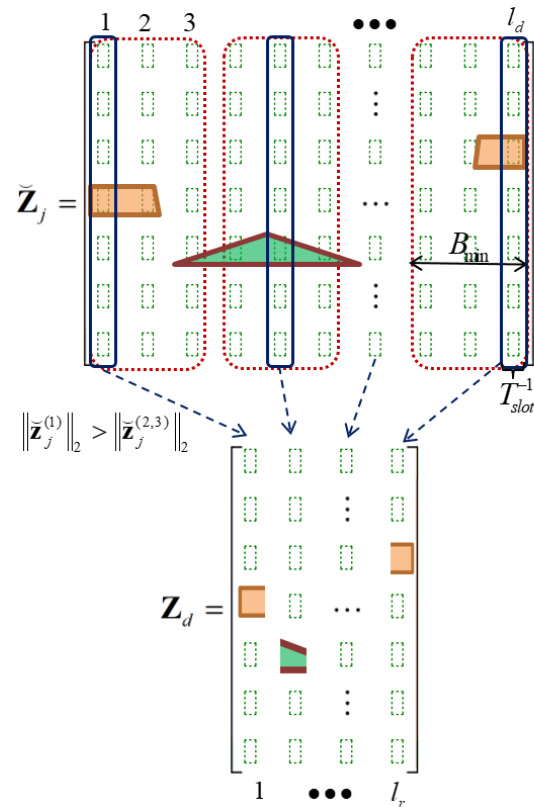


FIGURE 5. Selected columns of measurement matrix contain essential signal information for detecting a support set while reducing computational burden.

To address this scenario, we propose the sub-sampling method depicted in Fig. 5. This method generates subsets by classifying columns of $\check{\mathbf{Y}}_j$ at intervals less than the minimum signal bandwidth, B_{min} . For each subset, the sub-sampling method selects the column that has the maximum energy. When the subset comprises fewer columns than B_{min} , the method avoids the situation where several signals are present in a subset and do not overlap in columns. In this situation, only one of the signals within the sampling window

will be selected, whereas the others are missed. However, a sub-sampled matrix $\mathbf{Y}_d \in \mathbb{R}^{mq \times l_r}$, the union of selected columns, includes components of all of the signals while still reducing the size of the measurement matrix. The number of sub-sampled columns is calculated as follows:

$$l_r = \lceil l_d / \lfloor B_{\min} T_{sub} \rfloor \rceil, \quad (27)$$

where $\lfloor B_{\min} T_{sub} \rfloor$ is the element number of each subset. The notation \mathbf{Z}_j also becomes $\mathbf{Z}_d \in \mathbb{R}^{M \times l_r}$. Because this simple sub-sampling method works in one step before the following iterative MMV recovery algorithm, the computational complexity added by the sub-sampling is negligible. The reduced column l_r reduces computational complexity proportionally in the following MMV recovery algorithm. We chose the SOMP algorithm [15] as an example to verify the computational complexity benefits and the recovery performance of the support set.

B. DISCUSSION ON COMPUTATIONAL COMPLEXITY

To verify the computational complexity benefit of our sub-sampling method, we adopted the SOMP algorithm [15], which is commonly used to solve MMV problems. Note that the sub-sampling method is independent of SOMP. Although there are advanced MMV algorithms, such as MMV basic matching pursuit (M-BMP), Regularized MMV FOCUS (M-FOCUS), Bayesian, and group OMP (GOMP), these are inappropriate for our radar ES system because M-BMP has problems with signal reconstruction performance, and the other algorithms require high computational complexity. To implement our radar ES system with field programmable gate array, the high computational complexity becomes a problem. M-BMP works by matching a column of the sensing matrix with measurement vectors. However, according to terminal conditions, the algorithm provides an accurate solution with high sparsity or an inaccurate solution when the algorithm terminates with a predetermined sparsity amount [20]. Meanwhile, the other algorithms require higher computational complexities compared to SOMP [20]–[22]. Regularized M-FOCUS algorithms contain a concatenation of three matrices compared to the two matrices required of SOMP. Bayesian has a conversion of MMV to single measurement vector, and the dimension of the sensing matrix is increased with the number of measurements [22]. The number of rows and columns of the sensing matrix in GOMP are proportionally increased with the group size parameter [21]. From these reasons, we adopted SOMP because it not only shows the effect of our sub-sampling method but also it requires low complexity.

The SOMP is an iterative algorithm. At each iteration, the algorithm recovers the indices of nonzero rows of a signal matrix \mathbf{Z}_d , i.e., the support set, in (24) by matching the MMV matrix \mathbf{Y}_d with the bases of the sensing matrix \mathbf{C} . The procedure of SOMP is adjusted for the radar ES system to enhance the algorithm’s efficiency and reduce computational

complexity. The terminal condition is

$$\|\mathbf{Y}_d\|_2 \leq EPS. \quad (28)$$

When (28) is satisfied, SOMP determines that signals do not exist. Until this condition is satisfied, the algorithm continues to estimate a support among a set Λ , defined as row indices of \mathbf{Z}_d , which is expressed as follows:

$$J = \arg \max_{\Lambda} \left\| \mathbf{C}_{\Lambda}^H \mathbf{Y}_{residue}[i] \right\|_2, \quad (29)$$

where \mathbf{C}_{Λ} is a column of $\check{\mathbf{C}}$ and i is the iteration index. In the first iteration, the original MMV \mathbf{Y}_d replaces the residual matrix $\mathbf{Y}_{residue}$. From the conjugate symmetry of a real-valued radar signal, the selected and symmetric supports can be stored in S_i gathered at S , i.e., $S = \{S_i | i = 1, 2, \dots, N\}$. After estimating the support set S containing $2i$ elements, the residual of \mathbf{Y}_d is generated by the following equation:

$$\mathbf{Y}_{residue} = \mathbf{Y}_d - \mathbf{C}_S \cdot \mathbf{C}_S^{\dagger} \mathbf{Y}_d, \quad (30)$$

where \mathbf{C}_S^{\dagger} is a Moore–Penrose pseudoinverse of the outcome by extracting the columns of S from $\check{\mathbf{C}}$ in (24). Note that N real-valued radar signals can yield up to $4N$ supports. SOMP detects the support set Ω for the upmost $2N$ iterations instead of $4N$ iterations. The signal information is reconstructed as follows:

$$\check{\mathbf{Z}}_j[\check{f}_n] = \mathbf{C}_{\Omega}^{\dagger} \cdot \check{\mathbf{Y}}_j[\check{f}_n]. \quad (31)$$

The result of (31) can be used for the synthesis process explained in Section IV-C.

To verify computational complexities, we focus on the matrix multiplication and inverse operations (29) and (30) in the algorithm because these are the main factors that enlarge computational complexity. For a time slot, the sizes of the measurement and sensing matrices are $\check{m} \times l_d$ and $\check{m} \times M$, respectively, where $\check{m} = mq$. For (29), the computational complexity is $O(M\check{m}l_d)$. In (30), the computational complexities are different for each i -th iteration owing to $\mathbf{C}_S \in \mathbb{C}^{\check{m} \times 2i}$, but we can ignore this effect for easy verification as long as $i < \check{m}$. Thus, the computational complexity of (30) becomes $O(\check{m}^2(l_d + 2) + 16\check{m})$. As a result, the total complexity for N signals becomes

$$O(2N(l_d\check{m}(M + \check{m} + 2) + 16\check{m})). \quad (32)$$

Next, we compute the computational complexity of the pre-processing method. The computational complexity becomes

$$O(2N(l_r\check{m}(M + \check{m} + 2) + 16\check{m})), \quad (33)$$

where $l_r < l_d$ is the number of sub-sampled columns. In (33), we see the computational complexity reduced in proportion to the small l_r .

Fig. 6 plots the support recovery rate versus the number of sub-sampled columns of the measurement matrix. The support recovery rate is defined as one when the recovered support set is a subset of the original signal. We simulated

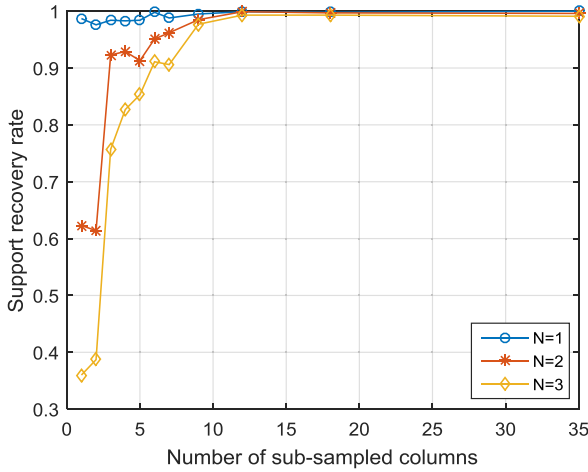


FIGURE 6. Support recovery rate versus the number of sub-sampled columns of measurement matrix using SOMP.

the support recovery rate along the number of subsets corresponding to the number of sub-sampled columns compared to the original columns, $l_d = 35$. The minimum bandwidth of the signal was $B_{\min} = 0.1f_p = 3.15MHz$. From (27), l_r is 12, which is a similar recovery performance to original full columns as shown in Fig. 6. This implies that the sub-sampled MMV \mathbf{Y}_d does not miss the signal information and includes all of the essential parts of the original MMV $\check{\mathbf{Y}}_j$. In this scenario, the sub-sampling method reduced computational complexity by a third. Consequently, we can conclude that this sub-sampling method proportionally reduces computational complexity while maintaining recovery performance.

VI. SIMULATIONS

Through simulations, we verified that the radar ES system can successfully trade-off between the reduction of computational complexity and degradation of signal reconstruction performance. For the simulations, we generated three pulsed radar signals whose carrier frequencies appeared randomly from $f_{\min} = 0.5GHz$ to $f_{\max} = 2GHz$. We gave as an input 5 dB pulsed radar signals where the signal to noise ratio (SNR) and signal type were not discussed. The SNR is defined as $10 \cdot \log(\|x\|_2^2 / \|n\|_2^2)$, where x and n are the input signals and noise vectors, respectively. We considered a four-channel MWC system; the remaining system parameters are listed in Table 1. The SOMP algorithm discussed in Section V was used to reconstruct the multiband signal.

First, we tested the improvement to the relative error in the reconstructed signal gained by the split-synthesis process of the radar ES system. In this simulation, the relative error was defined as follows:

$$relative_error[i] := \|x[i] - x_r[i]\|_2^2, \quad (34)$$

where $x[i]$ is the input radar value at the i -th time and x_r is the reconstructed radar vector. To clearly verify the reduction of the relative errors at the borders of the timeslots, we shortened

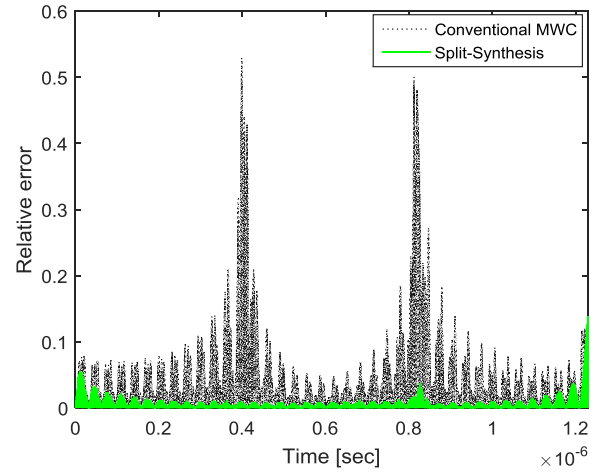


FIGURE 7. The relative errors at the borders of the timeslots are alleviated with the split-synthesis process of the proposed radar ES system.

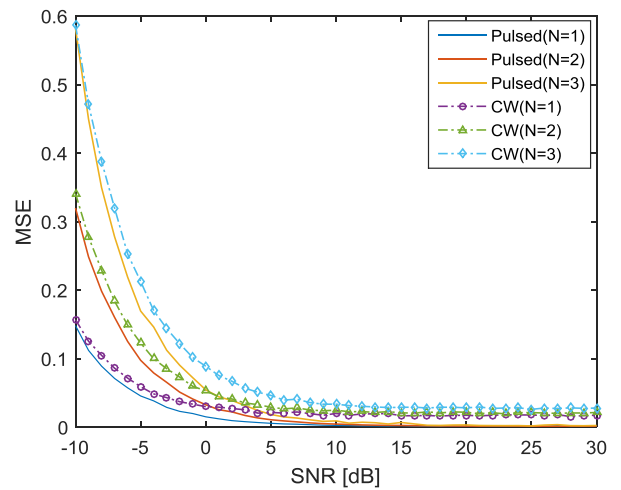


FIGURE 8. Reconstruction performance of radar ES system with pulsed signal and continuous wave (CW) for varying SNRs.

the acquisition period $GT_{slot} = 1.29\mu s$. As shown in Fig. 7, the errors among the timeslots are clearly reduced with the split-synthesis process. Although the trade-off parameter k discussed at the Section IV-A yields k additional columns in the channel-expanded measurement matrix, compared to the original column number $l_d = 35$, our simulation showed that even a small $k = 2$ yields some benefit. The Table 2 compares computational complexity between the conventional MWC and the split-synthesis process of the proposed radar ES system. Note that the order of computational complexity was discussed at the Section IV-A. With the small trade-off parameter $k = 2$, the split-synthesis of (22)-(26) can considerably reduce the computational complexity at the cost of negligible degradation of signal reconstruction as shown in the Fig. 7.

Second, we verified the robustness of noise along SNRs for continuous waves (CWs) and pulsed signals [9]. In this simulation, the MSE is defined as $\|x - x_r\|_2^2 / \|x\|_2^2$. As can be seen in Fig. 8, the pulsed signal was reconstructed better than

TABLE 2. Comparison of computational complexity.

	Conventional MWC	Split-Synthesis
Order of complexity	$O(G^2 l_d^2)$	$O((\tilde{l}_d + Gk) \log(\tilde{l}_d / G + k))$
Calculated complexity	11,025	155

The simulation parameters are $l_d = 35$, $G = 3$ and $k = 2$, which are also used in Fig. 7.

the CW. For the purpose of this study, our radar ES system can successfully monitor up to three radar signals above 2 dB as we pursued the detection of three radar signals under 5% of the MSE.

VII. CONCLUSION

In this study, we verified and compared the signal-acquisition performances among RSSR, RMPI, and MWC with a novel probability analysis. In the analysis, the MWC performed better than the other receivers. In addition, this analysis may be extended for comparison between CS-based sub-Nyquist receivers. Our proposed radar ES system with MWC was able to monitor incoming wideband signals in a simulation. In this ES system, the split-synthesis process considerably reduced the computational complexity with the trade-off parameter to alleviate the degradation of signal reconstruction. Pre-processing with sub-sampling before the MMV algorithm was able to proportionally reduce computational complexity while maintaining signal recovery performance. We plan to implement and test this signal-acquisition methods using the hardware that is currently being developed. As a future work, it would be meaningful to consider the problem of estimating the PDW of radar signals, including spatial location of carrier frequencies.

REFERENCES

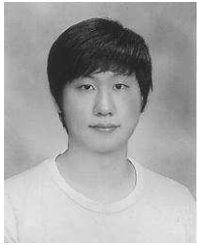
- [1] G. Schrick and R. G. Wiley, "Interception of LPI radar signals," in *Proc. Radar Conf.*, May 1990, pp. 108–111.
- [2] A. B. Carlson, P. B. Crilly, and J. C. Rutledge, *Communication Systems: An Introduction to Signals and Noise in Electrical Communication*, 4th ed. New York, NY, USA: McGraw-Hill, 2002.
- [3] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 375–391, Apr. 2010.
- [4] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Devices and Syst.*, vol. 5, no. 1, pp. 8–20, 2011.
- [5] J. Yoo, S. Becker, M. Monge, M. Loh, E. Candes, and A. E. Neyestanak, "Design and implementation of a fully integrated compressed-sensing signal acquisition system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2012, pp. 5325–5328.
- [6] J. Yoo et al., "A compressed sensing parameter extraction platform for radar pulse signal acquisition," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 3, pp. 626–638, Sep. 2012.
- [7] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [8] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
- [9] M. A. Richards, J. Scheer, W. A. Holm, and W. L. Melvin, Eds., *Principles of Modern Radar*. Raleigh, NC, USA: SciTech, 2010.
- [10] L. Bin, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 69–77, Nov. 2005.
- [11] X. Liu, W. Li, J. Wei, and L. Cheng, "Adaptable hybrid filter bank analog-to-digital converters for simplifying wideband receivers," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1525–1528, Jul. 2017.
- [12] A. V. Oppenheim and R. W. Schaffer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2010.
- [13] S. L. Miller and D. G. Childers, *Probability and Random Processes: With Applications to Signal Processing and Communications*, 1st ed. Burlington, NJ, USA: Elsevier, 2004.
- [14] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [15] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.
- [16] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4719–4734, Oct. 2011.
- [17] M. Mishali, A. Elron, and Y. C. Eldar, "Sub-Nyquist processing with the modulated wideband converter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 3626–3629.
- [18] Y. Tian, B. Wen, J. Tan, and Z. Li, "Study on pattern distortion and DOA estimation performance of crossed-loop/monopole antenna in HF radar," *IEEE Trans. Antennas Propag.*, vol. 65, no. 11, pp. 6095–6106, Nov. 2017.
- [19] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.
- [20] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.
- [21] A. T. Abebe and C. G. Kang, "Iterative order recursive least square estimation for exploiting frame-wise sparsity in compressive sensing-based MTC," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1018–1021, May 2016.
- [22] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 912–926, Sep. 2011.



JEONG PARK received the undergraduate degree in electrical engineering and computer science from the Kyungpook National University, Daegu, South Korea, in 2015. He is currently pursuing the M.S. degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea. His research interests include sub-Nyquist sampling, compressed sensing, estimation, and wireless communication.



JEHYUK JANG received the B.S degree in electronic engineering from the Kumoh National Institute of Technology, Gumi, South Korea, in 2014, and the M.S degree in information and communication engineering from the Gwangju Institute of Science and Technology, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science. His research interests include sub-Nyquist sampling and compressed sensing.




SANGHUN IM received the B.S. degree in electronics engineering from Soongsil University, Seoul, South Korea, in 2009, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2011 and 2016, respectively. He is currently with Hanwha Systems, South Korea. His current research interests include communication theories and signal processing for wireless communications, and physical layer security.



HEUNG-NO LEE (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He was with the HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member from 1999 to 2002. From 2002 to 2008, he was an Assistant Professor with the University of Pittsburgh, PA, USA. In 2009, he then moved to the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea, where he is currently affiliated. His areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He has received several prestigious national awards, including the Top 100 National Research and Development Award in 2012, the Top 50 Achievements of Fundamental Researches Award in 2013, and the Science/Engineer of the Month in 2014.

• • •

SCIENTIFIC REPORTS



OPEN

Visible and UV photo-detection in ZnO nanostructured thin films via simple tuning of solution method

Richa Khokhra¹, Bandna Bharti¹, Heung-No Lee² & Rajesh Kumar^{1,2}

This study demonstrates significant visible light photo-detection capability of pristine ZnO nanostructure thin films possessing substantially high percentage of oxygen vacancies (V_{OS}) and zinc interstitials (Zn_{IS}), introduced by simple tuning of economical solution method. The demonstrated visible light photo-detection capability, in addition to the inherent UV light detection ability of ZnO, shows great dependency of V_{OS} and Zn_{IS} with the nanostructure morphology. The dependency was evaluated by analyzing the presence/percentage of V_{OS} and Zn_{IS} using photoluminescence (PL) and X-ray photoelectron spectroscopy (XPS) measurements. Morphologies of ZnO viz. nanoparticles (NPs), nanosheets (NSs) and nanoflowers (NFs), as a result of tuning of synthesis method contended different concentrations of defects, demonstrated different photo-detection capabilities in the form of a thin film photodetector. The photo-detection capability was investigated under different light excitations (UV; 380–420 nm, white; $\lambda > 420$ nm and green; 490–570 nm). The as fabricated NSs photodetector possessing comparatively intermediate percentage of $V_{OS} \sim 47.7\%$ and $Zn_{IS} \sim 13.8\%$ exhibited superior performance than that of NPs and NFs photodetectors, and ever reported photodetectors fabricated by using pristine ZnO nanostructures in thin film architecture. The adopted low cost and simplest approach makes the pristine ZnO-NSs applicable for wide-wavelength applications in optoelectronic devices.

Photodetectors have a wide range of applications in many important areas such as; space communication, air quality monitoring, flame monitoring, industrial quality control, optical imaging, optoelectronic circuits, military surveillances etc.¹. Conventional photodetectors employ crystalline semiconductor materials such as; silicon, germanium, gallium arsenide etc. However, in these materials certain issues still need to be addressed; for instance, requirement of high temperature conditions for device fabrication, possibility of blurring, cross talk of optical signals between neighboring pixels² and limited freedom in material design. To overcome these problems, studies on inorganic semiconductor nanostructures^{3,4} such as; ZnS, InSe, CdS, CdSe etc. and metal-oxide semiconductor⁵ such as; ZnO⁶, CeO₂⁷, V₂O₅³ etc. could pave the way of fabricating a suitable photodetector. Nonetheless, these materials in nanostructure form provide a higher degree of freedom for material's properties tuning as well as the reduced dimensionality of the active device^{8,9}. However, as a key issue, these materials in their pristine form work only for ultra violet (UV) photo-detection applications, as allowed by their wide bandgap structure. Most of the studies correlate their wide bandgap with UV applications; however, in addition to the UV applications there are many areas that urgently require photodetector's sensitivity for visible-light region, and thus there is a great need to achieve a wide spectral response of the proposed nanostructured semiconductor materials. In other words, the widening of photodetector's spectral response (extended wavelength photo-detection) would enhance their application area. In this view, studies on the detection of visible spectrum by achieving a broadband photo-detection capability of nanostructured semiconductor materials, specifically metal-oxide semiconductors, have attracted a great attention in the last few years^{7,10}.

The wide spectral applications of metal-oxide semiconductor materials require tuning of optical properties (bandgap) of semiconductor nanostructures; therefore, normally they are doped with metals¹¹, non-metals¹², combined with other materials/functional groups^{13–15}, and formed as composites with another semiconductor materials^{1,7,16}. However, it is noteworthy that most of these processes applied for tuning of optical properties, require complicated and expensive equipments, and a complex device structure to achieve visible-light detection

¹Jaypee University of Information Technology, Wanknaghat, Solan, 173234, India. ²Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro Buk-gu, Gwangju, 61500, South Korea. Correspondence and requests for materials should be addressed to H.-N.L. (email: heungno@gist.ac.kr) or R.K. (email: rajesh.kumar@juit.ac.in)

response. Moreover, in these approaches, the requirement of high temperature and pressure conditions is another issue. In this concern of application, when considering morphology, the one dimensional (1D) nanostructures owing to their large surface to volume ratio and Debye length (that influence electronic/optical properties and thus exhibiting superior photosensitivity) show better performance in the photo-detection application^{3,5}.

While talking about metal-oxide semiconducting materials, investigated for photo-detection applications, the ZnO is found most promising candidate due to its many peculiar properties such as; high efficiency, low cost, non-toxicity, stability, high temperature operation capability, and environmental compatibility. Looking to the superior photosensitivity of morphologically 1D nanostructure, the ZnO itself is also studied mostly in 1D form with various modifications such as; decorating with gold nanoparticles¹⁷ and CdS¹⁸, doping with Cu¹⁹, Mn²⁰ and making heterostructures^{21,22} to show multi-spectral visible and UV light photo-detection capability. In principle, in the modifications, mid-gap electronic levels of dopant/s are introduced which generate charge carriers upon visible-light irradiation and thus make the material sensitive to visible light.

In a further and recent advancement¹⁰, undoped ZnO nanowires with a vertical alignment have been presented to exhibit an extended-range of visible light photo-detection capability upon annealing in hydrogen gas. The hydrogen annealing creates porosity on the surface of the nanowires that makes the assembled nanowire photodetector as a visible-light sensor by the phenomenon comprising antireflection, multiple scattering and defect state excitation induced mechanism. Similarly, the undoped ZnO structures have demonstrated visible-light activity upon vacuum deoxidation²³, where oxygen energy levels are introduced in the energy band of ZnO. The introduction of oxygen levels in the band gap enables ZnO as an active material for visible and UV-light photo-detection. Despite of many efforts, in the case of undoped ZnO nanostructures, it is still a challenging issue to make a simple and economical photodetector that could use an easily fabricated nanostructure to work in broad spectrum region (ultraviolet and visible), except the 1D nanostructures, and avoids the require sophisticated instrumentation. From the reported studies, it is ensured that for an undoped/pristine ZnO photodetector, it is only the multiple scattering or defects (V_o s, Zn_i s and antisites) that enables the visible-light/broadband response. Therefore, the tuning of morphology of undoped ZnO nanostructures, capable of broadband spectral-response through the combined effect of multiple scattering and formation of V_o s and Zn_i s, synthesized by the simple solution method would be highly valuable for making an economical photodetector.

In this work, different morphologies of *undoped ZnO nanostructures*; NSs, NFs and NPs were formed simply by low temperature chemical route engineering. The V_o s and Zn_i s were introduced intentionally during the formation, so as to avoid the cost effective approaches²³ which involve high temperature and vacuum conditions for inclusion of V_o s and Zn_i s. Thus generated V_o s and Zn_i s levels in the energy band of ZnO exhibited significant photo-response in the visible as well as UV spectral region. The photo-response of these nanostructures was investigated in terms of generated photovoltages by wide range of spectral illumination i.e. $\lambda \approx 380\text{--}420\text{ nm}$ (UV), $\lambda \approx 490\text{--}560\text{ nm}$ (green), and $\lambda > 420\text{ nm}$ (white light). We found that from the fabricated nanostructures, the NSs photodetector in a thin film form shows a faster rise and decay time both in the UV and visible spectrum region than that of the co-fabricated NPs and NFs photodetectors, and the ever reported sophisticated single ZnO nanowire based photodetector working only in UV region^{24–26}. Based upon the observations, simply fabricated NSs, a low-cost photodetector could be a highly competent candidate for the applications requiring detection of wide range spectrum.

Results and Discussion

Morphological and X-ray analysis. Surface morphology of the samples prepared by varying the synthesis parameters such as; variation in the concentration of precursor solutions, ratio, reaction time and solvent, was investigated using FE-SEM (Supplementary Figures S2 and S3 for C_2H_5OH medium and S4, S5 for H_2O medium). The large number of synthesis reaction performed using the precursor ' $ZnCl_2$ ' in solvent media C_2H_5OH and H_2O resulted mainly in three types of nanostructures i.e. NPs (for 15 minutes of reaction time in both the reaction media), NSs and NFs that were formed as the only distinguishable forms of the product as shown in Fig. 1. The solvent media C_2H_5OH and H_2O play a significant role in the different aggregations of initially formed nanoparticles that resulted in NSs and NFs after 4 hours of reaction time. The FE-SEM images show that an abrupt addition of $ZnCl_2$ solution in alkali solution resulted preliminary in the formation of NPs (Fig. 1a) for both the solvent media, which then aggregated differently as NSs (Fig. 1b) and NFs (Fig. 1c), respectively. The XRD patterns are shown in the right of Fig. 1. There are five prominent diffraction peaks, in all the cases, at diffraction angles $2\theta = 32.7^\circ, 34.5^\circ, 36.42^\circ, 47.44^\circ$ and 56.58° , which are indexed as lattice planes (100), (002), (101), (102), and (110) with the lattice constants ($a = 0.325\text{ nm}$ and $c = 0.5211\text{ nm}$), corresponding to Wurtzite crystal structure of ZnO. The size of nanocrystals estimated for maximum intensity peak, using Debye-Scherrer formula $D = 0.9\lambda/\beta\cos\theta$ is 17.41 nm, 21.57 nm and 23.29 nm for NP, NS and NF, respectively; with an average crystallite sized 20.16 nm, 24.44 nm and 20.44 nm for NP, NS and NF, respectively, indicates a successive growth mechanism of nanostructures evolving from the nanoparticles.

Optical absorbance of the samples obtained by UV-Visible spectroscopy is shown in Fig. 2. These absorption spectra appear to be extended from UV to visible-region, with a sharp UV excitonic peak at wavelength, 315 nm for NPs (Fig. 2a), 355 nm for NSs (Fig. 2b) and 365 nm for NFs (Fig. 2c). The presence of excitonic peak in UV region along with an extended absorption region reveals their UV as well as visible-light activity¹⁰. The shift in the excitonic peak from NPs (315 nm) to NSs (355 nm) and then to NFs (365 nm), corresponds to a red shift in the spectrum. With the shift in the excitonic peaks, the overall absorbance also increases from NPs to NSs and then to NFs (Fig. 2d). This result is analogous to the observations of enhanced absorbance in the porosity induced antireflections leading to visible-light photo-activity¹⁰. A rough estimation of the porosity order as NFs > NSs > NPs, can be made from the FE-SEM images of ZnO films (Fig. 1). Among these nanostructures, the NFs are expected to have comparatively more multiple reflections of light rays once they enter the film, and these large multiple

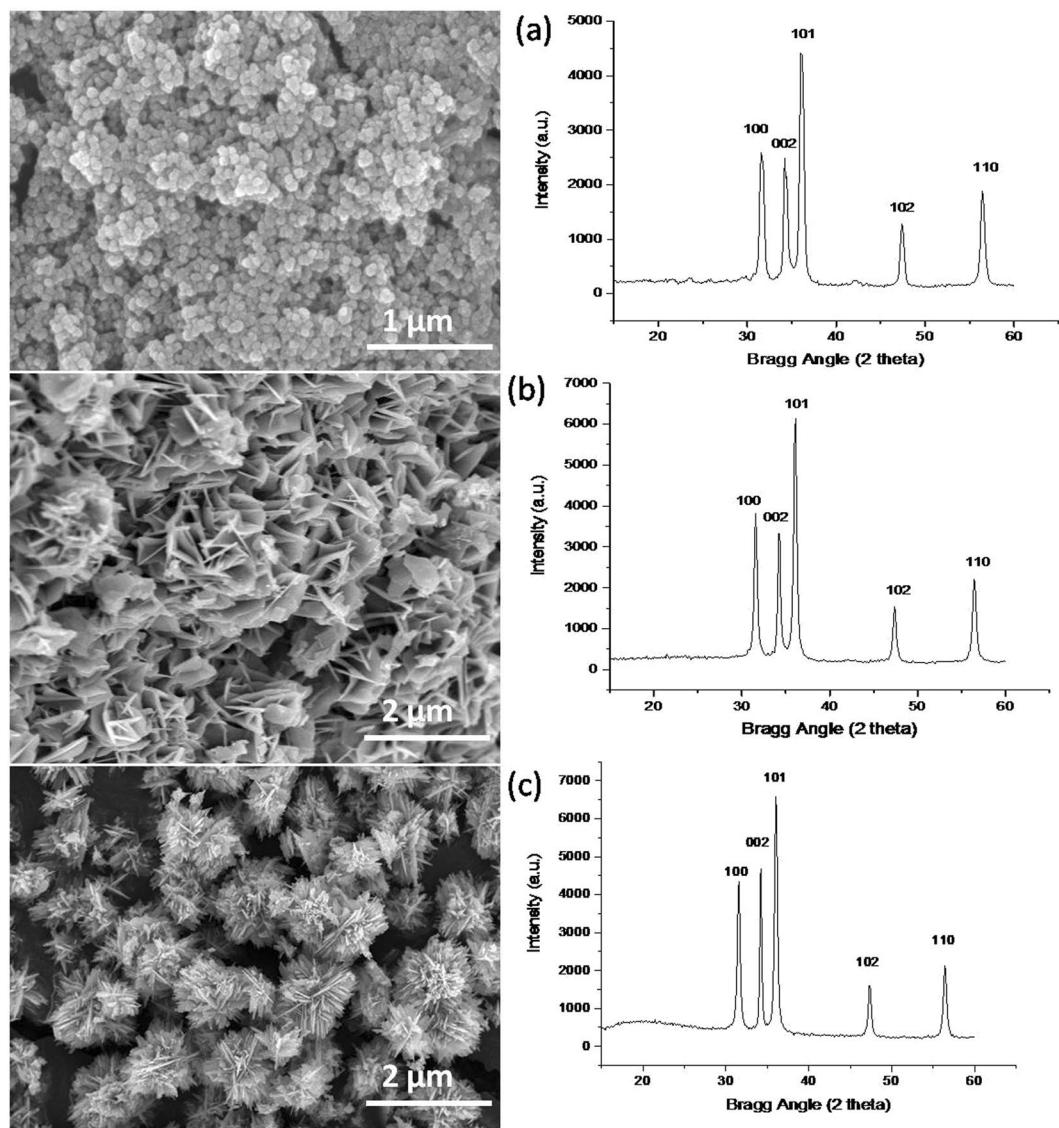


Figure 1. FE-SEM images and XRD results of ZnO nanostructures. (a) NPs formed in 15 minutes by using C_2H_5OH solvent, (b) ZnO-NSs formed in 4 hours using C_2H_5OH solvent and (c) ZnO-NFs formed in 4 hours using H_2O solvent. In the right side, there are XRD plots indicating similar crystallographic structures for all the morphologies.

reflections may lead to higher absorption¹⁰ in the NFs film in comparison with NSs and NPs films as shown in Fig. 2d.

PL and XPS studies. Emission/absorption in the UV and visible-region by ZnO nanostructures were investigated via analyzing defects states (vacancies, interstitials and antisite) in the studies conducted by PL and XPS experiments. Figure 3 shows a room temperature PL spectrum of NPs, NSs and NFs nanostructures. For excitation energy ~ 3.36 eV (corresponding to wavelength ~ 370 nm, equivalent to the typical band gap of ZnO), the PL spectra of NPs, NSs and NFs samples show near-band emission; respectively, at 386 nm (3.21 eV) (Fig. 3a), 393 nm (3.15 eV) (Fig. 3b) and 396 nm (3.12 eV) (Fig. 3c) which is attributed to the transitions from excitonic levels and/or zinc interstitials (Zn_i) to the conduction band (CB)²⁷ and is analogous to the previous studies^{28,29}. Along with these near-band emissions, all three samples exhibit visible emissions in the spectrum region 420–569 nm (Fig. 3d) consisting distinct peaks at 420 nm (2.95 eV), 456 nm (2.71 eV), 484 nm (2.56 eV), 511 nm (2.42 eV), 530 nm (2.34 eV) and 568 nm (2.43 eV). Basically, transition from VB to CB and from VB to shallow levels occurs upon photoexcitation in the PL, which then give the subsequent transitions; CB \rightarrow deep levels, shallow levels \rightarrow VB, shallow level \rightarrow deep levels and hole capture at deep levels gives violet, blue and green emissions according to energy levels difference. Since the exciting energy (3.36 eV) is equivalent to ZnO energy gap; therefore, the electrons excited from the valance band (VB) can jump to the CB as well as shallow defect levels.

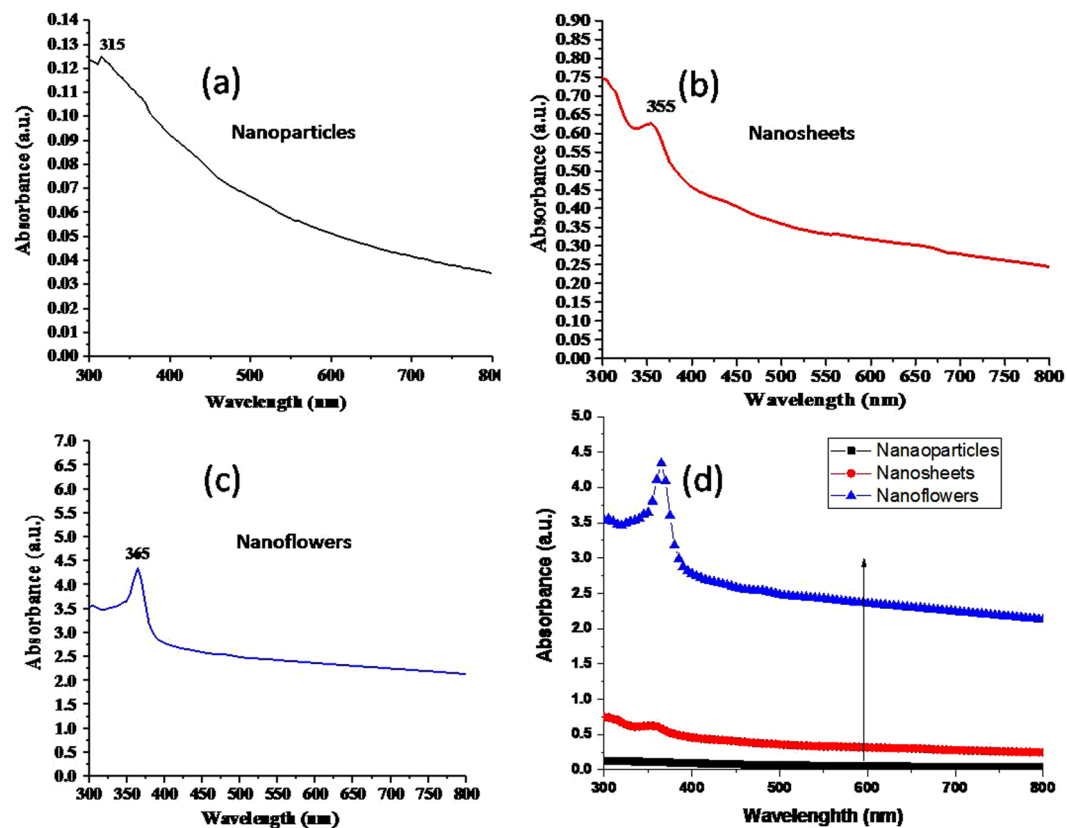


Figure 2. UV-Vis spectroscopy absorbance spectra. (a) NPs having excitonic peak at 315 nm and extended absorbance in the visible spectrum region, (b) NSs having excitonic peak at 355 nm and showing extended visible region, (c) NFs having excitonic peak at 365 nm with comparatively higher absorbance, and (d) shows increasing absorbance from NPs to NSs and then to NFs.

Generally, oxygen vacancies (V_o^* , V_o^+ and V_o^{++}) are considered as green emission centers, where V_o^* is neutralized oxygen vacancy that lies 0.86 eV below the CB, and the V_o^{++} a double ionized vacancy lies ~ 2.18 eV below the CB. The single ionized oxygen vacancy V_o^+ is reported to have two energy locations; 0.9 and 2.47 eV above the VB in the band gap^{30,31}. Besides the oxygen vacancies, oxygen antisite (O_{Zn}) located 2.33 eV below the CB do corresponds to green emission of shorter wavelength³⁰. Other defects such as; zinc interstitials Zn_i s, natural (Zn_i), singly and doubly ionized interstitials (Zn_i^+ , Zn_i^{++}) are responsible for blue emission^{27,32,33}. Zn_i being located 0.22 eV below the CB edge³⁴ gives violet emission around 390 nm in the PL. However, some studies show more deeper location of Zn_i (~ 0.37 eV below CB) to explain violet-blue emission²⁷. Other two Zn defects; Zn_i^+ and Zn_i^{++} lie 0.56 and 0.63 eV below the CB minima²⁷, respectively; whereby the transitions to zinc vacancy (V_{Zn}) and/or VB results in blue emissions of higher wavelengths³⁵. Except Zn_i s and V_{Zn} , the oxygen interstitial (O_i) generally located in the band gap at position 0.4 eV above the VB, also participates in the blue emission^{35,36}.

PL spectra in Fig. 3 show co-existence of violet emission peak in the range 386–398 nm (corresponding to excitonic emissions) and visible emission peak in the range 420–568 nm. All three types of nanostructures; NPs (3a), NSs (3b) and NFs (3c) show common visible light emission peaks at 420 nm (2.95 eV), 456 nm (2.71 eV), 484 nm (2.56 eV), 511 nm (2.42 eV), 530 nm (2.34 eV) and a low intensity peak at 568 nm (2.43 eV), whereas UV region emission peaks are located at different wavelengths. The UV emission peak shows a red shift in a order NPs \rightarrow NSs \rightarrow NFs as can be seen in Fig. 3d. This red shift in PL is analogues to the observed red shift in UV-Vis spectrum (Fig. 2), indicating slightly different excitonic energy levels in the band gap of synthesized nanostructures. As reported earlier ZnO possesses stable excitonic states just below its CB³³ minima, whereby transition to VB gives near band violet emissions. Thus in our case, the observed peak in PL of NPs at position 386 nm (3.21 eV) can be assigned to the transition from excitonic states to its VB. In other words, the decay of self-trapped exciton to the CB, causes near band violet emission as shown in Fig. 4. Similarly, the UV emissions in NSs and NFs at 393 nm (3.15 eV) and 398 nm (3.12 eV) are close to the electronic transition from a slightly lower energy excitonic state or Zn interstitial Zn_i (lying ~ 0.22 eV below the conduction band) to the VB. The possible transition scheme, corresponds to all the peaks in the PL shown in Fig. 4b, is given in Fig. 4a.

In the visible region, the observed blue emission peak at 420 nm (2.95 eV) corresponds to the transition from Zn_i to VB, considering that Zn_i lies ~ 0.41 eV deeper to the CB edge, this is alike to the previous reports²⁴, wherein energetic location of Zn_i is considered 0.38 eV deeper in the energy band. The another possibility is the transition CB \rightarrow oxygen interstitial O_i (located 0.4 eV above VB as proposed earlier³⁵) that also gives blue emission

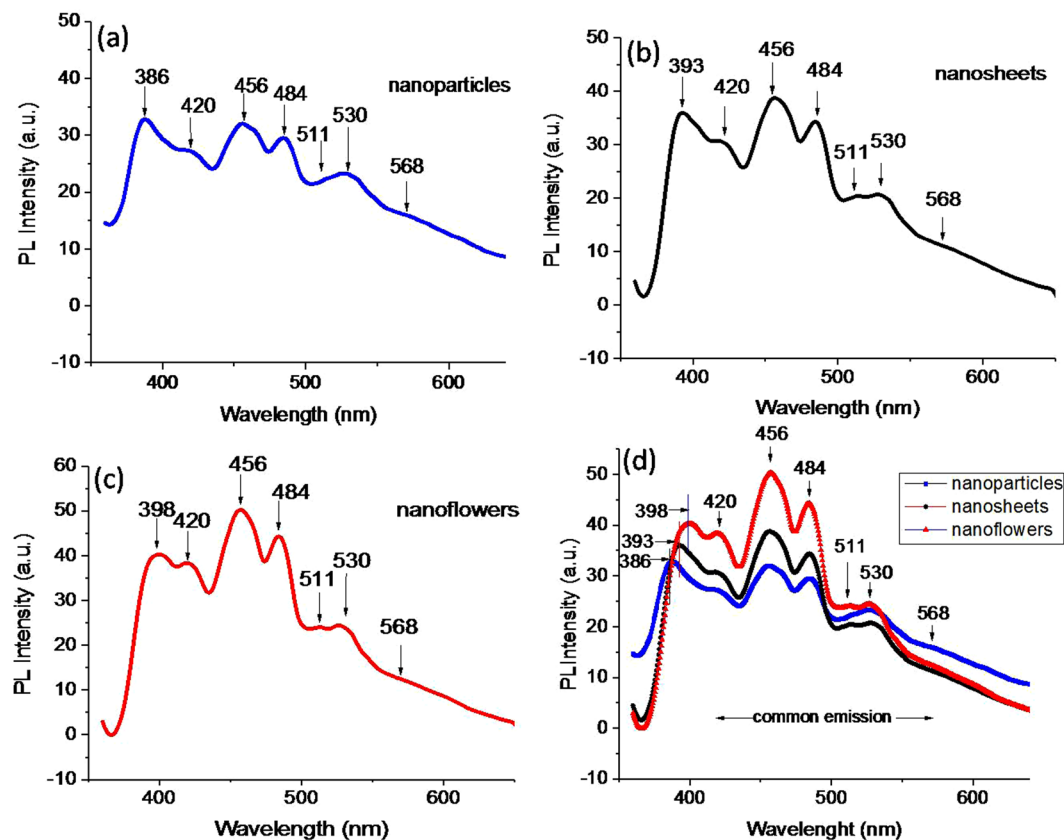


Figure 3. Photoluminescence spectra of nanostructures. (a) NPs show violet emissions at 386 nm, (b) NSs show violet emissions at 393 nm and (c) NFs show violet emission at 398 nm. Blue emissions at wavelengths positions 420, 456 and 484 nm and green emissions at wavelengths 511, 530 and low intensity peak at 568 nm are common in all nanostructures.

at 420 nm. In fact, the second one is rather in a good agreement as the energy level difference (2.96 eV) between CB edge and O_i is close to the obtained emission energy (2.95 eV). Nonetheless, the excitation energy (3.36 eV) is quite enough to pump the electrons from VB to CB that make a transition $CB \rightarrow O_i$ and emit blue radiation of 420 nm, indicating the latter transition more plausible. The another blue emission at ~ 456 nm (2.71 eV) is assigned to the transition from extended Zn_i states $\rightarrow VB$. The extended Zn_i states are generally; localized Zn_i states, Zn_i^{++} and complex defects, whose energetic locations deep in the band gap depend upon the fabrication methods^{27,35}. In our case, the blue emission (2.71 eV) is in close agreement to the energy of transition $Zn_i^{++} \rightarrow CB$ (2.73 eV), indicating the presence of Zn_i^{++} states in the band gap. Third blue emission peak located at 484 nm (2.56 eV) corresponds to the transition from $Zn_i^+ \rightarrow V_{Zn}$, as the energy difference between these levels (2.52 eV) is in close agreement to the observed emission energy.

Green emissions falling in the spectral region 511–568 nm, possesses three emissions peaks at 511 nm (2.42 eV), 530 nm (2.34 eV) and 568 nm (2.18 eV). When the electrons from CB, recombine with doubly ionized V_o^{++} located at an energy level 1.12 eV above the VB, generate green emission of wavelength 568 nm³⁷. The second green emissions peak centered at 530 nm may come from the transition $V_o^+ \rightarrow VB$ or $CB \rightarrow O_{Zn}$. The first transition ($V_o^+ \rightarrow VB$) occurs due to the formation of unstable V_o^+ state of V_o^+ by capturing electrons from CB³⁸. This unstable state, when recombine with photoexcited hole in the VB, would generate green emission around 530 nm³⁹. The second transition $CB \rightarrow O_{Zn}$ also has a strong possibility as the exciting energy is enough to pump the electrons to CB that after falling to O_{Zn} will give rise a green emission at 530 nm. Coming back to the V_o^+ states, in the energy band, there may be occurrence of complex V_o^+ states along with isolated V_o^+ centers. The complex states lying deeper in the band gap also give a possible explanation of the green emission around 530 nm, whereas the isolated V_o^+ states suitably explain the emission around 511 nm³⁸ through the transition $V_o^+ \rightarrow VB$. As mentioned previously, two possible energetic locations of isolated V_o^+ states are estimated theoretically at 0.9 and ~ 2.47 eV above the VB. The transition $V_o^+ \rightarrow VB$ (corresponding to 2.47 eV energy level position of V_o^+) gives 511 nm emission as shown in the scheme of Fig. 4a, and the transition $CB \rightarrow V_o^+$ (corresponding to 0.9 eV energy level position of V_o^+) will give 510 nm emission as shown in the scheme of Fig. 4a by dashed line. Both of the transitions have equal possibility to generate emission around 510 nm in the PL. All these observations in the PL, indicate that all of the samples possess defects states such as; V_o^s , Zn_i , O_i , V_{Zn} and O_{Zn} .

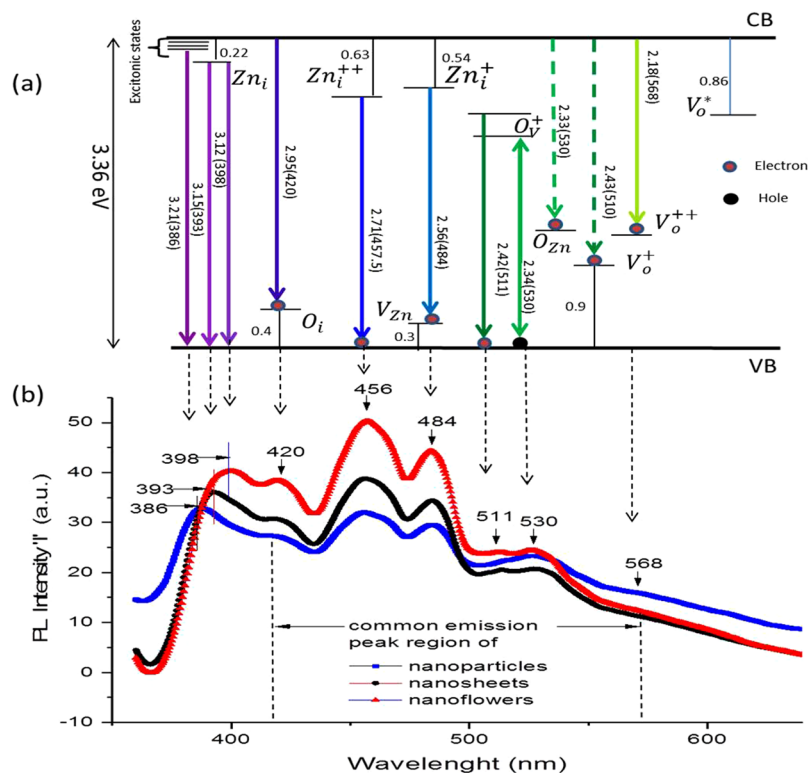


Figure 4. Emission scheme with PL spectra. (a) Schematic representation of emission scheme in the prepared samples of NPs, NSs and NFs nanostructures, (b) PL emission showing different peak position in violet region 386–398 nm, and common emission peaks in the visible region 420–568 nm in all samples.

XPS studies were performed to get information about chemical bonding and defects states present in the ZnO samples. Figure S6(a–c) shows XPS survey spectra recorded at room temperature for NPs, NSs and NFs samples. The overview of survey spectra reveals the presence of O1s and Zn2p ($Zn2p_{3/2}$ and $Zn2p_{1/2}$) peaks in all the samples. In order to further examination, the high-resolution peaks were deconvoluted in satellite components at different binding energies. Figure 5(a–c) illustrates high-resolution XPS spectra for all the samples corresponding to O1s core level. These spectra are fitted with three Gaussian peaks. In the NPs case (Fig. 5a), deconvoluted peaks are located at binding energies 530.8, 530.93 and 532.5 eV. Here, the lower binding energy peak is attributed to lattice oxygen (O_L) which contributes to the perfect hexagonal structure of ZnO lattice, the presence of middle peak at binding energy 530.93 eV is ascribed to vacancies (V_o s)⁴⁰ in ZnO lattice. The observation of V_o s supports the presence of green emission line in PL. The higher binding energy peak at 532.5 eV corresponds to chemisorbed oxygen (OH^- , $-CO_3$, adsorbed H_2O , and O_2 (O_C))^{41,42}. Similarly, the peaks in high-resolution deconvoluted XPS spectrum of NSs and NFs, when analyzed do correspond to O_L , V_o s and chemisorbed oxygen. However, there is shift in the binding energy values of the corresponding deconvoluted peaks for NSs and NFs with respect to that of NPs, which is ascribed to the difference in their morphologies and synthesis approaches^{43,44}. Further, the change in the percentage of oxygen content related to each deconvoluted peak was estimated by the change in percentage area of the peaks. From the calculations of percentage area, we found that the lattice oxygen in NPs is about 17%, whereas in case of NSs and NFs it reduces to ~ 12%. At the same time, the percentage area of oxygen vacancies (V_o^* , V_o^+ and V_o^{++})^{45,46}, increases from 21.8% (in NPs) to 47.7% (in NSs) and 54.5% (in NFs), and that of the chemisorbed oxygen decreases from NPs (60.3%) to NSs (40.9%) and then to NFs (37.6%).

Figure 6(a–c) represents high-resolution XPS spectra for Zn peaks in the NPs, NSs and NFs samples, respectively. In these spectra, the energy separation between two Zn components; $Zn2p_{3/2}$ and $Zn2p_{1/2}$ for NPs is 23.11 eV, NSs is 23.09 eV and NFs is 23.13 eV, which are in agreement with the reported values of ZnO^{47,48}. The difference of binding energies of these deconvoluted peaks for NPs, NSs and NFs (given in Table S2) is ascribed to the difference in their morphologies and synthesis approaches^{43,44}. The peak $Zn2p_{3/2}$ in NPs sample is deconvoluted in three peaks at binding energies 1022.02 eV (lower side), 1022.41 eV (middle) and 1022.35 eV (higher), also the peak $Zn2p_{1/2}$ is deconvoluted in three peaks at binding energies 1044.92 eV (lower), 1045.51 eV (middle) and 1045.64 eV (higher) as shown in Fig. 6a. The middle peaks (1022.41 and 1045.51 eV) have an energy difference of 23.1 eV, which is in good agreement with the spin-orbit splitting value of divalent Zn bounded in ZnO structure^{49,50}, and thus suggesting that the middle peak corresponds to lattice Zn. The lower energy peaks centered at 1022.02 and 1044.92 eV correspond to metallic Zn in the sample⁵¹. And, third peaks centered at higher energies 1022.35 and 1045.64 eV, correspond to +2 oxidation state of Zn due to the presence of $Zn(OH)_2$ or/and ionized Zn_i interstitials. It is found^{37,52} that the binding energy location of satellite peak of $2p_{3/2}$ lying between

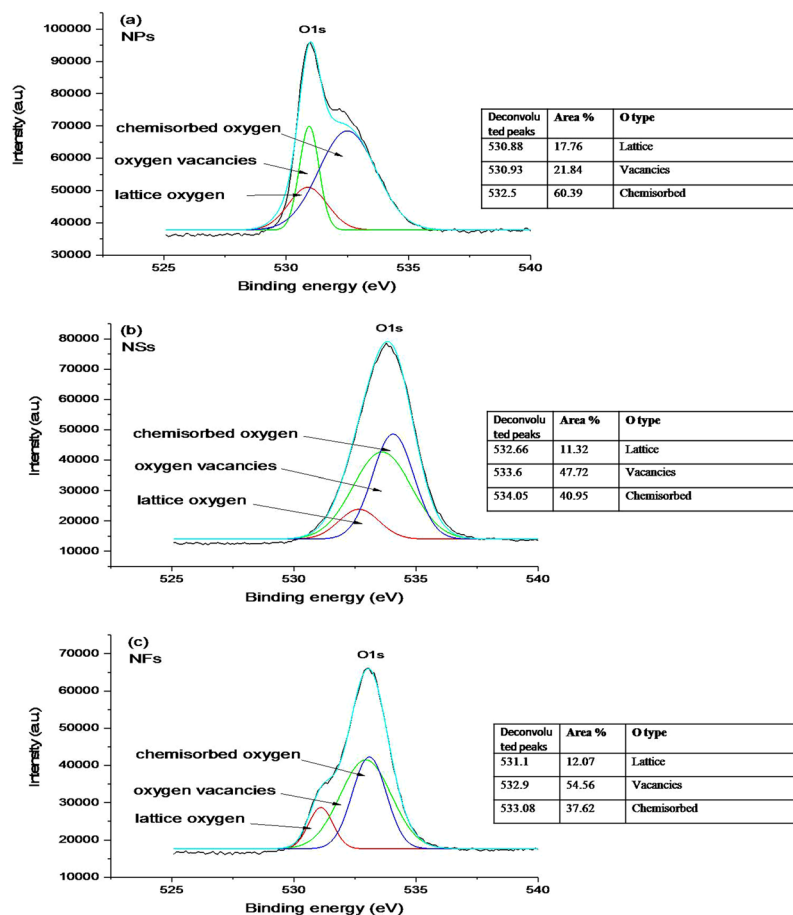


Figure 5. High-resolution XPS spectra for O1s. (a) ZnO-NPs different percentages of lattice oxygen, oxygen vacancies and chemisorbed oxygen, (b) ZnO-NSs and (c) ZnO-NFs show variations in the peak positions, and constituent percentages are given in respective tables.

1022.70 and 1021.80 eV corresponds to $\text{Zn}(\text{OH})_2$, and therefore the +2 oxidation state is due to the presence of $\text{Zn}(\text{OH})_2$. Whereas in our case for NPs sample, the satellite peak $2p_{3/2}$ (at 1022.35 eV) lies between the referred energy limits, which suggests the presence of $\text{Zn}(\text{OH})_2$ or +2 oxidation state by hydroxide form of Zn. This can be further correlated with the corresponding peak in the O1s spectrum of NPs (shown in Fig. 6a), which is assigned as chemisorbed oxygen (generally in OH^- and/or water molecules). In the O1s spectrum, the obtained higher percentage (60.39%) of chemisorbed oxygen can be assigned to OH^- group as confirmed from $\text{Zn}2p_{3/2}$ spectrum; however, there should co-exist a small percentage of interstitial Zn as well, as to produce emission in the corresponding PL spectrum (as the PL of nanoparticles shows the presence of interstitials). Moreover, the lower PL emission intensity of NPs than that of NSs and NFs indicates comparatively smaller interstitial percentage in NPs. Next, in case of NSs, each of the peaks $\text{Zn}2p_{3/2}$ and $\text{Zn}2p_{1/2}$ (Fig. 6b) is deconvoluted in three satellite peaks at energies 1022.81, 1023.62, 1024.16 eV and 1045.37, 1046.6, 1048.66 eV, respectively. The middle peaks centered at binding energies 1023.62 eV and 1046.6 eV have energy difference of 22.98 (approximately 23.0 eV), that again corresponds to lattice Zn with two valance state in ZnO. The satellite peak at lower energy side is assigned to metallic Zn alike to that of the NPs sample. However, in this sample the higher energy satellite peaks 1024.16 and 1048.66 eV located at significantly higher energies do not correspond to $\text{Zn}(\text{OH})_2$; instead it indicates that the Zn atom is surrounded by more than one oxygen atoms and is occupied at interstitial positions³⁷. These interstitial Zn could be neutral Zn_i , extended states of Zn_i ; single and double ionized zinc interstitials (Zn_i^+ and Zn_i^{++}). Similarly, in the samples of NFs, the Zn2p spectrum is deconvoluted in three peaks as shown in Fig. 6c. Here also, the higher energy satellite peaks are located at much higher energies as 1024.34 and 1074.04 eV should correspond to the interstitial Zn similar to that of NSs sample. However, the area percentage in NFs is more than that of NSs, suggesting a higher content of interstitials in NFs, which is in good agreement with PL observation (Fig. 4b) showing higher intensity, revealing the higher Zn interstitials in NFs sample.

Performance of ZnO photodetectors. To study the performance of ZnO photodetectors, photo-response was measured in terms of photovoltage by applying a bias voltage ' $V_b = 5\text{ V}$ ' as shown in Figure S1. Figure 7 shows photovoltage versus time plots for all the three types of nanostructures; NPs (Fig. 7a), NSs (Fig. 7b) and NFs (Fig. 7c) thin film photodetectors under the illuminations; violet, white and green. In the photovoltage measurements, rise time is the time required in 90% rise of photovoltage from its initial value (after switching ON

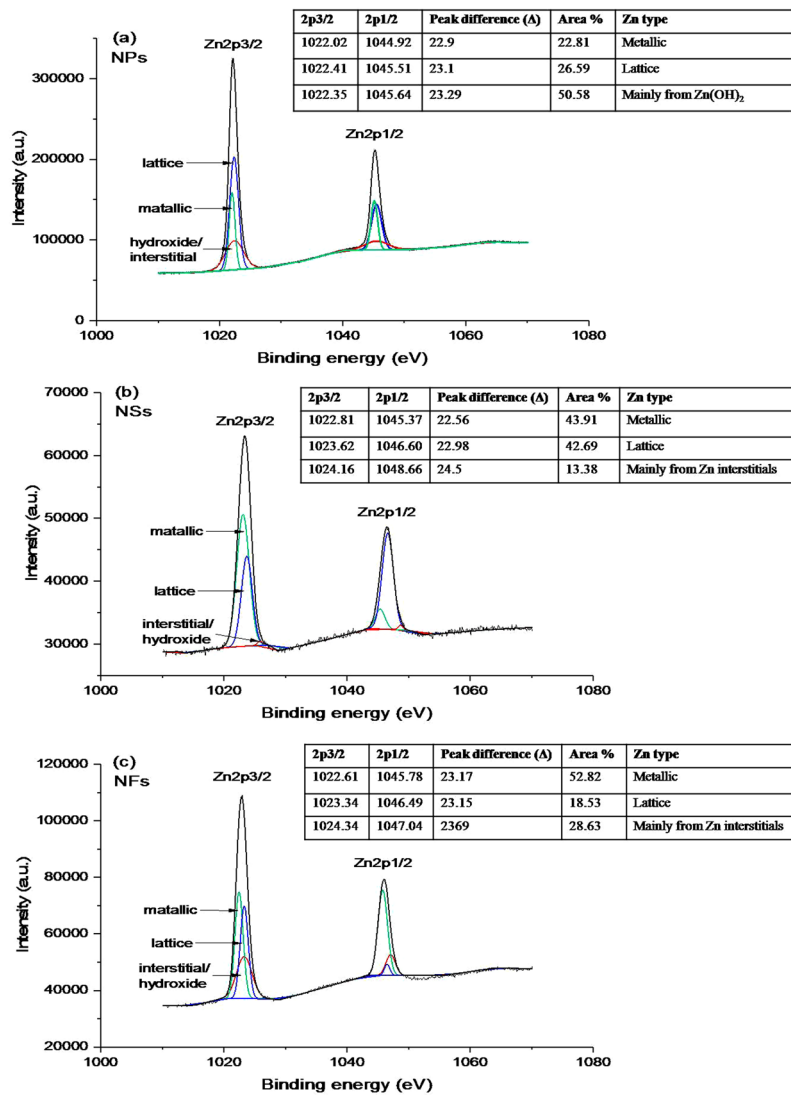


Figure 6. High-resolution XPS spectra of Zn2p. **(a)** ZnO-NPs possessing dominating Zn peak at higher energy side is due to the hydroxide form as mentioned in table, **(b)** ZnO-NSs and **(c)** ZnO-NFs spectra have dominating interstitial Zn peak at higher energy. The position of peaks and their area percentages are mentioned in the corresponding tables.

the illumination), and fall time is the time taken in the falling of maximum photovoltage to 10% (after switching ‘OFF’ the illumination)⁵³. For each photodetector, the period of ON and OFF time was taken different by considering their response time as different, and also to eliminate the heating effects on the sample surface^{39,54}. In the experiments, the ON and OFF time was controlled using a camera shutter. The photo-response of all ZnO nanostructure photodetectors (in the region ultraviolet to visible) is given in Table 1.

The mechanism of photodetector response with light illumination can be explained by desorption and adsorption of O₂ and/or H₂O on its surface^{26,55–58}. In a dark condition, O₂ and H₂O molecules present in the air get adsorbed on ZnO surface, which by capturing electrons from the conduction band of ZnO becomes negatively charged ions (chemisorbed) that results in the increased depletion barrier height. When the illumination is turned on, the chemisorbed O₂ gets H₂O is/are desorbed by two ways; (i) capturing a photo-generated hole and/or (ii) direct photo-excitation of captured electron to the conduction band of ZnO²⁴. The desorption mechanism occurs in accordance to the energy level of illuminating radiation (above-band or below-band energy). While using above-band illumination ‘UV radiation’, electron and hole are generated directly since the illumination energy is higher than the band gap of ZnO. Thus the produced photo-holes migrate to the chemisorbed ions sites, and release the electron by neutralizing the ions. The desorption of ions decreases barrier height and releases electrons in the conduction band of ZnO. These released electrons along with the photo-electrons would enhance the concentration of carriers in the CB of ZnO, and thus give rise to its photoconductivity. In case of below-band illumination, desorption of O₂/H₂O occurs by direct photo-excitation of the captured electrons to the conduction band of ZnO that also increases photoconductivity. In the present study, we used above-band (UV) as well as below-band (white and green) illuminations which resulted in the generation of photovoltage. Thus the observed

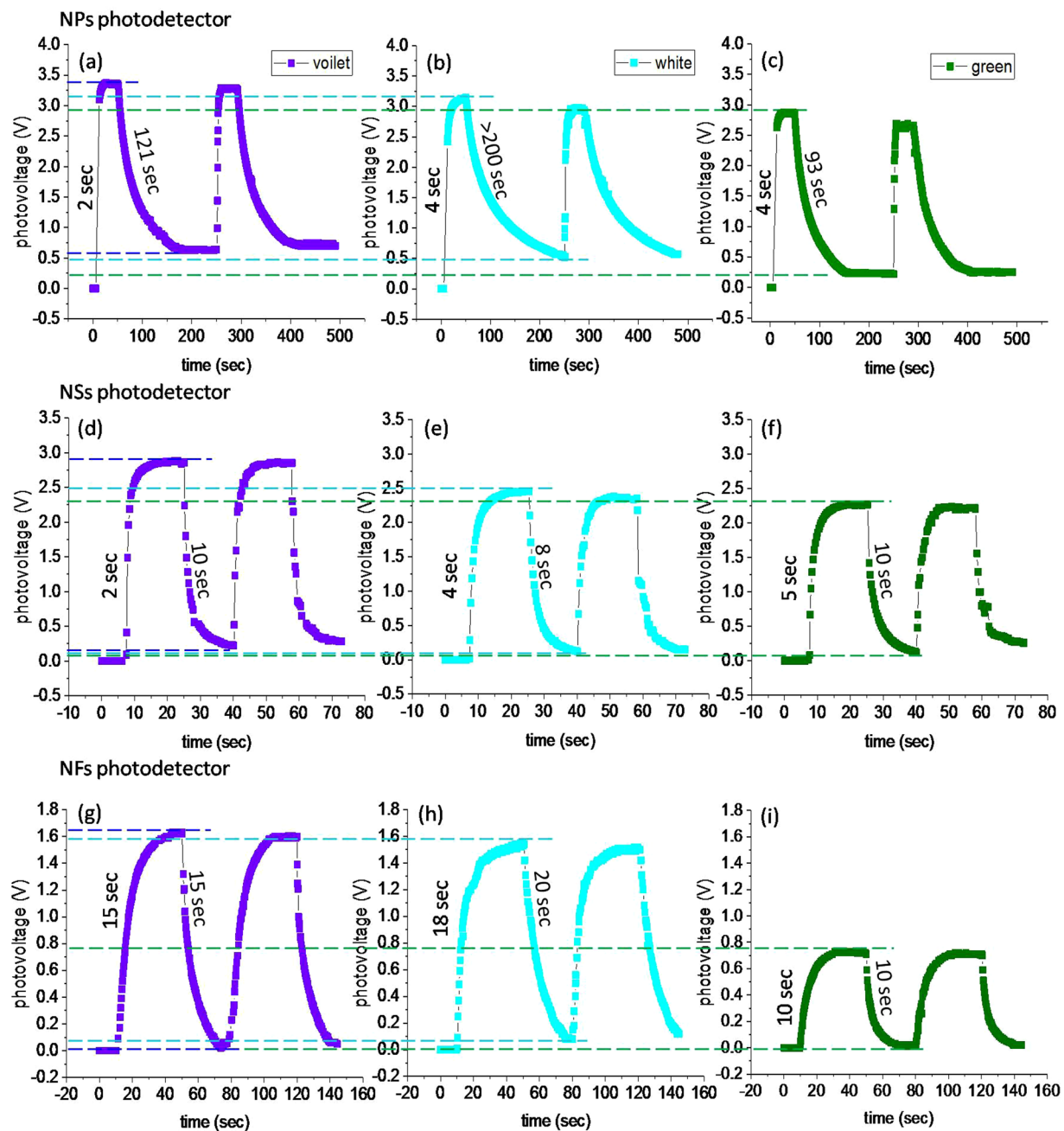


Figure 7. Response of nanostructured thin film photodetectors. (a–c) NPs photodetector, (d–f) NSs photodetector and (g–i) NSs photodetector for UV, white and green illuminations, dashed lines are plotted for clear observation of photo-generated ON state saturation voltages and OFF state dark voltages by different illuminations. In each photodetector, the illumination ON and OFF time is taken as short as required to avoid the heating effects. The applied bias voltage in all three cases is 5 V.

photovoltage might involve both of the desorption mechanisms explained above, according to the illumination conditions.

First of all, we compare photo-response curves corresponding with different illuminations in the NPs photodetector. In the NPs photodetector, the photovoltage initially at ~ 0 V (dark voltage) reaches the saturation voltage 3.35 V within 2 sec upon UV exposure (Fig. 7a), and at 3.12 V and 2.85 V within 4 sec after white and green illuminations (Fig. 7(b,c)), respectively. Here, the illumination ON and OFF time were taken as 50 sec and 200 sec, respectively. While looking at the photo-response curve of NPs photodetector, the rise time corresponding to different illuminations have small variations, whereas their decay time have large difference, and are much longer in comparison with the corresponding rise time as 93 sec for green, 121 sec for UV and greater than 200 sec for white illumination. After turning OFF the illuminations, the photodetector does not achieve its original dark

Structures/ photodetectors	Saturation photovoltage for each illumination (V)	Growth time for each illumination (sec)	Decay time for each illumination (sec)	Surface area (m ² /gm)
NPs	Violet: 3.35	UV ~ 2	Violet~121	20.5
	White: 3.12	White ~ 4	White > 200	
	Green: 2.85	Green ~ 4	Green ~ 93	
NSs	Violet: 2.87	UV ~ 2	Violet ~ 10	11.8
	White: 2.45	White ~ 4	White ~ 8	
	Green: 2.27	Green ~ 5	Green ~ 10	
NFs	Violet: 1.62	UV ~ 15	Violet ~ 15	10.0
	White: 1.53	White ~ 18	White ~ 20	
	Green: 0.72	Green ~ 10	Green ~ 10	

Table 1. Shows comparison of the performance of ZnO nanostructures-based photodetectors under different spectral illumination using 5 V bias voltage.

voltage; instead it remains at a minimum voltage such as; 0.635 V for UV, 0.241 V for green illuminations, whereas for white illumination, it remains unsaturated even after 200 sec as can be seen in Fig. 7a.

The observed difference in the photovoltage rise time with UV, white and green illuminations can be understood by the difference in their illumination energies. UV illumination being an above-band energy, generates carriers both by desorption of adsorbed ions as well as direct excitation of electrons from VB to CB, whereas in case of white and green illuminations the charge carriers are generated only by desorption of the adsorbed ions. Thus for UV illumination, the reduction in depletion region height should be larger, enhancing charge carriers mobility⁵⁹, and hence the fast rise in the photovoltage (Table 1). In the response curves, the ON state saturation voltage is also a reflection of different illumination energies. Based upon the energy levels of the illuminations, different concentrations of charge carriers are generated that give rise to the different values of saturation photovoltages as shown in Fig. 7. The decay of photovoltage after turning OFF the illuminations is due to re-adsorption of O₂/H₂O. The re-adsorption/decay curve shows two regions, fast decaying and slow decay regions. The fast decay is due to an instantaneous electron capture by adsorbed chemisorbed O₂/H₂O on the surface V_{os}, and slower decay is due to the adsorption of O₂/H₂O deep into the surface between nano-crystallites⁶⁰. The latter involves rate-limited diffusion and rearrangement of O₂/H₂O in a closed packed structure for adsorption on the surface that makes the process slower⁶⁰. Just after turning OFF the illuminations, the decrease in photovoltage is faster and almost similar for all three illuminations (UV, white and green), which successively becomes slower on the latter stage and acquires different decreasing rate for each illumination.

A model governing the rate-limited diffusion controlled adsorption process can be represented as⁶⁰,

$$\frac{dNi(t)}{dt} = \frac{Ns - Ni(t)}{\tau} \quad (1)$$

where Ni represents density of charged ions (after capturing electrons), Ns is saturation density of ionized species which prevents further ionization and τ is adsorption rate.

Since for the green illumination, the photo-generated voltage is comparatively lower due to small number of generated charge carriers. As soon as the illumination is turned OFF, the produced smaller number of electrons will be captured by adsorbed O₂/H₂O at a faster rate than that for UV illumination⁵⁹. Nevertheless, by UV illumination, electrons and holes are separated away in space that also increases their recombination life time²⁵. However, the slow decay response in case of white illumination is unclear. When looking to the minimum value of dark voltage in the OFF state, there exists a shift in the dark voltage for each illumination. This may be ascribed to the presence of neutralized oxygen on the surface of ZnO-NPs. The neutralized oxygen residing on the surface would occupy a part of the surface, and thus will not allow the newly coming oxygen for electron capture, and thus resulting in a shift of photovoltage⁶⁰. The observed shift in the dark voltage is in accordance with the energy levels of illuminating radiation. The UV radiation, being a high energy, would neutralize more ions as compared with white and green illuminations, and therefore shows a larger shift in dark voltage (Fig. 8(a-c)).

Now let's compare the photo-response of NPs, NSs and NFs photodetectors with respect to different illuminations (UV, white and green). The NSs and NPs photodetectors show similarity in their photovoltage rise time as 2 sec for UV, 4 sec for white and 5 sec for green; however, their ON state saturation voltages are different as shown in Fig. 8(a-c). NSs photodetector has smaller value of ON state saturation voltage than that of NPs photodetector. The NFs photodetector shows further smaller value of ON state saturation photovoltages as 1.62 V for UV, 1.53 V for white and 0.72 V for green, whereas the photovoltage rise time is longer than that of NPs and NSs. The difference in the saturation photovoltages can be correlated with the different densities of illumination centers/defects in the nanostructures as detected by PL spectra (Fig. 4b). The observed PL intensities in the order, $I_{NF} > I_{NS} > I_{NP}$ are representative of densities of illumination centers/defects in the respective nanostructure. These intensities/densities are adverse to the ON state saturation photovoltages. Indubitably, not all the peaks observed in PL do correspond to the generated photovoltage⁶¹, rather it indicates the possibility that a corresponding photovoltage may be generated upon illumination when any/all of the defect states participate in the photovoltage generation mechanism. The observed adverse effect of defect states over the ON state saturation voltage indicates that there should be capturing/scattering of photo-generated charge carriers during transportation^{62,63}. In other words, the higher is defect density, lower is photo-generated voltage or vice versa. In whole of the process; however, unfortunately we could not obtain variation of defect states within a single morphology (either of NPs, NSs or NFs),

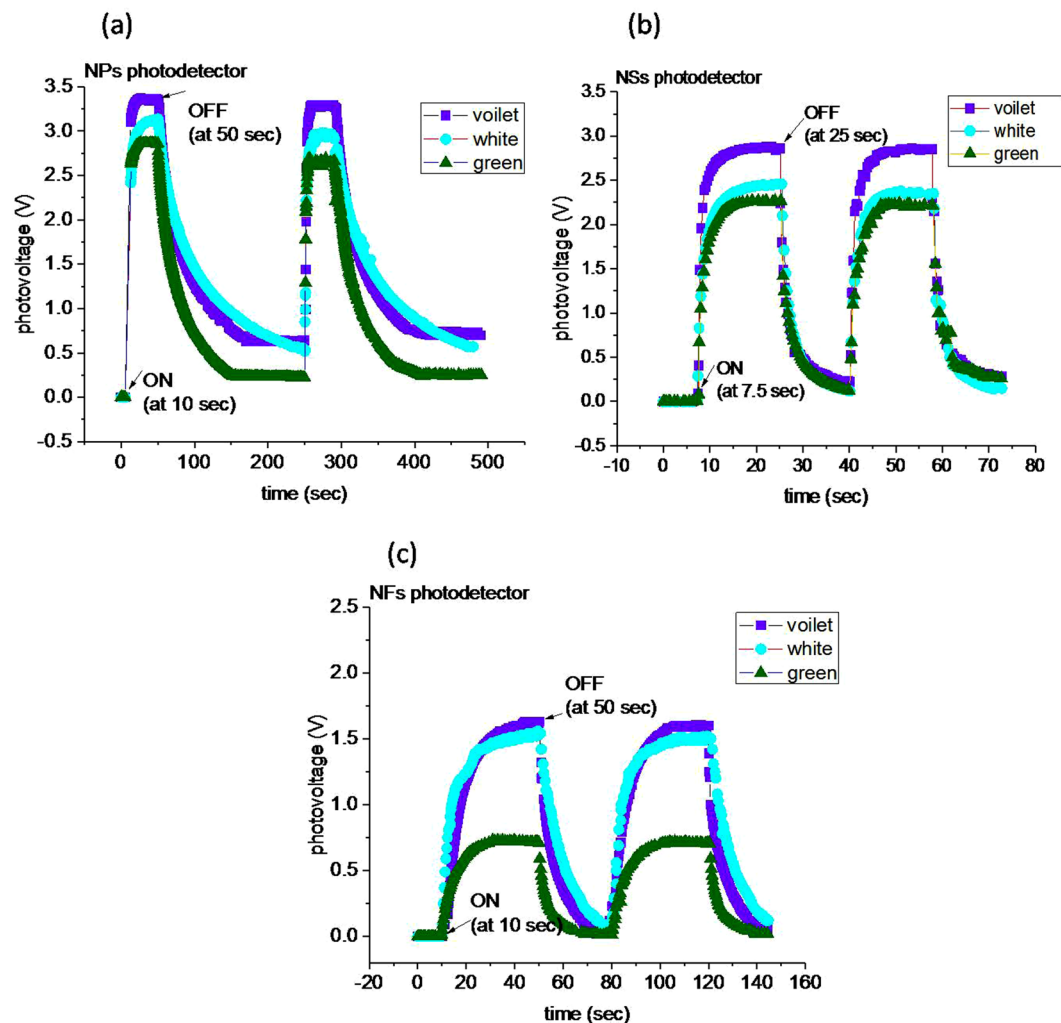


Figure 8. Comparison of photovoltage vs time curves of ZnO photodetectors thin films for UV, white and green illuminations. (a) NPs photodetector response showing high photo-generated voltage, (b) NSs photodetector shows fast response for both rise and fall time, and (c) NFs photodetector is slow for rise time and intermediate for fall time, also in this case the photo-generated voltage is smaller than that of NPs and NSs photodetectors.

as that could have been a further better examination of the effect of defects states variation over the ON state saturation voltage.

For decay/off state recovery time of photovoltages, it is known that the decay time depends mainly upon three factors; the available defect states ($V_{o,s}$ and $Zn_{i,s}$ in our case), surface area and adsorbate (O_2 or H_2O). The XPS results (Figs 5a and 6a) demonstrate that NPs possess smaller content of $V_{o,s}$ (21.8%) and smaller content of $Zn_{i,s}$ (as the main contribution being from hydroxide formation), whereas larger content of adsorbed O_2/H_2O (60.39%) along with a large surface area ($20.5 \text{ m}^2/\text{g}$) in comparison to that of NSs and NFs. The smaller percentage of $V_{o,s}$ and $Zn_{i,s}$ spreading over a larger surface area would initially promote fast chemisorption process of O_2/H_2O after turning OFF the illumination. The fast-initial decay can also be seen in NSs and NFs photodetectors, which latter on becomes slower and different in all the photodetectors. The fast decay, after turning OFF the illumination, indicates that initially the decay is essentially through a chemisorption process that is considered as a fast process. The surface states are then saturated by initial chemisorption process, and O_2/H_2O diffuse through the inter-crystallite deep into the surface, where they can find sites for adsorption; this corresponds to the slower decay⁶⁴ of photovoltages. In NPs, the content of adsorbed O_2/H_2O (60.39%) over the 21.8% of $V_{o,s}$ and a smaller percentage of $Zn_{i,s}$ is larger, this shows that during the slower decay step, a large amount of O_2/H_2O , in comparison to the chemisorption, is physisorbed on the surface of NPs. Therefore, physisorption process is dominating in the case of NPs photodetector. Since the physisorption process involves rearrangement of adsorbate on the surface, so it will result in an increasing decay time of NPs photodetector (Fig. 8a). On the other hand, NSs photodetector, in comparison with NPs photodetector, has higher content of $V_{o,s}$ (47.7%) and $Zn_{i,s}$ (~13.3%), and smaller content of adsorbed (O_2/H_2O ~40.95%) along with smaller surface area ($11.8 \text{ m}^2/\text{g}$) would decay prominently through the chemisorption process due to the abundance of defect states to facilitate chemisorption process. Thus

the larger content of V_o s and Zn_i s than the adsorbed O_2/H_2O leads to dominating fast chemisorption process that results in faster decay of NSs photodetector (Fig. 8b). Next, the NFs photodetector decay time shows totally different behavior, which is adverse to the trend obtained from NPs to NSs photodetectors. In this case, an increased decay time despite of higher content of V_o s (54.5%) and Zn_i s (28.6%), and smaller content of adsorbate (37.6%), surface area ($10\text{ m}^2/\text{g}$) is observed. Here, the excess of defect states (V_o s and Zn_i s) appears to create adverse effect on photoconductivity, similar to the reduction in the ON state saturation photovoltage (peak voltage in the Fig. 8(a,c)), and thus increases photovoltage decay time in NFs photodetector (Fig. 8c). The photovoltage rise time of NFs photodetector (15 sec for UV, 18 sec for white and 10 sec for green illuminations) is also longer than that of both the NSs and NPs photodetectors (Table 1). Here also, the deterministic factor appears to be the different content of V_o s and Zn_i s in nanostructures⁶⁴; as the higher content of V_o s and Zn_i s in nanostructure photodetector results in longer photovoltage rise time. The NPs have lower percentage of V_o s and Zn_i s that would desorb rapidly and thus fast release of charge carriers upon illumination.

From the above discussed observations, the UV and visible photo-response of thin film NSs is found better in comparison with NPs and NFs photodetectors (as compared in Table 1). Another important feature of NSs photodetector is its similarity in the rates of decay curves for both UV and visible illuminations despite of having different saturation voltages. The photo-response of the NSs photodetector in film architecture is even better than that of a sophisticated architectures consisting a single ZnO nanowire that performs only in UV illumination^{24–26,57}.

Conclusion

As a conclusion, a significant visible light photo-detection along with UV photodetection is achieved in pristine ZnO nanostructure based thin films. The defect states V_o s and Zn_i s along with morphology giving rise to the visible-light photo response are generated simply by tuning of solution method. The results of PL, XPS and photovoltages show that although the defect states V_o s and Zn_i s are responsible for exhibiting photo response in the visible-region of spectrum but at the same time their excess reduces the performance of photodetector as observed ~54% of V_o s and ~28% of Zn_i s in NFs photodetector, which is probable by scattering/capturing of charge carriers during their transportation. The ZnO-sheet based photodetector possessing moderate amount of V_o s (~47%) and Zn_i s (~13%) in comparison with ZnO-particle and ZnO-flower photodetectors, shows faster response to photovoltage growth time and decay time under the UV as well as visible-light illumination. The ZnO nanosheet based photodetector can be used as an efficient material for photodetector applications in broad-band spectral applications.

Materials and Methods

Synthesis of ZnO-NPs, ZnO-NSs, ZnO-NFs, and fabrication of photodetectors. The simple chemical route was tuned for the formation of surfactant free ZnO nanostructures. Detailed synthesis experiments listed in the Tables S1 and S2 of the supplementary information were performed with the variation of synthesis time, precursors, solvents and their molar ratio that resulted in the different ZnO morphologies. In the detailed experiments, a new approach was adopted to introduce V_o s and Zn_i s in the lattice of ZnO which enabled the ZnO nanostructures active for the visible light photo-detection. In the approach, an abrupt/fast mixing of precursor solution with alkali solution in the reaction chamber resulted in the formation of V_o s and Zn_i s as identified by XPS and PL studies. From the detailed synthesis experiments, we found that surfactant free ZnO-NSs and ZnO-NFs resulted only for specific conditions of precursor concentration (0.5 M), precursor's molar ratio (1:1) and reaction time (4 hours). For NSs synthesis, 0.5 M solution of zinc chloride ($ZnCl_2$) (purity 99.99%, Sigma-Aldrich, USA) and 0.5 M solution of sodium hydroxide (NaOH) (purity 98%, Merck India Ltd.) were dissolved in ethyl alcohol (C_2H_5OH) separately. Then the prepared precursor solution of $ZnCl_2$ was added abruptly in the solution of NaOH under vigorous stirring conditions at room temperature (~30 °C). After 4 hours of reaction time, the formed precipitate was collected, filtered, and washed with deionized water and C_2H_5OH to remove Cl^- and Na^+ ions, which was finally dried at 60 °C. In the second set of experiments for the synthesis of NFs, all the conditions were kept same as that in case of NSs formation, the only solvent C_2H_5OH was replaced with deionized water. In this case, ZnO-NFs rather than NSs were obtained for the specific condition of precursor concentrations 0.5 M, precursor's molar ratio 1:1 and reaction time 4 hours. The NPs were formed in both of the synthesis cases for precursor concentration 0.1 M, molar ratio of precursors 1:4 (NaOH: $ZnCl_2$) and reaction time 15 min. In order to fabricate the photodetector of as synthesized ZnO nanostructures, they were dispersed in C_2H_5OH solutions and sonicated for 30 minutes. Then these dispersed nanostructures were spray coated on ultrasonically cleaned glass substrates to make uniform films of thickness about 4 μm . For spray coating, N_2 was used as a carrier gas in the nozzle at spray rate of 1 ml/min. After drying, silver (Ag) interdigitated fingers were printed on the surface of films to make electrical contacts. The width of Ag interdigital electrodes was taken about 1 mm with fringe separation of 2 mm as shown in Figure S1.

Characterizations of materials and photodetectors. The morphology of ZnO nanostructures was investigated by field-emission electron microscopy (FE-SEM, Hitachi S-4700, Tokyo, Japan), optical properties (absorbance and bandgap) were investigated by using UV-Vis spectrophotometer (Perkin-Elmer Lambda 750) and structural study was done by X-ray diffractometer (XRD) (Rigaku, radiation $Cu\ K\alpha$, $\lambda = 1.5406\text{ \AA}$). Photoluminescence (PL) (LS-55 Luminescence, Perkin Elmer, Germany), and X-ray photoelectron spectroscopy (XPS) were used to investigate defect states in the nanostructures. To estimate photovoltage in the visible region, a mercury lamp (power 100 watt) was used as a light source, whose intensity on the surface of photodetectors was adjusted 1 mW cm^{-2} . The photovoltage of fabricated photo-detectors was recorded at room temperature using digital multimeter. In the photo-detection experiments, bias voltage of 5 V was applied. UV diode and optical filters with transmittance wavelength in the range 380–420 nm, 490–560 nm and $\lambda > 420\text{ nm}$ (white light) were used to select different spectrum regions, respectively. The 'ON' and 'OFF' state of incident radiations

were controlled by using a camera shutter. Schematic layout of the experimental set-up for the measurement of photo-generated voltage is shown in Figure S1.

References

- Razeghi, M. & Rogalski, A. Semiconductor ultraviolet detectors. *J. Appl. Phys.* **79**, 7433–7473 (1996).
- Zhan, Z. Y., Zheng, L. X., Pan, Y. Z., Sun, G. Z. & Li, L. Self-powered, visible-light photodetector based on thermally reduced graphene oxide-ZnO (rGO-ZnO) hybrid nanostructure. *J. Mater. Chem.* **22**, 2589–2595 (2012).
- Zhai, T. *et al.* Recent developments in one-dimensional inorganic nanostructures for photodetectors. *Adv. Funct. Mater.* **20**, 4233–4248 (2010).
- Kim, Y., Kim, S. J., Cho, S., Hong, B. H. & Jang, D. High-performance ultraviolet photodetectors based on solution-grown ZnS nanobelts sandwiched between graphene layers. *Sci. Rep.* **5**, 12345 (2015).
- Shen, G. & Chen, D. One-dimensional nanostructures for photodetectors. *Recent Pat. Nanotechnol.* **4**, 20–31 (2010).
- Lin, Y., Wang, J., Sun, B., Blakesley, J. C. & Greenham, N. C. Solution-processed ultraviolet photodetectors based on colloidal ZnO nanoparticles. *Nano Lett.* **8**, 1649–1653 (2008).
- Rajendran, S. *et al.* Ce³⁺-ion induced visible-light photocatalytic degradation and electrochemical activity of ZnO/CeO₂ nanocomposite. *Nat. Publ. Gr.* 1–11, <https://doi.org/10.1038/srep31641> (2016).
- Jie, J. S. *et al.* Photoconductive characteristics of single-crystal CdS nanoribbons. *Nano Lett.* **6**, 1887–1892 (2006).
- Soci, C. *et al.* ZnO nanowire UV photodetectors with high internal gain. *Nano Lett.* **7**, 1003–1009 (2007).
- Gupta, K., Lin, J. T., Wang, R. C. & Liu, C. P. Porosity-induced full-range visible-light photodetection via ultrahigh broadband antireflection in ZnO nanowires. *NPG Asia Mater.* **8**, e314 (2016).
- Subramanian, V., Wolf, E. & Kamat, P. V. Semiconductor - Metal Composite Nanostructures. To What Extent Do Metal Nanoparticles Improve the Photocatalytic Activity of TiO₂ Films? *J. Phys. Chem. B* **105**, 11439–11446 (2001).
- Liu, J. *et al.* Self-assembling TiO₂ nanorods on large graphene oxide sheets at a two-phase interface and their anti-recombination in photocatalytic applications. *Adv. Funct. Mater.* **20**, 4175–4181 (2010).
- Hsu, C. Y. *et al.* H Supersensitive, ultrafast, broad-band light-harvesting scheme employing carbon nanotube/TiO₂ core-shell nanowire geometry. *ACS Nano* **6**, 6687–6692 (2012).
- Qiao, H. *et al.* Broadband photodetectors based on graphene-Bi₂Te₃ heterostructure. *ACS Nano* **9**, 1886–1894 (2015).
- Liu, C.-H., Chang, Y.-C., Norris, T. B. & Zhong, Z. Graphene photodetectors with ultra broadband and high responsivity at room temperature. *Nat. Nanotechnol.* **9**, 273–8 (2014).
- Mane, R. S., Lee, W. J., Pathan, H. M. & Han, S.-H. Nanocrystalline TiO₂/ZnO thin films: fabrication and application to dye-sensitized solar cells. *J. Phys. Chem. B* **109**, 24254–24259 (2005).
- Kouklin, N. Cu-doped ZnO nanowires for efficient and multispectral photodetection applications. *Adv. Mater.* **20**, 2190–2194 (2008).
- Rakshit, T., Mondal, S. P., Manna, I. & Ray, S. K. CdS-decorated ZnO nanorod heterostructures for improved hybrid photovoltaic devices. *ACS Appl. Mater. Interfaces.* **4**, 6085–6095 (2012).
- Prabhakar, R. R. *et al.* Efficient multispectral photodetection using Mn doped ZnO nanowires. *J. Mater. Chem.* **22**, 9678 (2012).
- Bora, T., Zoepfl, D. & Dutta, J. Importance of Plasmonic Heating on Visible Light Driven Photocatalysis of Gold Nanoparticle Decorated Zinc Oxide Nanorods. *Sci. Rep.* **6**, 26913 (2016).
- Mandal, S., Sambasivarao, K., Dhar, A. & Ray, S. K. Photoluminescence and electrical transport characteristics of ZnO nanorods grown by vapor-solid technique. *J. Appl. Phys.* **106**, 024103 (2009).
- Roy, N., Chowdhury, A. & Roy, A. Observation of negative differential resistance and electrical bi-stability in chemically synthesized ZnO nanorods. *J. Appl. Phys.* **115**, 223502 (2014).
- Lu, Y. *et al.* Production of visible activity and UV performance enhancement of ZnO photocatalyst via vacuum deoxidation. *Appl. Catal. B Environ.* **138–139**, 26–32 (2013).
- Li, Q. H., Gao, T., Wang, Y. G. & Wang, T. H. Adsorption and desorption of oxygen probed from ZnO nanowire films by photocurrent measurements. *Appl. Phys. Lett.* **86**, 1–3 (2005).
- Bao, J. *et al.* Photoinduced oxygen release and persistent photoconductivity in ZnO nanowires. *Nanoscale Res. Lett.* **6**, 404 (2011).
- Chen, Q. *et al.* Passivation of surface states in the ZnO nanowire with thermally evaporated copper phthalocyanine for hybrid photodetectors. *Nanoscale* **5**, 4162–4165 (2013).
- Zeng, H. *et al.* Blue luminescence of ZnO nanoparticles based on non-equilibrium processes: Defect origins and emission controls. *Adv. Funct. Mater.* **20**, 561–572 (2010).
- Mandal, S., Goswami, M. L. N., Das, K., Dhar, A. & Ray, S. K. Temperature dependent photoluminescence characteristics of nanocrystalline ZnO films grown by sol-gel technique. *Thin Solid Films* **516**, 8702–8706 (2008).
- Ma, J. G. *et al.* Preparation and characterization of ZnO particles embedded in SiO₂ matrix by reactive Magnetron sputtering. *J. Appl. Phys.* **97**, 103509 (2005).
- Lim, K. *et al.* Temperature-driven structural and morphological evolution of zinc oxide nano-coalesced microstructures and its defect-related photoluminescence properties. *Materials (Basel)* **9**, 300 (2016).
- Vlasenko, L. S. & Watkins, G. D. Optical detection of electron paramagnetic resonance in room-temperature electron-irradiated ZnO. *Phys. Rev. B - Condens. Matter Mater. Phys.* **71**, 1–6 (2005).
- Kundu, T., Karak, N., Barik, P. & Saha, S. Optical Properties of ZnO Nanoparticles Prepared by Chemical Method Using Poly (VinylAlcohol) (PVA) as Capping Agent. *Ijsee.Org* 19–24 (2011).
- Epie, E. N. & Chu, W. K. Ionoluminescence study of Zn and O implanted ZnO crystals: An additional perspective. *Appl. Surf. Sci.* **371**, 28–34 (2016).
- Ahn, C. H., Kim, Y. Y., Kim, D. C., Mohanta, S. K. & Cho, H. K. A comparative analysis of deep level emission in ZnO layers deposited by various methods. *J. Appl. Phys.* **105**, 1–6 (2009).
- Vempati, S., Mitra, J. & Dawson, P. One-step synthesis of ZnO nanosheets: a blue-white fluorophore. *Nanoscale Res. Lett.* **7**, 470 (2012).
- Bandopadhyay, K. & Mitra, J. Zn interstitials and O vacancies responsible for n-type ZnO: what do the emission spectra reveal? *RSC Adv.* **5**, 23540–23547 (2015).
- Kayaci, F., Vempati, S., Donmez, I., Biyikli, N. & Uyar, T. Role of zinc interstitials and oxygen vacancies of ZnO in photocatalysis: a bottom-up approach to control defect density. *Nanoscale* **6**, 10224 (2014).
- Vanheusden, K. *et al.* Mechanisms behind green photoluminescence in ZnO phosphor powders. *J. Appl. Phys.* **79**, 7983 (1996).
- Ghosh, A. & Choudhary, R. N. P. Optical emission and absorption spectra of Zn-ZnO core-shell nanostructures. *J. Exp. Nanosci.* **5**, 134–142 (2010).
- Janotti, A. & Van de Walle, C. G. Fundamentals of zinc oxide as a semiconductor. *Rep Prog Phys* **72**, 126501 (2009).
- Kunat, M., Girol, S. G., Burghaus, U. & Wöll, C. The Interaction of Water with the Oxygen-Terminated, Polar Surface of ZnO. *J. Phys. Chem. B* **107**, 14350–14356 (2003).
- Bharti, B., Kumar, S., Lee, H.-N. & Kumar, R. Formation of oxygen vacancies and Ti³⁺ state in TiO₂ thin film and enhanced optical properties by air plasma treatment. *Sci. Rep.* **6**, 32355 (2016).

43. Al-Gaashani, R., Radiman, S., Daud, A. R., Tabet, N. & Al-Douri, Y. XPS and optical studies of different morphologies of ZnO nanostructures prepared by microwave methods. *Ceram. Int.* **39**, 2283–2292 (2013).
44. Islam, M. N., Ghosh, T. B., Chopra, K. L. & Acharya, H. N. XPS and X-ray diffraction studies of aluminum-doped zinc oxide transparent conducting films. *Thin Solid Films* **280**, 20–25 (1996).
45. van Dijken, A., Meulenkamp, E. A., Vanmaekelbergh, D. & Meijerink, A. The kinetics of the radioactive and nonradioactive processes in nanocrystalline ZnO particles upon photoexcitation. *J. Phys. Chem. B* **104**, 1715–1723 (2000).
46. Wang, Y. *et al.* Optical modulation of persistent photoconductivity in ZnO nanowires. *Appl. Phys. Lett.* **98** (2011).
47. Das, D. & Mondal, P. Low temperature grown ZnO:Ga films with predominant c-axis orientation in wurtzite structure demonstrating high conductance, transmittance and photoluminescence. *RSC Adv.* **6**, 6144–6153 (2016).
48. Moulder, J. F. Handbook of X-ray photoelectron spectroscopy: a reference book of standard spectra for identification and interpretation of XPS data. Edited by Jill Chastain, and Roger C. King. Eden Prairie, MN: Physical Electronics, (1995).
49. Hameed, A. S. H. *et al.* Impact of alkaline metal ions Mg²⁺, Ca²⁺, Sr²⁺ and Ba²⁺ on the structural, optical, thermal and antibacterial properties of ZnO nanoparticles prepared by the co-precipitation method. *J. Mater. Chem. B* **1**, 5950–5962 (2013).
50. Hou, L. *et al.* Self-sacrifice template fabrication of hierarchical mesoporous bi-component-active ZnO/ZnFe₂O₄ sub-microcubes as superior anode towards high-performance lithium-ion battery. *Adv. Funct. Mater.* **25**, 238–246 (2015).
51. Mai, N. T., Thuy, T. T., Mott, D. M. & Maenosono, S. Chemical synthesis of blue-emitting metallic zinc nano-hexagons. *Cryst Eng Comm* **15**, 6606 (2013).
52. NIST X-ray photoelectron spectroscopy database. Measurement Services Division of the National Institute of Standards and Technology (NIST) Technology Services (2008).
53. Decoster, D. & Harari, J. In *Optoelectronic Sensors* Wiley-ISTE (2010).
54. Shapira, Y., Cox, S. M. & Lichtman, D. Photodesorption from powdered zinc oxide. *Surface Science*. **50**, 503–514 (1975).
55. Law, J. B. K. & Thong, J. T. L. Simple fabrication of a ZnO nanowire photodetector with a fast photoresponse time. *Appl. Phys. Lett.* **88**, 13–15 (2006).
56. Yan, C., Singh, N. & Lee, P. S. Wide-bandgap Zn₂GeO₄ nanowire networks as efficient ultraviolet photodetectors with fast response and recovery time. *Appl. Phys. Lett.* **96** (2010).
57. Li, Q. H., Wan, Q., Liang, Y. X. & Wang, T. H. Electronic transport through individual ZnO nanowires. *Appl. Phys. Lett.* **84**, 4556–4558 (2004).
58. Afal, A., Coskun, S. & Emrah Unalan, H. All solution processed, nanowire enhanced ultraviolet photodetectors. *Appl. Phys. Lett.* **102** (2013).
59. Li, Y., Della Valle, F., Simonnet, M., Yamada, I. & Delaunay, J. J. Competitive surface effects of oxygen and water on UV photoresponse of ZnO nanowires. *Appl. Phys. Lett.* **94**, 1–3 (2009).
60. Tansley, T. L. & Neely, D. F. Adsorption, desorption and conductivity of sputtered zinc oxide thin films. *Thin Solid Films*. **2**, 95–107 (1984).
61. Keem, K. *et al.* Photocurrent in ZnO nanowires grown from Au electrodes. *Appl. Phys. Lett.* **84**, 4376–4378 (2004).
62. Zheng, Z., Gan, L., Zhang, J., Zhuge, F. & Zhai, T. An Enhanced UV-Vis-NIR and Flexible Photodetector Based on Electrospun ZnO Nanowire Array/PbS Quantum Dots Film Heterostructure. *Adv. Sci.* 1600316, <https://doi.org/10.1002/advs.201600316> (2016).
63. Aga, R. S. Jr *et al.* Enhanced photoresponse in ZnO nanowires decorated with CdTe quantum dot. *Appl. Phys. Lett.* **91**, 1–4 (2007).
64. Ahn, S. E. *et al.* Photoresponse of sol-gel-synthesized ZnO nanorods. *Appl. Phys. Lett.* **84**, 5022–5024 (2004).

Acknowledgements

This work was supported by National Research Foundation of South Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2015R1A2A1A05001826), and research grant for Nanotechnology Lab of Jaypee University of Information Technology, Waknaghat, India.

Author Contributions

R. Khokhra fabricated the samples and did their characterizations. B. Bharti helped in XPS studies of the samples. R. Kumar wrote the manuscript and supervised the work. H.N. Lee co-supervised the work and helped in characterizations.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-017-15125-x>.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

An Information-Theoretic Study for Joint Sparsity Pattern Recovery With Different Sensing Matrices

Sangjun Park, Nam Yul Yu, *Member, IEEE*, and Heung-No Lee, *Senior Member, IEEE*

Abstract—In this paper, we study a support set reconstruction problem for multiple measurement vectors (MMV) with different sensing matrices, where the signals of interest are assumed to be jointly sparse and each signal is sampled by its own sensing matrix in the presence of noise. Using mathematical tools, we develop upper and lower bounds of the failure probability of the support set reconstruction in terms of the sparsity, the ambient dimension, the minimum signal-to-noise ratio, the number of measurement vectors, and the number of measurements. These bounds can be used to provide guidelines for determining the system parameters for various compressed sensing applications with noisy MMV with different sensing matrices. Based on the bounds, we develop necessary and sufficient conditions for reliable support set reconstruction. We interpret these conditions to provide theoretical explanations regarding the benefits of taking more measurement vectors. We then compare our sufficient condition with the existing results for noisy MMV with the same sensing matrix. As a result, we show that noisy MMV with different sensing matrices may require fewer measurements for reliable support set reconstruction, under a sublinear sparsity regime in a low noise-level scenario.

Index Terms—Compressed sensing, support set reconstruction, joint sparsity structure, multiple measurement vectors model.

I. INTRODUCTION

CONVENTIONALLY, signals sensed from sensors such as microphones and imaging devices are sampled following the Shannon and Nyquist sampling theory [1] at a rate higher than twice the maximum frequency for signal reconstruction. As the number of samples decided by this theory is often large, the samples go through a compression stage before being stored. Therefore, taking numerous samples, where most of them will be discarded in this stage, is inefficient. Because compressed sensing (CS) [2]–[7] removes the inefficiency, CS has been applied in various areas such as wireless communications [8]–[11], spectrometers [12], multiple input multiple output (MIMO) radars [13], magnetic resonance imaging [14], and imaging/signal processing [15]–[17].

The CS theory states that signals that are sparsely representable in a certain basis are compressively sampled and reconstructed from what we thought is incomplete

information. Let $\mathbf{x} \in \mathbb{R}^N$ be a K -sparse vector with a support set $\mathcal{I} := \{i | x(i) \neq 0\}$ whose indices indicate the positions of the nonzero coefficients of \mathbf{x} . It is compressively sampled by a model called *single measurement vector (SMV)* as follows:

$$\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{n} \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^M$ is a (noisy) measurement vector, $\mathbf{F} \in \mathbb{R}^{M \times N}$ is a sensing matrix, and $\mathbf{n} \in \mathbb{R}^M$ is a noise vector, whose elements are independent and identically distributed (i.i.d) Gaussian with a zero mean and a σ^2 variance. Once the support set is correctly reconstructed, then (1) can be well-posed, which allows us to obtain an accurate estimate of \mathbf{x} using the least square approach. We thus aim to focus on the support set reconstruction problem.

A. Information-Theoretic Works for CS With SMV

Works [18]–[23] have studied the support set reconstruction problem from an information-theoretic perspective. For reliable support set reconstruction, sufficient and necessary conditions were established in the linear and sublinear sparsity regimes.

For support set reconstruction, Wainwright [18] used the union bound to establish a sufficient condition on the number of measurements M for a maximum likelihood (ML) decoder and used Fano's inequality [24] to obtain a necessary condition on M . This ML decoder was analyzed by Fletcher *et al.* [19] to establish a necessary condition on M . Aeron *et al.* [20] used Fano's inequality to form necessary conditions on both M and σ^2 . Then, they used the union bound to obtain sufficient conditions on both M and σ^2 for their sub-optimal decoder. Akcakaya and Tarokh [21] used the union and the large deviation bounds based on empirical entropies to get sufficient conditions on M for their joint typical decoder. They used the converse of the channel coding theorem to get necessary conditions on M . Scarlett *et al.* [22] extended this decoder [21] with the assumption that the distribution of the support set is provided. For a uniform distribution case, their necessary and sufficient conditions are equivalent to those of [21]. However, they are better for a non-uniform distribution case. Scarlett and Cevher [23] linked the support set reconstruction with the problem of coding over a mixed channel, where information spectrum methods were used to obtain necessary and sufficient conditions on M .

B. Information-Theoretic Works for CS With MMV

CS has many applications in wireless sensor networks (WSNs) [8]–[11] and MIMO radars [13]. In these

Manuscript received April 1, 2016; revised November 14, 2016; accepted January 17, 2017. Date of publication May 15, 2017; date of current version August 16, 2017. This work was supported by the National Research Foundation of Korea through the South Korean Government under Grant NRF-2015R1A2A1A05001826. This paper was presented at the 2012 International Symposium on Information Theory.

The authors are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: sjpark1@gist.ac.kr; nyu@gist.ac.kr; heungno@gist.ac.kr).

Communicated by H. Rauhut, Associate Editor for Signal Processing.

Digital Object Identifier 10.1109/TIT.2017.2704111

applications, the signals of interest $\mathbf{x}^s \in \mathbb{R}^N$, $s = 1, 2, \dots, S$ are often modeled as *jointly K -sparse vectors*, implying that $\mathcal{I} = \mathcal{I}^1 = \mathcal{I}^2 \dots = \mathcal{I}^S$, where \mathcal{I}^s is the support set of \mathbf{x}^s and $|\mathcal{I}| = K$, which is referred to as a *joint sparsity structure*.

There are two models for sampling jointly K -sparse vectors. The first model is called *multiple measurement vectors (MMV) with the same sensing matrix* [25], in which they are sampled by the same sensing matrix. The second model is named as MMV with *different sensing matrices* [8], [9], in which each one is sampled by its own sensing matrix.

The authors of [26]–[28] have conducted information-theoretic research to obtain conditions under which the support set of both the models was reconstructed with a high probability. In noisy MMV with the same sensing matrix, Tang and Nehorai [26] used the hypothesis theory to obtain necessary and sufficient conditions on both the number of measurements M and the number of measurement vectors S , and proved that the success probability of the support set reconstruction increases with S , if $M = \Omega(K \log \frac{N}{K})$. Jin and Rao [27] exploited the communication theory to establish necessary and sufficient conditions on M and demonstrated the benefits of the joint sparsity structure based on their conditions. A detailed comparison between the results of our paper and those of [27] will be presented in Section IV. Finally, Duarte *et al.* [28] studied noiseless MMV with different sensing matrices, and formed necessary and sufficient conditions on M . However, it is difficult to apply the conditions to noisy MMV with different sensing matrices.

Meanwhile, works [8], [29], [30] have presented conditions of practical algorithms for a reliable support set reconstruction. In noiseless MMV with the same sensing matrix, Blanchard and Davies [30] obtained conditions for a reliable reconstruction from rank aware orthogonal matching pursuit (OMP). In noisy MMV with the same sensing matrix, Kim *et al.* [29] created compressive MUSIC, and presented its sufficient condition. In noiseless MMV with different sensing matrices, Baron *et al.* [8] produced trivial pursuit (TP) and distributed compressed sensing-simultaneous OMP (DCS-SOMP). By analyzing TP with the assumption that each sensing matrix contains i.i.d. Gaussian elements and that the nonzero values of each sparse vector are i.i.d. Gaussian variables, they demonstrated that with $M \geq 1$, TP reconstructs the support set as S is sufficiently large. They conjectured that $M \geq K + 1$ suffice for DCS-SOMP to reconstruct the support set as S is sufficiently large, based on its empirical results.

To the best of our knowledge, no information-theoretic study has been published to get necessary and sufficient conditions for reliable support set reconstruction in noisy MMV with different sensing matrices. Besides, these conditions have not been provided from the practical recovery algorithms for CS with noisy MMV with different sensing matrices.

C. Motivations of This Paper

CS with noisy MMV with different sensing matrices has been applied in many applications and the benefits facilitated by the joint sparsity structure have been empirically reported in [10] and [14]. In WSNs, Caione *et al.* [10] used the joint sparsity structure to reduce the number of

transmitted bits per sensor and reported that each sensor can reduce its transmission cost. In magnetic resonance imaging (MRI), Wu *et al.* [14] modeled multiple diffusion tensor images (DTIs) as jointly sparse vectors. They exploited the joint sparsity structure to reduce the number of samples per DTI, while retaining the reconstruction quality. Using the joint sparsity structure, they also empirically reported that the reconstruction quality of each DTI can be improved for a fixed number of samples per DTI.

To theoretically explain the above empirical benefits facilitated by the joint sparsity structure, theoretical tools are required to measure the performance of CS with noisy MMV with different sensing matrices. Such tools can be useful as guidelines for determining the system parameters in various CS applications with noisy MMV with different sensing matrices. For example, if the number of samples per DTI is fixed in the MRI [14], the theoretical tools may enable us to determine the number of DTIs required for achieving a given reconstruction quality. Thus, the first motivation of this paper is to provide the theoretical tools by establishing sufficient and necessary conditions for reliable support set reconstruction.

Next, for noiseless MMV with the same sensing matrix, let $\mathbf{Y}_A = \mathbf{F} \times [\mathbf{x}^1 \ \mathbf{x}^2 \ \dots \ \mathbf{x}^S] \in \mathbb{R}^{M \times S}$. Also, for noiseless MMV with different sensing matrices, let $\mathbf{Y}_B = [\mathbf{F}^1 \mathbf{x}^1 \ \mathbf{F}^2 \mathbf{x}^2 \ \dots \ \mathbf{F}^S \mathbf{x}^S] \in \mathbb{R}^{M \times S}$. Then, all the elements of \mathbf{Y}_B are uncorrelated because all the sensing matrices are independent. In contrast, those of \mathbf{Y}_A are correlated because they are taken from the same sensing matrix. Now, we consider a case where we set $S > K$ and $M > K$. Then, it is clear that $\text{rank}(\mathbf{Y}_B) = \min(S, M)$ with a high probability and $\text{rank}(\mathbf{Y}_A) \leq K$. Therefore, for this case, we conclude that $\text{rank}(\mathbf{Y}_B) > \text{rank}(\mathbf{Y}_A)$. This implies that a more reliable support set reconstruction can be expected in noiseless MMV with different sensing matrices for this case. Thus, the second motivation is to verify this perception in the presence of noise, by comparing our results with the existing ones in noisy MMV with the same sensing matrix [27].

D. Contributions of This Paper

The contributions of this paper are as follows: First, we derive upper and lower bounds of a failure probability of the support set reconstruction from Lemmas 1 and 2, by exploiting Fano's inequality [24] and the Chernoff bound [31]. These bounds are used for measuring the performance of CS with noisy MMV with different sensing matrices.

Next, we develop necessary and sufficient conditions for reliable support set reconstruction. Theorem 1 states that

$$M > K \left(1 + \frac{1}{Sf(\text{SNR}_{\min})} \right)$$

suffices to achieve reliable support set reconstruction in the *linear sparsity* regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$, and it also states that

$$M > K \left(1 + \frac{1}{Sf(\text{SNR}_{\min})} \log \frac{N}{K} \right)$$

suffices to achieve reliable support set reconstruction in the *sublinear sparsity* regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = 0$, where

$f(\text{SNR}_{\min})$ is an increasing function with respect to the minimum signal-to-noise ratio SNR_{\min} defined in (4). Next, for a finite S, N, K , and SNR_{\min} , Theorem 3 states that

$$M < \frac{2K \log \frac{N}{K} - 2 \log 2}{S \log(1 + K \times \text{SNR}_{\min})}$$

is necessary for reliable support set reconstruction. The necessary and sufficient conditions can be useful as guidelines to determine the system parameters of CS applications with noisy MMV with different sensing matrices. Corollaries 1 and 2 indicate that reliable support set reconstruction is possible as sufficiently many measurement vectors S for a fixed M are taken at a low SNR_{\min} . For a fixed N and K , Theorem 2 shows that $M \geq K + 1$ measurements suffice for reconstructing the support set, as S is sufficiently large. Then, for a fixed N, K , and $M = K + 1$, Corollary 3 provides a sufficient condition on S for reliable support set reconstruction. We provide theoretical explanations of the benefits of the joint sparsity structure, which conform with the empirical results of CS applications with noisy MMV with different sensing matrices [10], [14]. Finally, we compare the sufficient condition (11) with the known one (26) for noisy MMV with the same sensing matrix [27]. Therefore, we demonstrate that if $S \geq K$, noisy MMV with different sensing matrices may require fewer measurements M for reliable support set reconstruction than noisy MMV with the same sensing matrix under a low noise-level scenario. It confirms the superiority of MMV with different sensing matrices.

II. NOTATIONS, SYSTEM MODEL & PROBLEM FORMULATION

A. Notations

The following notations will be used in the whole paper.

1. \mathbb{P} , \mathbb{E} and \mathbb{V} denote the probability, expectation and (co)variance, respectively.
2. A small (capital) bold letter \mathbf{f} (\mathbf{F}) is a vector (matrix).
3. A sub-vector (sub-matrix) formed by the elements (columns) of \mathbf{f} (\mathbf{F}) indexed by a set \mathcal{I} is denoted by $\mathbf{f}_{\mathcal{I}}$ ($\mathbf{F}_{\mathcal{I}}$).
4. For a given matrix \mathbf{F} , its inversion, transpose, trace and the i th eigenvalue are denoted by \mathbf{F}^{-1} , \mathbf{F}^T , $\text{tr}[\mathbf{F}]$ and $\lambda_i(\mathbf{F})$, respectively. Also, its orthogonal projection matrix is defined by

$$\mathbf{Q}(\mathbf{F}) := \mathbf{I}_M - \mathbf{F}(\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \quad (2)$$

where $\mathbf{Q}(\mathbf{F})$ maps an arbitrary vector to the space orthogonal onto the space spanned by the columns of \mathbf{F} .

5. For given sets \mathcal{I} and \mathcal{J} , the relative complements of \mathcal{J} in \mathcal{I} is denoted as $\mathcal{J} \setminus \mathcal{I}$. The cardinality of a set \mathcal{I} is denoted by $|\mathcal{I}|$.
6. For a given function $f(x)$, its n th derivation with respect to x is denoted by $f^{(n)}(x)$.
7. The *linear sparsity regime* is defined by $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$.
8. The *sublinear sparsity regime* is defined by $\lim_{N \rightarrow \infty} \frac{K}{N} = 0$.

9. The expression $f(x) = \Omega(g(x))$ denotes $|f(x)| \geq c|g(x)|$ as $x \rightarrow \infty$ for a constant $c > 0$.

B. System Model

Let $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^S$ be jointly K -sparse vectors with a support set \mathcal{I} that belongs to

$$\mathcal{S} := \{\mathcal{H} | \mathcal{H} \subset \{1, 2, \dots, N\}, |\mathcal{H}| = K\}.$$

Thus, the number of nonzero coefficients of each sparse vector is K , the indices of the nonzero coefficients of all the sparse vectors are the same and the indices belong to the support set.

In noisy MMV with different sensing matrices, each sparse vector is sampled by its own sensing matrix, i.e.,

$$\mathbf{y}^s = \mathbf{F}^s \mathbf{x}^s + \mathbf{n}^s \quad s = 1, 2, \dots, S \quad (3)$$

where all the sensing matrices have i.i.d. Gaussian elements with a zero mean and a unit variance, and all the noise vectors have i.i.d. Gaussian elements with a zero mean and a σ^2 variance. We assume that all the noise vectors and all the sensing matrices are mutually independent. Then, we let x_{\min} be the smallest nonzero magnitude of all the sparse vectors and SNR_{\min} be the minimum signal-to-noise ratio given by

$$\text{SNR}_{\min} := x_{\min}^2 / \sigma^2. \quad (4)$$

C. Problem Formulation

We extend Akcakaya and Tarokh [21]'s decoder for noisy MMV with different sensing matrices. It takes all the measurement vectors as its input and yields a support set decision as its output

$$d : \{\forall_s (\mathbf{y}^s, \mathbf{F}^s)\} \mapsto \hat{\mathcal{I}} \in \mathcal{S}, \quad s = 1, 2, \dots, S.$$

Its decision rules are given in Definition 1.

Definition 1: All the measurement vectors $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^S\}$ and a set $\mathcal{J} \in \mathcal{S}$ are δ jointly typical if the rank of $\mathbf{F}_{\mathcal{J}}^s$, $s = 1, \dots, S$, is K and

$$\left| \left(\sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{J}}^s) \mathbf{y}^s\|_2^2 \right) - S(M - K)\sigma^2 \right| < SM\delta. \quad (5)$$

As each sensing matrix contains i.i.d. Gaussian elements, the rank of each $\mathbf{F}_{\mathcal{J}}^s$, $s = 1, \dots, S$, is K with a high probability. The decision rule is to find sets that satisfy (5) for all the given measurement vectors and $\delta > 0$. In the entire paper, the support set is denoted by \mathcal{I} and any incorrect support set is denoted by \mathcal{J} , where their cardinalities are K , i.e., $|\mathcal{I}| = |\mathcal{J}| = K$.

We define the failure events, wherein the joint typical decoder fails to reconstruct the correct support set. First,

$$\mathcal{E}_{\mathcal{I}}^c := \left\{ \left| \left(\sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s) \mathbf{y}^s\|_2^2 \right) - S(M - K)\sigma^2 \right| \geq SM\delta \right\} \quad (6)$$

implies that the correct support set is not δ jointly typical with all the measurement vectors. Next, for any $\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}$,

$$\mathcal{E}_{\mathcal{J}} := \left\{ \left| \left(\sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{J}}^s) \mathbf{y}^s\|_2^2 \right) - S(M - K)\sigma^2 \right| < SM\delta \right\} \quad (7)$$

implies that an incorrect support set is δ jointly typical with all the measurement vectors. Based on these failure events, we define a failure probability and give its upper bound as follows:

$$\begin{aligned} p_{err} &:= \mathbb{P} \left\{ \hat{\mathcal{I}} \neq \mathcal{I} \mid \mathbf{x}^1, \dots, \mathbf{x}^S \right\} \\ &= \mathbb{P} \left\{ \mathcal{E}_{\mathcal{I}}^c \cup \bigcup_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathcal{E}_{\mathcal{J}} \right\} \\ &\leq \mathbb{P} \left\{ \mathcal{E}_{\mathcal{I}}^c \right\} + \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P} \left\{ \mathcal{E}_{\mathcal{J}} \right\} \end{aligned} \quad (8)$$

where $\mathbb{P} \left\{ \mathcal{E}_{\mathcal{I}}^c \right\}$ is taken with respect to all the noise vectors and $\mathbb{P} \left\{ \mathcal{E}_{\mathcal{J}} \right\}$ is taken with respect to all the noise vectors and all the sensing matrices. We establish Lemmas 1 and 2 given in Appendix A to give upper bounds of the probabilities of the failure events. Combining these lemmas with (8) yields

$$\begin{aligned} p_{err} &\leq \mathbb{P} \left\{ \mathcal{E}_{\mathcal{I}}^c \right\} + \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P} \left\{ \mathcal{E}_{\mathcal{J}} \right\} \\ &\leq 2p(d_1) + \binom{N}{K} p(d_{2,\alpha^*} - 1) \end{aligned}$$

where p is defined in (31), $d_1 = \frac{M\delta}{(M-K)\sigma^2}$, $d_{2,\alpha^*} = \frac{(M-K)\sigma^2 + M\delta}{(M-K)\alpha^*}$, and $\alpha^* = \sigma^2 + x_{\min}^2$.

It is of interest to examine why $\mathbb{P} \left\{ \mathcal{E}_{\mathcal{I}}^c \right\}$ depends only on the noise vectors. As shown in Lemma 3, the random variable to define the event $\mathcal{E}_{\mathcal{I}}^c$ in (6) is $\sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s) \mathbf{y}^s\|_2^2 / \sigma^2$, where the measurement vector in (3) consists of the two parts: the noise part \mathbf{n}^s and the signal part $\mathbf{F}_{\mathcal{I}}^s \mathbf{x}_{\mathcal{I}}^s$. The signal part belongs to the space spanned by the columns of $\mathbf{F}_{\mathcal{I}}^s$. Then, as specified in (2), the orthogonal projection matrix $\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s)$ maps the measurement vector to the space orthogonal onto the space spanned by the columns of $\mathbf{F}_{\mathcal{I}}^s$. Thus, the random variable is a function of the noise vectors only.

III. MAIN RESULTS

As the main contribution of this paper, this section presents sufficient and necessary conditions on M for reliable support set reconstruction, i.e., p_{err} converges to zero, in noisy MMV with different sensing matrices. We then interpret the conditions to demonstrate the benefits facilitated by the joint sparsity structure.

A. Sufficient Conditions on M

In [18] and [21], the authors have shown that fewer measurements M for a reliable support set reconstruction are required for noisy SMV in the linear sparsity regime, compared to the sublinear sparsity regime. Based on the results of [18] and [21], we are motivated to examine if the same result can be observed in noisy MMV with different sensing matrices.

Theorem 1: For any $\rho > 1$, we let $\delta = \rho^{-1} (1 - K/M) x_{\min}^2$. If the number of measurements satisfies

$$M > K + v_1 \frac{K}{S} \quad (9)$$

then the failure probability p_{err} defined in (8) converges to zero in the linear sparsity regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$, where

$$v_1 = - \frac{2(1 - \log \beta)}{\log \left(1 - \frac{1 - \rho^{-1}}{1 + \text{SNR}_{\min}^{-1}} \right) + \frac{1 - \rho^{-1}}{1 + \text{SNR}_{\min}^{-1}}} > 0. \quad (10)$$

Also, under the same conditions on ρ and δ , if the number of measurements satisfies

$$M > K + v_2 \frac{K}{S} \log \frac{N}{K} \quad (11)$$

then the failure probability p_{err} defined in (8) converges to zero in the sublinear sparsity regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = 0$, where

$$v_2 = - \frac{2}{\log \left(1 - \frac{1 - \rho^{-1}}{1 + \text{SNR}_{\min}^{-1}} \right) + \frac{1 - \rho^{-1}}{1 + \text{SNR}_{\min}^{-1}}} > 0. \quad (12)$$

Proof: The proof is given in Appendix C.

In terms of N , K , and S , the asymptotic order of the sufficient condition on M for the linear sparsity regime is $\Omega \left(K + \frac{K}{S} \right)$, whereas the order for the sublinear sparsity regime is $\Omega \left(\frac{K}{S} \log \frac{N}{K} \right)$. It confirms that fewer measurements are required in the linear sparsity regime, compared to the sublinear sparsity regime. Next, from the sufficient conditions, we observe an inverse relationship between M and S , owing to the joint sparsity structure. This relationship implies that taking more measurement vectors S reduces the number of required measurements M for reliable support set reconstruction. Then, the relationship can be used for explaining the empirical results of Caione *et al.* [10] and Wu *et al.* [14]. In [10], the authors have reported that the number of transmitted bits per sensor could be inversely reduced by the number of sensors, which implies that the transmission cost of each sensor could be saved. The result can be confirmed by our inverse relationship by considering S and M as the number of sensors and the number of transmitted bits per sensor, respectively. In [14], S and M are considered as the number of DTIs and the number of samples of each DTI, respectively. Again, it has been observed from [14] that the joint sparsity structure enabled the number of samples of each DTI to be inversely reduced by the number of DTIs, reducing the acquisition time for each DTI. These results can be confirmed by our inverse relationship.

Theorem 2: For any $\rho > 1$, we let $\delta = \rho^{-1} (1 - K/M) x_{\min}^2$, N and K be fixed. If the number of measurements satisfies $M \geq K + 1$, the failure probability p_{err} defined in (8) converges to zero as the number of measurement vectors is increased to the infinity.

Proof: The proof is given in Appendix C.

Theorem 2 suggests that with $M \geq K + 1$, reliable support set reconstruction for noisy MMV with different sensing matrices is possible when a large number of measurement vectors is available. The sufficient conditions in Theorem 1, i.e., (9) and (11) have SNR_{\min} values as shown in (10) and (12). They disappear in the sufficient condition of Theorem 2, i.e., $M \geq K + 1$. The support set reconstruction problem becomes

robust against noise when the number of measurement vectors is large.

B. Discussions on the Sufficient Conditions

We now examine the effect of SNR_{\min} on the sufficient conditions of Theorem 1. The aim is to determine the relationship among S , M and SNR_{\min} for reliable support set reconstruction.

Corollary 1: For any $\rho > 1$, we let $\delta = \rho^{-1}(1 - K/M)x_{\min}^2$. The sufficient conditions of Theorem 1 are rewritten as

$$M > K + \left(\frac{\sqrt[3]{S} + (\sqrt{S} \times \text{SNR}_{\min})^{-1}}{1 - \rho^{-1}} \right)^2 4K \log \frac{N}{K} \quad (13)$$

in the sublinear sparsity regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = 0$, and

$$M > K + \left(\frac{\sqrt[3]{S} + (\sqrt{S} \times \text{SNR}_{\min})^{-1}}{1 - \rho^{-1}} \right)^2 4K(1 - \log \beta) \quad (14)$$

in the linear sparsity regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$.

Proof: The proof is given in Appendix D.

Corollary 1 suggests that for a fixed M , reliable support set reconstruction is possible as the number of measurement vectors S is increased to infinity, although SNR_{\min} is low. Namely, we observe a noise reduction effect, which shows that using the joint sparsity structure leads to an increase in SNR_{\min} or a decrease in σ^2 by the square root of S . This effect can explain the improvement in the reconstruction quality of the DTIs, as empirically reported in [14].

We then improve our noise reduction effect by considering that SNR_{\min} is larger than a certain value.

Corollary 2: For any $\rho > 3$, we let $\delta = \rho^{-1}(1 - K/M)x_{\min}^2$ and $\alpha = 2/3$. If

$$\text{SNR}_{\min} \geq \frac{\alpha}{1 - \rho^{-1} - \alpha} = \frac{2\rho}{\rho - 3}, \quad (15)$$

the sufficient conditions of Theorem 1 are rewritten as

$$M > K + \frac{S^{-1} + (S \times \text{SNR}_{\min})^{-1}}{1 - \rho^{-1}} 4K \log \frac{N}{K} \quad (16)$$

in the sublinear sparsity regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = 0$, and

$$M > K + \frac{S^{-1} + (S \times \text{SNR}_{\min})^{-1}}{1 - \rho^{-1}} 4K(1 - \log \beta) \quad (17)$$

in the linear sparsity regime, i.e., $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$.

Proof: The proof is given in Appendix D.

First of all, Corollary 2 requires $\rho > 3$ to ensure that the lower bound in (15) is positive. A simple computation

shows that Corollary 2 requires fewer measurements in both the regimes compared to Corollary 1 because

$$\begin{aligned} \left(\frac{\sqrt[3]{S} + (\sqrt{S} \times \text{SNR}_{\min})^{-1}}{1 - \rho^{-1}} \right)^2 &= S^{-1} \left(\frac{1 + \text{SNR}_{\min}^{-1}}{1 - \rho^{-1}} \right)^2 \\ &\geq S^{-1} \left(\frac{1 + \text{SNR}_{\min}^{-1}}{1 - \rho^{-1}} \right) \\ &= \frac{S^{-1} + (S \times \text{SNR}_{\min})^{-1}}{1 - \rho^{-1}} \end{aligned}$$

where the second inequality is owing to $\frac{1 + \text{SNR}_{\min}^{-1}}{1 - \rho^{-1}} = \frac{1}{t} > 1$ for any $\rho > 3$ and t defined in (61). Besides, Corollary 2 improves the noise reduction effect observed in Corollary 1 by showing that SNR_{\min} is increased by S for the region of SNR_{\min} in (15).

Theorem 2 suggests, it is to be noted, that $M = K + 1$ is sufficient for reliable support set reconstruction if S is sufficiently large with a fixed N and K . Then, it would be interesting to determine how large S should be required for achieving the minimum number of measurements at each sensor, i.e., $M = K + 1$. In wireless sensor networks [34], energy sources used in sensors are very limited due to limitation of sensor sizes. Thus, minimizing the energy used for transmission of data at each sensor which often leads to extending the lifetime of the sensor battery is a value of importance. This point is noted in Caione *et al.* [10] as an advantage of using distributed compressed sensing on joint sparse model-2 signal ensembles (see Section V there). Corollary 3 which aims to provide a sufficient condition on S for achieving $M = K + 1$ thus is motivated.

Corollary 3: Let N and K be fixed and finite. For any $\rho > 1$, we let $\delta = \rho^{-1}(K + 1)^{-1}x_{\min}^2$ and $M = K + 1$. If the number of measurement vectors satisfies

$$S > \underbrace{\left(\log \left(\binom{N}{K} + 2 \right) - \log \varepsilon \right)}_{:=S^*} \times \max \left[\left| \frac{1}{\log \mu_{\mathcal{I}}} \right|, \left| \frac{1}{\log \mu_{\mathcal{J}}} \right| \right] \quad (18)$$

reliable support set reconstruction is possible, i.e., $p_{err} < \varepsilon$ for sufficiently small $\varepsilon \in (0, 1)$, where $\log \mu_{\mathcal{I}}$ and $\log \mu_{\mathcal{J}}$ are defined in (63) and (65), respectively. The sufficient condition on S is decreasing with respect to SNR_{\min} .

Proof: The proof is given in Appendix D.

To the best of our knowledge, the sufficient conditions on S for a reliable support set reconstruction have not yet been developed. A similar result has been reported by Tang and Nehorai [26], in which they reported that $M = \Omega(K \log \frac{N}{K})$ and $S = \frac{\log N}{\log \log N}$ suffice for a reliable support set reconstruction in noisy MMV with the same sensing matrix, as N is sufficiently large.

It is of interest to examine whether the sufficient condition S^* in (18) is good. For this, we implement the joint typical decoder in (5) and conduct experiments for different values of SNR_{\min} and K , for a fixed $N = 50$. We count the number of failure occurrences, wherein the joint typical

decoder fails to reconstruct the support set. We obtain the smallest S^{emp} such that the ratio of the failure occurrences is smaller than $\varepsilon = 0.01$. By comparing S^{emp} with S^* in (18), we see that S^* approaches S^{emp} , as SNR_{\min} is sufficiently large. For example, we see that $S^{emp} = 8$ and $S^* = 12$ at $\text{SNR}_{\min} = 20$ [dB], $K = 2$, and $S^{emp} = 5$ and $S^* = 6$ at $\text{SNR}_{\min} = 30$ [dB], $K = 2$. A similar trend is observed with a bigger K , i.e., at $K = 5$. For example, we see that $S^{emp} = 12$ and $S^* = 19$ at $\text{SNR}_{\min} = 20$ [dB], and $S^{emp} = 7$ and $S^* = 10$ at $\text{SNR}_{\min} = 30$ [dB].

Fletcher *et al.* [19] have reported that the ML decoder requires $M = K + 1$ measurements for a reliable support set reconstruction in noisy SMV, when the signal-to-noise ratio is sufficiently large. This result can be observed from Corollary 3. Specifically, we assume that SNR_{\min} is sufficiently large for a fixed N and K . Then, from (63) and (65), it is easy to see that

$$\begin{aligned} \lim_{\text{SNR}_{\min} \rightarrow \infty} \log \mu_{\mathcal{I}} &= -\infty, \\ \lim_{\text{SNR}_{\min} \rightarrow \infty} \log \mu_{\mathcal{J}} &= 2^{-1} (1 - \rho^{-1} - \log \rho). \end{aligned}$$

Hence, (18) is simplified to

$$S > \left(\log \left(\binom{N}{K} + 2 \right) - \log \varepsilon \right) \times \left| 2 \left(1 - \rho^{-1} - \log \rho \right)^{-1} \right|. \quad (19)$$

Note that N , K , and ε are fixed. Thus, for a large ρ , we have

$$\left| 1 - \rho^{-1} - \log \rho \right| \gg 2 \left(\log \left(\binom{N}{K} + 2 \right) - \log \varepsilon \right), \quad (20)$$

which leads to $S \geq 1$. This result suggests that the joint typical decoder requires $M = K + 1$ measurements for reliable support set reconstruction in noisy SMV, whenever SNR_{\min} is sufficiently large and ρ satisfies (20).

C. Necessary Condition on M

We specify a necessary condition that must be satisfied by a decoder for reliable support set reconstruction in noisy MMV with different sensing matrices. Unlike the sufficient conditions of Theorem 1, the necessary condition is presented for a finite N and K .

We begin by transforming (3) into

$$\underbrace{\begin{bmatrix} \mathbf{y}^1 \\ \vdots \\ \mathbf{y}^S \end{bmatrix}}_{=: \mathbf{y} \in \mathbb{R}^{SM}} = \underbrace{\begin{bmatrix} \mathbf{F}^1 & & \\ & \ddots & \\ & & \mathbf{F}^S \end{bmatrix}}_{=: \tilde{\mathbf{F}} \in \mathbb{R}^{SM \times SN}} \underbrace{\begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^S \end{bmatrix}}_{=: \mathbf{x} \in \mathbb{R}^{SN}} + \underbrace{\begin{bmatrix} \mathbf{n}^1 \\ \vdots \\ \mathbf{n}^S \end{bmatrix}}_{=: \mathbf{n} \in \mathbb{R}^{SM}} \quad (21)$$

where \mathbf{x} is an SK -sparse vector belonging to an infinite set

$$\mathcal{X}_{x_{\min}} := \left\{ \mathbf{x} \in \mathbb{R}^{SN} \mid |x(i)| \geq x_{\min}, \forall i \in \mathcal{I}, |\mathcal{I}| = SK \right\}$$

where $x(i)$ is the i th element of \mathbf{x} and \mathcal{I} is the support set of \mathbf{x} . Owing to the joint sparsity structure, the number of possible support sets is $\binom{N}{K}$. Then, we define a failure probability as:

$$p_{err} := \mathbb{E}_{\tilde{\mathbf{F}}} \sup_{\mathbf{x} \in \mathcal{X}_{x_{\min}}} \mathbb{P} \left\{ \hat{\mathcal{I}} \neq \mathcal{I} \mid \mathbf{x}, \tilde{\mathbf{F}} \right\} \quad (22)$$

where $\hat{\mathcal{I}}$ is an estimate of the support set based on \mathbf{y} and $\tilde{\mathbf{F}}$ in (21). Then, Lemma III-3 of [20] yields

$$\sup_{\mathbf{x} \in \mathcal{X}_{x_{\min}}} \mathbb{P} \left\{ \hat{\mathcal{I}} \neq \mathcal{I} \mid \mathbf{x}, \tilde{\mathbf{F}} \right\} \geq \min_{\hat{\mathbf{x}} \in \mathcal{X}_{x_{\min}}} \max_{\mathbf{x} \in \mathcal{X}_{x_{\min}}} \mathbb{P} \left\{ \hat{\mathbf{x}} \neq \mathbf{x} \mid \mathbf{x}, \tilde{\mathbf{F}} \right\} \quad (23)$$

where $\hat{\mathbf{x}}$ is an estimate for \mathbf{x} based on \mathbf{y} and $\tilde{\mathbf{F}}$ in (21) and

$$\mathcal{X}_{x_{\min}} := \left\{ \mathbf{x} \in \mathbb{R}^{SN} \mid x(i) = x_{\min}, \forall i \in \mathcal{I}, |\mathcal{I}| = SK \right\}$$

which is a finite set. Assume that \mathbf{x} is uniformly distributed over this finite set. Applying Fano's inequality [24] to (23) yields

$$\begin{aligned} \max_{\mathbf{x} \in \mathcal{X}_{x_{\min}}} \mathbb{P} \left\{ \hat{\mathbf{x}} \neq \mathbf{x} \mid \mathbf{x}, \tilde{\mathbf{F}} \right\} &\geq \mathbb{P} \left\{ \hat{\mathbf{x}} \neq \mathbf{x} \mid \tilde{\mathbf{F}} \right\} \\ &\geq 1 - \frac{\mathbb{I}(\mathbf{x}; \mathbf{y} \mid \tilde{\mathbf{F}}) + \log 2}{\log(|\mathcal{X}_{x_{\min}}| - 1)} \end{aligned} \quad (24)$$

where \mathbf{x} and $\hat{\mathbf{x}}$ belong to the finite set $\mathcal{X}_{x_{\min}}$ and $\mathbb{I}(\mathbf{x}; \mathbf{y})$ is the mutual information between \mathbf{x} and \mathbf{y} . We get a necessary condition on M to ensure that the lower bound in (24) is bounded away from zero, as follows:

Theorem 3: Let N and K are fixed and finite. In (21), if the number of measurements satisfies

$$M < \frac{2K \log \frac{N}{K} - 2 \log 2}{S \log(1 + K \times \text{SNR}_{\min})} \quad (25)$$

then the failure probability p_{err} defined in (22) is bounded away from zero.

Proof: The proof is given in Appendix C.

IV. RELATIONS TO THE EXISTING INFORMATION-THEORETIC RESULTS

A. Relations to Noisy MMV With the Same Sensing Matrix [27]

Jin and Rao [27] have exploited the Chernoff bound to obtain a tight sufficient condition on M for a reliable support set reconstruction for noisy MMV with the same sensing matrix in the sublinear sparsity regime. Owing to the complicated form of their sufficient condition, they could not clearly show the benefits facilitated by the joint sparsity structure. Thus, they simplified their condition under scenarios such as: *i)* a low noise-level scenario and *ii)* a scenario with S identical sparse vectors. In Table I, we summarize our sufficient conditions on M , and compare them to that of [27] under the low noise-level scenario in the sublinear sparsity regime.

First, in a low noise-level scenario, as shown in Table I, the sufficient condition [27] for noisy MMV with the same sensing matrix is

$$M = \Omega \left(\frac{K \log N}{\min(K, S)} \right). \quad (26)$$

If $S < K$, the sufficient conditions (11) and (26) have the same order, implying that there is no significant performance gap in the support set reconstruction between the models. However, if $S > K$, (26) is $M = \Omega(\log N)$, whereas (11) is

TABLE I
SUFFICIENT CONDITIONS ON M FOR SUPPORT SET RECONSTRUCTION

	This paper	Yuzhe and Rao [27]
Linear sparsity regime $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$	$M = \Omega(K + \frac{K}{S})$	Not presented
Sublinear sparsity regime $\lim_{N \rightarrow \infty} \frac{K}{N} = 0$	$M = \Omega(\frac{K}{S} \log \frac{N}{K})$	$M = \Omega(\frac{K \log N}{\min(K, S)})$
N and K are finite ($\text{SNR}_{\min} \rightarrow \infty$ or $S \rightarrow \infty$)	$M \geq K + 1$	Not presented

$M = \Omega(\frac{K}{S} \log N)$. It implies that noisy MMV with different sensing matrices is superior to noisy MMV with the same sensing matrix or $S > K$, with respect to M for reliable support set reconstruction. The result of this comparison supports the perception presented in Section I-C, wherein a more reliable support set reconstruction could be expected in a noiseless MMV with different sensing matrices owing to the linear independency of the measurement vectors. Moreover, it validates the perception, even in the presence of noise.

Second, we consider a scenario with S identical sparse vectors. Then, the sufficient condition of [27] is

$$M = \Omega\left(\frac{K \log N}{\log(1 + S \|\mathbf{x}\|_2^2 / \sigma^2)}\right). \quad (27)$$

From (27), we observe that σ^2 is reduced by a factor of S . However, the noise reduction effect for noisy MMV with the same sensing matrix requires a restriction, where all the sparse vectors should be identical, which can be hardly achieved in practice. In contrast, the noise reduction effect for noisy MMV with different sensing matrices does not require this restriction, as shown in Corollaries 1 and 2.

B. Relations to Noisy SMV [21]

Akçakaya and Tarokh [21] have used the joint typical decoder to establish the sufficient conditions on M for a reliable support set reconstruction in noisy SMV. They exploited the exponential inequalities [32] to obtain the upper bounds on the sum of the weighted chi-square random variables. In this subsection, we demonstrate that the approaches developed in this paper are superior to the use of the exponential inequalities. Thus, we use the exponential inequalities to generalize their bounds for noisy MMV with different sensing matrices. We give Propositions 1 and 2 to prove that the generalized bounds are worse than the bounds of Lemmas 1 and 2.

Proposition 1: For any positive δ , we have

$$\mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq 2p(d_1) \leq 2p_{1,\text{exp}}$$

where both $p(d_1)$ and d_1 are given in Lemma 1, and

$$p_{1,\text{exp}} := \exp\left(-\frac{S\delta^2}{4\sigma^4} \frac{M^2}{M - K + 2\delta M / \sigma^2}\right). \quad (28)$$

Proof: The proof is given in Appendix E.

Proposition 2: For any $\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}$ and any $\delta > 0$ such that

$$0 < \delta < (1 - K/M) x_{\min, \mathcal{J}}^2, \quad (29)$$

we have

$$\mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} \leq p(d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1) \leq p_{2, \mathcal{J}, \text{exp}}$$

where both $p(d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1)$ and $d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})}$ are given in Lemma 2 and

$$p_{2, \mathcal{J}, \text{exp}} := \exp\left(-\frac{S^2(M - K)}{4 \sum_{s=1}^S \alpha_{\mathcal{J}, s}^2} \left(x_{\min, \mathcal{J}}^2 - \frac{M\delta}{M - K}\right)^2\right) \quad (30)$$

and $\alpha_{\mathcal{J}, s}$ is defined in (39) and $x_{\min, \mathcal{J}}^2$ is defined in (43).

Proof: The proof is given in Appendix E.

If $S = 1$, we can see that $p_{1,\text{exp}}$ and $p_{2, \mathcal{J}, \text{exp}}$ are equivalent to the bounds of Akçakaya and Tarokh [21]. Propositions 1 and 2 state that the bounds on the failure probability of Lemmas 1 and 2 are tighter than the bounds of [21] for noisy SMV.

V. CONCLUSIONS

We have studied a support set reconstruction problem for CS with noisy MMV with different sensing matrices. The union and Chernoff bounds have been used to obtain the upper bound of the failure probability of the support set reconstruction, and Fano's inequality has been used to obtain the lower bound of this failure probability. As we have obtained the upper bound by analyzing an exhaustive search decoder, the bound is used to measure the performance of CS with noisy MMV with different sensing matrices. We have then developed the necessary and sufficient conditions in terms of the sparsity K , the ambient dimension N , the number of measurements M , the number of measurement vectors S , and the minimum signal-to-noise ratio SNR_{\min} . They can be useful as guidelines to determining the system parameters in various CS applications with noisy MMV with different sensing matrices.

The conditions are interpreted to provide theoretical explanations for the benefits facilitated by the joint sparsity structure in noisy MMV with different sensing matrices:

- i. From the sufficient conditions of Theorem 1, we have observed an inverse relationship between M and S . Due to the inverse relation, we take fewer measurements M per each measurement vector for reliable support set reconstruction by taking more measurement vectors S .
- ii. From the sufficient conditions of Corollaries 1 and 2, we have observed a noise reduction effect, which shows that the usage of the joint sparsity structure results in an increase in SNR_{\min} or a decrease in σ^2 by a factor of S . Therefore, the support set reconstruction can be robust against noise as the number of measurement vectors is increased to infinity.
- iii. From Theorem 2, we have shown that $M = K + 1$ is achieved for a fixed N and K , as S is sufficiently large. From Corollary 3, we have provided the sufficient condition on S to reconstruct the support set for a fixed N , K , and $M = K + 1$.

The theoretical explanations confirm the benefits of the joint sparsity structure, as empirically shown in CS applications with noisy MMV with different sensing matrices [10], [14].

We have compared our sufficient conditions for noisy MMV with different sensing matrices with the other existing results [27] for noisy MMV with the same sensing matrix. For a low-level noise scenario with $S \geq K$, we have shown that the number of measurements for reliable support set reconstruction for noisy MMV with different sensing matrices is lesser than that for noisy MMV with the same sensing matrix. Also, [27] has shown the noise reduction effect. This was achieved under a rather restricted condition though, i.e., all sparse vectors are the same. While such a restricted condition is not required in the noisy MMV problem with *different* sensing matrices studied in this paper, the noise reduction effect has also been observed, which is a significant improvement.

APPENDIX A LEMMAS 1 AND 2

This section presents Lemmas 1 and 2, which give upper bounds of the probabilities of the failure events defined in (6) and (7), respectively. Also, for simplicity, we define

$$p(x) = \exp\left(-\frac{S(M-K)}{2}x\right)(1+x)^{\frac{S(M-K)}{2}}. \quad (31)$$

Lemma 1: For any positive δ , we have

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} &\leq 2 \exp\left(-\frac{S(M-K)}{2}d_1\right)(1+d_1)^{\frac{S(M-K)}{2}} \\ &= 2p(d_1) \end{aligned} \quad (32)$$

where the function p is defined in (31), and

$$d_1 := \frac{M\delta}{(M-K)\sigma^2} > 0. \quad (33)$$

Proof: From (6), we have

$$\mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} = \mathbb{P}\{Z_{\mathcal{I}} \leq W_1\} + \mathbb{P}\{Z_{\mathcal{I}} \geq W_2\} \quad (34)$$

where $Z_{\mathcal{I}}$ is defined in Lemma 3, and

$$W_i = S(M-K) + (-1)^i SM\delta/\sigma^2, \quad i = 1, 2.$$

Applying the Chernoff bound [31] to (34) yields

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} &\leq \sum_{i=1}^2 \exp(-t_i W_i) \mathbb{E}[\exp(t_i Z_{\mathcal{I}})] \\ &= \sum_{i=1}^2 \underbrace{\exp(-t_i W_i) (1 - 2t_i)^{-S(M-K)/2}}_{=: f(t_i; W_i)} \end{aligned} \quad (35)$$

where the equality is from Lemma 3, $t_1 < 0$ and $t_2 \in (0, \frac{1}{2})$. As each $f(t_i; W_i)$ is convex, $t_i = t_i^*$ at $f^{(1)}(t_i; W_i) = 0$ yields the minimizer of $f(t_i; W_i)$, where

$$t_i^* = 2^{-1} \left(1 - W_i^{-1} S(M-K)\right), \quad i = 1, 2.$$

Thus, $f(t_i; W_i) \geq f(t_i^*; W_i)$ for each i . If $W_1 \leq 0$, it is clear that $\mathbb{P}\{Z_{\mathcal{I}} \leq W_1\} = 0$ because $Z_{\mathcal{I}}$ is quadratic. Thus,

$$\mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} = \mathbb{P}\{Z_{\mathcal{I}} \geq W_2\} \leq f(t_2^*; W_2) = p(d_1) \quad (36)$$

where $p(d_1)$ and d_1 are defined in (32) and (33), respectively. If $W_1 > 0$ then $f(t_1^*; W_1) \leq f(t_2^*; W_2)$ because

$$\begin{aligned} \log f(t_1^*; W_1) - \log f(t_2^*; W_2) \\ = S(M-K) [d_1 + 2 \log(1-d_1) - 2 \log(1+d_1)] < 0. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} &= f(t_1^*; W_1) + f(t_2^*; W_2) \leq 2f(t_2^*; W_2) \\ &= 2 \exp\left(-\frac{S(M-K)}{2}d_1\right)(1+d_1)^{\frac{S(M-K)}{2}}. \end{aligned} \quad (37)$$

Finally, combining (36) and (37) leads to (32). \blacksquare

Lemma 2: Let $\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}$ and a matrix $\mathbf{R}_{\mathcal{J}}$ be

$$\mathbf{R}_{\mathcal{J}} = \begin{bmatrix} \alpha_{\mathcal{J},1} \mathbf{I}_{M-K} & & \\ & \ddots & \\ & & \alpha_{\mathcal{J},S} \mathbf{I}_{M-K} \end{bmatrix} \quad (38)$$

where

$$\alpha_{\mathcal{J},s} := \sigma^2 + \|\mathbf{x}_{\mathcal{I} \setminus \mathcal{J}}^s\|_2^2 > 0. \quad (39)$$

Consider any positive δ such that

$$0 < \delta < (1 - K/M) (\lambda_{\min}(\mathbf{R}_{\mathcal{J}}) - \sigma^2)$$

where $\lambda_{\min}(\mathbf{R}_{\mathcal{J}})$ is the smallest eigenvalue of $\mathbf{R}_{\mathcal{J}}$. Then,

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} &\leq \exp\left(-\frac{S(M-K)}{2} (d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1)\right) d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})}^{\frac{S(M-K)}{2}} \\ &= p(d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1) \\ &\leq p(d_{2,\alpha^*} - 1) \end{aligned} \quad (40)$$

where the function p is defined in (31),

$$d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} := \frac{(M-K)\sigma^2 + M\delta}{(M-K)\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} \in (0, 1), \quad (41)$$

$$\alpha^* := \sigma^2 + x_{\min}^2, \quad (42)$$

and

$$x_{\min}^2 = \min_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \min_{s \in \{1, 2, \dots, S\}} \underbrace{\|\mathbf{x}_{\mathcal{I} \setminus \mathcal{J}}^s\|_2^2}_{=: x_{\min, \mathcal{J}}^2}. \quad (43)$$

Proof: From (7), we have

$$\mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} = \mathbb{P}\{Z_{\mathcal{J}} < W_1\} - \mathbb{P}\{Z_{\mathcal{J}} < W_2\} \leq \mathbb{P}\{Z_{\mathcal{J}} < W_1\} \quad (44)$$

where $Z_{\mathcal{J}}$ is defined in Lemma 4, and

$$W_i = S(M-K)\sigma^2 - (-1)^i SM\delta, \quad i = 1, 2. \quad (45)$$

Applying the Chernoff bound [31] to (44) yields for $t < 0$,

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} &\leq \exp(-t W_1) \mathbb{E}[\exp(t Z_{\mathcal{J}})] \\ &= \exp(-t W_1) \prod_{i=1}^{S(M-K)} (1 - 2t \lambda_i(\mathbf{R}_{\mathcal{J}}))^{-1/2} \\ &\leq \exp(-t W_1) (1 - 2t \lambda_{\min}(\mathbf{R}_{\mathcal{J}}))^{-S(M-K)/2} \\ &=: f(t; W_1) \end{aligned} \quad (46)$$

where the equality is from Lemma 4 and the second inequality is due to that all the eigenvalues are positive. We then define a function $h(t) := \log f(t; W_1)$. Then,

$$h^{(2)}(t) = 2S(M-K) \lambda_{\min}^2(\mathbf{R}_{\mathcal{J}}) (1 - 2t \lambda_{\min}(\mathbf{R}_{\mathcal{J}}))^{-2} > 0$$

which implies that h is convex with respect to t . It leads to that f in (46) is logarithmically convex. Thus $t = t^*$ at $f^{(1)}(t; W_1) = 0$ yields the minimizer of $f(t; W_1)$ where

$$t^* = 2^{-1} \left(\lambda_{\min}^{-1}(\mathbf{R}_{\mathcal{J}}) - W_1^{-1} S(M - K) \right) < 0.$$

Substituting t^* in (46) yields

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} &\leq f(t^*; W_1) \\ &= \exp\left(-\frac{S(M-K)}{2} \left(d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1\right)\right) d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})}^{\frac{S(M-K)}{2}} \\ &= p(d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1) \end{aligned} \quad (47)$$

where $d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})}$ is defined in (41) and p is defined in (31).

Next, let $\beta = 2^{-1} S(M - K)$ and $x = d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})}$ in the upper bound (47). Then, we have $p(x - 1) = x^\beta \exp(-\beta(x - 1))$, where

$$\frac{\partial p(x - 1)}{\partial x} = \beta x^{\beta-1} \exp(-\beta(x - 1)) (x^{-1} - 1) > 0 \quad (48)$$

and

$$\frac{\partial x}{\partial \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} = -x < 0. \quad (49)$$

Due to (48) and (49),

$$\begin{aligned} \frac{\partial p(x - 1)}{\partial \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} &= \frac{\partial p(x - 1)}{\partial x} \frac{\partial x}{\partial \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} \\ &= -\beta x^{\beta-1} \exp(-\beta(x - 1)) (x^{-1} - 1) < 0 \end{aligned}$$

which shows that the upper bound in (47) is decreasing with respect to $\lambda_{\min}(\mathbf{R}_{\mathcal{J}})$. Then, remind that the matrix in (38) is the covariance matrix of a multivariate Gaussian vector \mathbf{b} in (58). Then for any incorrect support set, its smallest eigenvalue can be easily computed and lower bounded by

$$\lambda_{\min}(\mathbf{R}_{\mathcal{J}}) = \min_{s \in \{1, 2, \dots, S\}} \alpha_{\mathcal{J}, s} = \sigma^2 + x_{\min, \mathcal{J}}^2 \geq \alpha^* \quad (50)$$

where $x_{\min, \mathcal{J}}^2$ is defined in (43) and α^* is defined in (42). Thus, for any incorrect support set $\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}$, we conclude that

$$\mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} \leq p(d_{2, \lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1) \leq p(d_{2, \alpha^*} - 1)$$

which completes the proof. \blacksquare

APPENDIX B LEMMAS 3 AND 4

First of all, we give the Scharf's theorem [33] to compute the moment generating function of a quadratic random variable. We then make Lemmas 3 and 4 to give the moment generating functions of the random variables of $\mathcal{E}_{\mathcal{I}}^c$ and $\mathcal{E}_{\mathcal{J}}$ that were used in the proofs of Lemmas 1 and 2, respectively.

Scharf's Theorem [33, p. 64]: Let $\mathbf{b} \in \mathbb{R}^N$ be a multivariate Gaussian vector with a mean \mathbf{m} and a covariance \mathbf{R} . Then a random variable $Q \triangleq (\mathbf{b} - \mathbf{m})^T (\mathbf{b} - \mathbf{m})$ is quadratic with $\mathbb{E}[Q] = \text{tr}[\mathbf{R}]$, $\mathbb{V}[Q] = 2\text{tr}[\mathbf{R}^T \mathbf{R}]$ and for any t

$$\mathbb{E}[\exp(tQ)] = \prod_{i=1}^N (1 - 2t\lambda_i(\mathbf{R}))^{-1/2}.$$

Lemma 3: In (6), define a quadratic random variable

$$Z_{\mathcal{I}} := \sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s) \mathbf{y}^s\|_2^2 / \sigma^2. \quad (51)$$

Then, $\mathbb{E}[Z_{\mathcal{I}}] = S(M - K)$, $\mathbb{V}[Z_{\mathcal{I}}] = 2S(M - K)$ and for any $0 < t < 0.5$,

$$\mathbb{E}[\exp(tZ_{\mathcal{I}})] = (1 - 2t)^{-S(M-K)/2}. \quad (52)$$

Proof: The orthogonal projection matrix is decomposed as

$$\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s) = \mathbf{U}_{\mathcal{I}}^s \mathbf{D}^s (\mathbf{U}_{\mathcal{I}}^s)^T$$

where \mathbf{D}^s is a diagonal matrix, whose first $M - K$ diagonals are ones and the remains are zeros, and $\mathbf{U}_{\mathcal{I}}^s$ is a unitary matrix. Then,

$$\begin{aligned} Z_{\mathcal{I}} &= \sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s) \mathbf{y}^s\|_2^2 / \sigma^2 = \sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{I}}^s) \mathbf{n}^s\|_2^2 / \sigma^2 \\ &= \sum_{s=1}^S \left\| \mathbf{D}^s \underbrace{(\mathbf{U}_{\mathcal{I}}^s)^T \mathbf{n}^s / \sigma^2}_{=: \mathbf{w}^s} \right\|_2^2 = \sum_{s=1}^S \|\mathbf{D}^s \mathbf{w}^s\|_2^2 \end{aligned} \quad (53)$$

where \mathbf{w}^s is a multivariate Gaussian vector with mean $\mathbf{0}_M$ and covariance \mathbf{I}_M . Since the first $M - K$ diagonal elements of each diagonal matrix are ones, we have

$$\begin{aligned} Z_{\mathcal{I}} &= \sum_{s=1}^S \|\mathbf{D}^s \mathbf{w}^s\|_2^2 = \sum_{s=1}^S \sum_{i=1}^{M-K} |w^s(i)|^2 \\ &= \sum_{s=1}^S (\mathbf{w}_{\mathcal{P}}^s)^T \mathbf{w}_{\mathcal{P}}^s = \mathbf{w}^T \mathbf{w} \end{aligned} \quad (54)$$

which is quadratic, where

$$\mathbf{w}_{\mathcal{P}}^s = [w^s(1) \quad w^s(2) \quad \dots \quad w^s(M - K)]^T$$

and

$$\mathbf{w} = [(\mathbf{w}_{\mathcal{P}}^1)^T \quad (\mathbf{w}_{\mathcal{P}}^2)^T \quad \dots \quad (\mathbf{w}_{\mathcal{P}}^S)^T]^T. \quad (55)$$

In (53), \mathbf{w}^s is determined by $\mathbf{U}_{\mathcal{I}}^s$ and \mathbf{n}^s . Since the elements of $\mathbf{U}_{\mathcal{I}}^s$ and \mathbf{n}^s are independent, \mathbf{w}^i and \mathbf{w}^j are mutually independent for any $1 \leq i \neq j \leq S$. The covariance matrix of \mathbf{w} is an identity matrix. Thus, applying the Scharf's theorem to $Z_{\mathcal{I}}$ completes the proof. \blacksquare

Lemma 4: In (7), for any $\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}$, define a quadratic random variable

$$Z_{\mathcal{J}} := \sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{J}}^s) \mathbf{y}^s\|_2^2. \quad (56)$$

Then, $\mathbb{E}[Z_{\mathcal{J}}] = \text{tr}[\mathbf{R}_{\mathcal{J}}]$, $\mathbb{V}[Z_{\mathcal{J}}] = 2\text{tr}[\mathbf{R}_{\mathcal{J}}^T \mathbf{R}_{\mathcal{J}}]$ and for any t ,

$$\mathbb{E}[\exp(tZ_{\mathcal{J}})] = \prod_{i=1}^{S(M-K)} (1 - 2t\lambda_i(\mathbf{R}_{\mathcal{J}}))^{-1/2},$$

where $\mathbf{R}_{\mathcal{J}}$ is given in (38).

Proof: Similar to the proof of Lemma 3,

$$\mathbf{Q}(\mathbf{F}_{\mathcal{J}}^s) = \mathbf{U}_{\mathcal{J}}^s \mathbf{D}^s (\mathbf{U}_{\mathcal{J}}^s)^T$$

where \mathbf{D}^s is a diagonal matrix, whose first $M - K$ diagonals are ones and the remains are zeros, and $\mathbf{U}_{\mathcal{J}}^s$ is a unitary matrix. Then,

$$\begin{aligned} Z_{\mathcal{J}} &= \sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{J}}^s) \mathbf{y}^s\|_2^2 = \sum_{s=1}^S \|\mathbf{Q}(\mathbf{F}_{\mathcal{J}}^s) \mathbf{c}^s\|_2^2 \\ &= \sum_{s=1}^S \left\| \mathbf{D}^s \underbrace{(\mathbf{U}_{\mathcal{J}}^s)^T \mathbf{c}^s}_{=\mathbf{b}^s} \right\|_2^2 = \sum_{s=1}^S \|\mathbf{D}^s \mathbf{b}^s\|_2^2 \end{aligned} \quad (57)$$

where \mathbf{b}^s is a multivariate Gaussian vector with mean $\mathbf{0}_M$ and

$$\mathbb{V}[\mathbf{b}^s] = \left(\sigma^2 + \|\mathbf{x}_{\mathcal{I} \setminus \mathcal{J}}^s\|_2^2 \right) \mathbf{I}_M$$

and $\mathbf{c}^s = \mathbf{n}^s + \sum_{u \in \mathcal{I} \setminus \mathcal{J}} \mathbf{f}_u^s x^s(u)$. Since the first $M - K$ diagonal elements of each diagonal matrix are ones, we have

$$\begin{aligned} Z_{\mathcal{J}} &= \sum_{s=1}^S \|\mathbf{D}^s \mathbf{b}^s\|_2^2 = \sum_{s=1}^S \sum_{i=1}^{M-K} |b^s(i)|^2 \\ &= \sum_{s=1}^S (\mathbf{b}_{\mathcal{P}}^s)^T \mathbf{b}_{\mathcal{P}}^s = \mathbf{b}^T \mathbf{b} \end{aligned} \quad (58)$$

which is quadratic, where

$$\mathbf{b}_{\mathcal{P}}^s = [b^s(1) \quad b^s(2) \quad \dots \quad b^s(M-K)]^T$$

and

$$\mathbf{b} = [(\mathbf{b}_{\mathcal{P}}^1)^T \quad (\mathbf{b}_{\mathcal{P}}^2)^T \quad \dots \quad (\mathbf{b}_{\mathcal{P}}^S)^T]^T.$$

In (57), \mathbf{b}^s is determined by $\mathbf{U}_{\mathcal{J}}^s$, \mathbf{n}^s and $\{\mathbf{f}_u^s : u \in \mathcal{I} \setminus \mathcal{J}\}$. Since the elements of $\mathbf{U}_{\mathcal{J}}^s$, \mathbf{n}^s and $\{\mathbf{f}_u^s : u \in \mathcal{I} \setminus \mathcal{J}\}$ are independent, \mathbf{b}^i and \mathbf{b}^j are mutually independent for any $1 \leq i \neq j \leq S$. The covariance matrix of \mathbf{b} is diagonal as shown in (38). Thus, applying the Scharf's theorem to $Z_{\mathcal{J}}$ completes the proof. ■

APPENDIX C

PROOFS OF THEOREMS 1, 2 AND 3

A. Proof of Theorem 1

It is clear that K goes to infinity as N goes to infinity in the linear sparsity regime. Then, let $M = cK$ where $c > 1$. From (32),

$$\log \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq 2^{-1}SK(c-1) \underbrace{(\log(1+d_1) - d_1)}_{=:A} + \log 2$$

where $A < 0$ due to (33). Thus,

$$\lim_{N \rightarrow \infty} \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq \lim_{K \rightarrow \infty} \exp\left(2^{-1}SK(c-1)A + \log 2\right) = 0$$

implying that the probability that the correct support set is not δ jointly typical with all the measurement vectors vanishes.

Next, from (40),

$$\begin{aligned} \log \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} &\leq \log\left(\binom{N}{K} p(d_{2,a^*} - 1)\right) \\ &= \log\binom{N}{K} + 2^{-1}SK(c-1) \underbrace{(\log(1-t) + t)}_{=: \gamma} \\ &\leq K \underbrace{\left(1 + \log \frac{N}{K} + 2^{-1}S(c_1 - 1)\gamma\right)}_{=: \eta} \end{aligned} \quad (59)$$

where the last inequality is due to

$$\binom{N}{K} \leq \exp\left(K \log \frac{Ne}{K}\right). \quad (60)$$

In (59), $\gamma < 0$ for any t where

$$t = \frac{1 - \rho^{-1}}{1 + \text{SNR}_{\min}^{-1}} \in (0, 1). \quad (61)$$

If $c > 1 + S^{-1}v_1$, then $\eta < 0$, which yields

$$\lim_{N \rightarrow \infty} \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} \leq \lim_{K \rightarrow \infty} \exp(K\eta) = 0$$

implying that the probability that all incorrect support sets are δ jointly typical with all the measurement vectors vanishes. Thus the failure probability p_{err} defined in (8) converges to zero if M satisfies (9).

Next, the remain is to derive (11) in the sublinear sparsity regime. Similarly, let $M = K + cK \log \frac{N}{K}$ where $c > 1$. From (32),

$$\log \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq 2^{-1}ScK \log \frac{N}{K} \underbrace{(\log(1+d_1) - d_1)}_{=:A} + \log 2$$

where $A < 0$ due to (33). Thus,

$$\lim_{N \rightarrow \infty} \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq \lim_{N \rightarrow \infty} \exp\left(2^{-1}ScKA \log \frac{N}{K} + \log 2\right) = 0$$

implying that the probability that the correct support set is not δ jointly typical with all the measurement vectors vanishes.

Then, from (40),

$$\begin{aligned} \log \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} &\leq \log\left(\binom{N}{K} p(d_{2,a^*} - 1)\right) \\ &= \log\binom{N}{K} + 2^{-1}ScK \underbrace{(\log(1-t) + t)}_{=: \gamma} \log \frac{N}{K} \\ &\leq K \underbrace{\left(1 + 2^{-1}Sc\gamma\right)}_{=: \eta} \log \frac{N}{K} + K \end{aligned}$$

where the last inequality is due to the bound in (60) and $\gamma < 0$ for any t in (61). If $c > S^{-1}v_2$, then $\eta < 0$, which yields

$$\lim_{N \rightarrow \infty} \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} \leq \lim_{N \rightarrow \infty} \exp\left(K\eta \log \frac{N}{K} + K\right) = 0$$

implying that the probability that all incorrect support sets are δ jointly typical with all the measurement vectors vanishes. Thus, the failure probability p_{err} defined in (8) converges to zero if M satisfies (11), which completes the proof. ■

B. Proof of Theorem 2

From Lemma 1,

$$\mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq 2 \underbrace{\left(\exp\left(-\frac{M-K}{2}d_1\right) (1+d_1)^{\frac{M-K}{2}}\right)^S}_{=: \mu_{\mathcal{I}}}. \quad (62)$$

If $M \geq K + 1$, we have

$$\log \mu_{\mathcal{I}} = 2^{-1} (M - K) (\log(1 + d_1) - d_1) < 0 \quad (63)$$

due to (33), which implies $\mu_{\mathcal{I}} < 1$. From Lemma 2,

$$\mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} \leq \left(\underbrace{\exp\left(-\frac{M-K}{2} (d_{2,\alpha^*} - 1)\right) d_{2,\alpha^*}^{\frac{M-K}{2}}}_{=:\mu_{\mathcal{J}}} \right)^S. \quad (64)$$

Similarly, if $M \geq K + 1$, we have

$$\log \mu_{\mathcal{J}} = 2^{-1} (M - K) (\log(1 - t) + t) < 0 \quad (65)$$

due to (61), which implies $\mu_{\mathcal{J}} < 1$. Thus, we conclude

$$\lim_{S \rightarrow \infty} p_{\text{err}} \leq 2 \lim_{S \rightarrow \infty} \mu_{\mathcal{I}}^S + \binom{N}{K} \lim_{S \rightarrow \infty} \mu_{\mathcal{J}}^S = 0$$

for $M \geq K + 1$ which completes the proof. ■

C. Proof of Theorem 3

The mutual information in (24) is bounded by

$$\begin{aligned} \mathbb{I}(\mathbf{x}; \mathbf{y} | \tilde{\mathbf{F}}) &= h(\mathbf{y} | \tilde{\mathbf{F}}) - h(\mathbf{y} | \mathbf{x}, \tilde{\mathbf{F}}) \leq h(\mathbf{y}) - h(\mathbf{n}) \\ &\leq \sum_{i=1}^{SM} h(y_i) - h(\mathbf{n}) \\ &\leq 2^{-1} SM \left(\log(2\pi e (Kx_{\min}^2 + \sigma^2)) - \log(2\pi e \sigma^2) \right) \\ &= 2^{-1} SM \log(1 + K \times \text{SNR}_{\min}) \end{aligned}$$

where $h(\mathbf{x})$ is the differential entropy of \mathbf{x} , and $h(\mathbf{x} | \mathbf{y})$ is the conditional entropy of \mathbf{x} given \mathbf{y} . The last inequality is due to that the Gaussian distribution maximizes the differential entropy. The denominator in (24) is bounded by

$$\log(|\mathcal{X}_{\{x_{\min}\}}| - 1) = \log\left(\binom{N}{K} - 1\right) > K \log \frac{N}{K}$$

for sufficiently large N . Then,

$$\begin{aligned} p_{\text{err}} &= \mathbb{E}_{\tilde{\mathbf{F}}} \sup_{\mathbf{x} \in \mathcal{X}_{\{x_{\min}\}}} \mathbb{P}\{\hat{\mathcal{I}} \neq \mathcal{I} | \mathbf{x}, \tilde{\mathbf{F}}\} \\ &\geq \mathbb{E}_{\tilde{\mathbf{F}}} \min_{\hat{\mathbf{x}} \in \mathcal{X}_{\{x_{\min}\}}} \max_{\mathbf{x} \in \mathcal{X}_{\{x_{\min}\}}} \mathbb{P}\{\hat{\mathbf{x}} \neq \mathbf{x} | \mathbf{x}, \tilde{\mathbf{F}}\} \\ &> 1 - \frac{2^{-1} SM \log(1 + K \times \text{SNR}_{\min}) + \log 2}{K \log \frac{N}{K}}. \quad (66) \end{aligned}$$

From (66), the failure probability is bounded away from zero by zero if (25) is satisfied, which completes the proof. ■

APPENDIX D

PROOFS OF COROLLARIES 1, 2 AND 3

A. Proof of Corollary 1

From the inequality $\log(1 + x) \leq \frac{2x}{2+x}$ for $x \in (-1, 0]$,

$$v_2 = -\frac{2}{\log(1-t) + t} < \frac{4-2t}{t^2} < \frac{4}{t^2} \quad (67)$$

where t is defined in (61). Then,

$$\frac{v_2}{S} < \frac{4}{St^2}. \quad (68)$$

From (61),

$$\sqrt{S}t = \frac{1 - \rho^{-1}}{-\sqrt{S} + (\sqrt{S} \times \text{SNR}_{\min})^{-1}}. \quad (69)$$

Combining (11), (68) and (69) leads to (13). This approach is used to get (14) using the following equality

$$v_1 = v_2 (1 - \log \beta) \quad (70)$$

where $\lim_{N \rightarrow \infty} \frac{K}{N} = \beta \in (0, 1/2)$, which completes the proof. ■

B. Proof of Corollary 2

Substituting $\alpha = \frac{2}{3}$ in (15), and rearranging the result with respect to t can yield $\frac{2}{3} \leq t < 1$, where t is defined in (61). Then from (67), a simple computation yields that

$$v_2 < \frac{4-2t}{t^2} \leq \frac{4}{t}$$

which immediately yields that

$$\frac{v_2}{S} < \frac{4}{St}. \quad (71)$$

where

$$St = \frac{1 - \rho^{-1}}{S^{-1} + (S \times \text{SNR}_{\min})^{-1}}. \quad (72)$$

Combining (11), (71) and (72) leads to (16). This approach is used to get (17) using (70), which completes the proof. ■

C. Proof of Corollary 3

We assume that $\mu_{\mathcal{I}} \geq \mu_{\mathcal{J}}$ and

$$p_{\text{err}} \leq \mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} + \sum_{\mathcal{J} \in \mathcal{S} \setminus \mathcal{I}} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} \leq \left(\binom{N}{K} + 2\right) \mu_{\mathcal{I}}^S < \varepsilon < 1. \quad (73)$$

Then, if the number of measurement vectors satisfies

$$S > \frac{\log \varepsilon - \log\left(\binom{N}{K} + 2\right)}{\log \mu_{\mathcal{I}}} > 0, \quad (74)$$

(73) is achieved for small ε , and hence, reliable support set reconstruction is possible. If $\mu_{\mathcal{I}} < \mu_{\mathcal{J}}$, we obtain inequalities similar to (73) and (74) by replacing $\mu_{\mathcal{I}}$ by $\mu_{\mathcal{J}}$, where

$$S > \frac{\log \varepsilon - \log\left(\binom{N}{K} + 2\right)}{\log \mu_{\mathcal{J}}} > 0. \quad (75)$$

Combining (74) and (75) yields (18).

Next, a simple computation yields that for any d_1 in (33),

$$\frac{\partial \log \mu_{\mathcal{I}}}{\partial d_1} = -\frac{d_1}{2(1+d_1)} < 0$$

where $\log \mu_{\mathcal{I}}$ is given in (63). From (33), we see $d_1 \propto \text{SNR}_{\min}$ that leads to $\log \mu_{\mathcal{I}} \propto \text{SNR}_{\min}^{-1}$. Also, for any t in (61),

$$\frac{\partial \log \mu_{\mathcal{J}}}{\partial t} = -\frac{t}{2(1-t)} < 0$$

where $\log \mu_{\mathcal{J}}$ is given in (65). From (61), we see $t \propto \text{SNR}_{\min}$ that leads to $\log \mu_{\mathcal{J}} \propto \text{SNR}_{\min}^{-1}$. Hence, the sufficient condition on S in (18) turns out to be a decreasing function with respect to SNR_{\min} , which completes the proof. ■

APPENDIX E
PROOFS OF PROPOSITIONS 1 AND 2

First of all, we introduce the exponential inequalities [32], and use them in the proofs of Propositions 1 and 2.

A. The Exponential Inequalities [32]

Let Y_i , $i = 1, 2, \dots, D$ be i.i.d. Gaussian variables with a zero mean and a unit variance. Then, let α_i , $i = 1, 2, \dots, D$ be non-negative. We set

$$|\alpha|_\infty = \sup |\alpha_i|, \quad |\alpha|_2^2 = \sum_{i=1}^D \alpha_i^2$$

and let

$$Y = \sum_{i=1}^D \alpha_i (Y_i^2 - 1). \quad (76)$$

Then, the following inequalities hold for any positive x

$$\mathbb{P}\{Y \geq 2|\alpha|_2 \sqrt{x} + 2|\alpha|_\infty x\} \leq \exp(-x) \quad (77)$$

$$\mathbb{P}\{Y \leq -2|\alpha|_2 \sqrt{x}\} \leq \exp(-x). \quad (78)$$

B. Proof of Proposition 1

In the proof of Lemma 3, $Z_{\mathcal{I}}$ is represented by

$$Z_{\mathcal{I}} = \sum_{s=1}^S \sum_{i=1}^{M-K} w^s(i)^2$$

where $w^s(i)$ is Gaussian with a zero mean and a unit variance. Define a random variable Y as

$$Y = Z_{\mathcal{I}} - S(M-K) = \sum_{s=1}^S \sum_{i=1}^{M-K} (w^s(i)^2 - 1)$$

which is of the form of (76). Then,

$$\mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} = \underbrace{\mathbb{P}\{Y \leq -SM\delta/\sigma^2\}}_{=:A} + \underbrace{\mathbb{P}\{Y \geq SM\delta/\sigma^2\}}_{=:B}.$$

Combining A with (78) gives

$$\begin{aligned} \mathbb{P}\{Y \leq -SM\delta/\sigma^2\} &= \mathbb{P}\{Y \leq -2\sqrt{S(M-K)}x\} \\ &\leq \underbrace{\exp\left(-\frac{SM^2\delta^2}{4(M-K)\sigma^4}\right)}_{=:C} \end{aligned}$$

and combining B with (77) gives

$$\begin{aligned} \mathbb{P}\{Y \geq SM\delta/\sigma^2\} &= \mathbb{P}\{Y \geq 2\sqrt{S(M-K)}x + 2x\} \\ &\leq p_{1,\text{exp}} \end{aligned}$$

where $p_{1,\text{exp}}$ is defined in (28). It is readily seen that $p_{1,\text{exp}} \geq C$, which leads to $\mathbb{P}\{\mathcal{E}_{\mathcal{I}}^c\} \leq 2p_{1,\text{exp}}$.

Next, from (32) and (28),

$$\log p(d_1) = 2^{-1}S(M-K)(\log(1+d_1) - d_1)$$

and

$$\log p_{1,\text{exp}} = -2^{-1}S(M-K)d_1^2(2+4d_1)^{-1}$$

where $d_1 > 0$ is defined in (33). Then, we have

$$\log \frac{p(d_1)}{p_{1,\text{exp}}} = \frac{S(M-K)}{2} \underbrace{\left(\log(1+d_1) - d_1 + d_1^2(2+4d_1)^{-1}\right)}_{=:g(d_1)}.$$

For any $d_1 > 0$, $\frac{\partial g(d_1)}{\partial d_1} = -d_1^2(2+3d_1)(1+d_1)^{-1}(1+2d_1)^{-2} < 0$ and $\max_{d_1>0} g(d_1) = 0$. Thus, we conclude $\log \frac{p(d_1)}{p_{1,\text{exp}}} \leq 0$, which completes the proof. ■

C. Proof of Proposition 2

In the proof of Lemma 4, $Z_{\mathcal{J}}$ is represented by

$$\begin{aligned} Z_{\mathcal{J}} &= \sum_{s=1}^S \sum_{i=1}^{M-K} b^s(i)^2 \\ &= \sum_{s=1}^S \sum_{i=1}^{M-K} \alpha_{\mathcal{J},s} g^s(i)^2 \end{aligned}$$

where $\alpha_{\mathcal{J},s}$ is defined in (39) and $g^s(i)$ is Gaussian with a zero mean and a unit variance. Define a new random variable Y as

$$\begin{aligned} Y &= Z_{\mathcal{J}} - S(M-K) \\ &= \sum_{s=1}^S \sum_{i=1}^{M-K} \alpha_{\mathcal{J},s} (g^s(i)^2 - 1) \end{aligned}$$

which is of the form of (76). Then, from (44)

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_{\mathcal{J}}\} &\leq \mathbb{P}\left\{Y < SM\delta - (M-K) \sum_{s=1}^S \|\mathbf{x}_{\mathcal{I} \setminus \mathcal{J}}^s\|_2^2\right\} \\ &\leq \mathbb{P}\left\{Y < \underbrace{SM\delta - S(M-K)x_{\min,\mathcal{J}}^2}_{=:A}\right\} \\ &\leq p_{2,\mathcal{J},\text{exp}} \end{aligned} \quad (79)$$

where $p_{2,\mathcal{J},\text{exp}}$ is defined in (30), the last inequality is due to (78). Due to (29), A is negative. Thus the exponential inequality of (78) gives the upper bound $p_{2,\mathcal{J},\text{exp}}$.

Next, from (40) and (30),

$$\log p(d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1) = 2^{-1}S(M-K)(t + \log(1-t))$$

and

$$\begin{aligned} \log p_{2,\mathcal{J},\text{exp}} &\geq -\frac{S(M-K)}{4} \left(\frac{x_{\min,\mathcal{J}}^2 - \frac{M\delta}{M-K}}{x_{\min,\mathcal{J}}^2 + \sigma^2}\right)^2 \\ &= -4^{-1}S(M-K)t^2 \end{aligned}$$

where $t \in (0, 1)$, is defined in (61) and the inequality is due to (50). Then,

$$\log \frac{p(d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1)}{p_{2,\mathcal{J},\text{exp}}} \leq \frac{S(M-K)}{4} \underbrace{(t^2 + 2t + 2\log(1-t))}_{=:g(t)}.$$

For any $t \in (0, 1)$, $\frac{\partial g(t)}{\partial t} = -2t^2(1-t)^{-1} < 0$ and

$\max_{t \in (0, 1)} g(t) = 0$. We conclude $\log \frac{p(d_{2,\lambda_{\min}(\mathbf{R}_{\mathcal{J}})} - 1)}{p_{2,\mathcal{J},\text{exp}}} \leq 0$. It completes the proof. ■

REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
- [2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [3] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

- [4] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [5] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
- [6] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
- [7] E. J. Candès and M. B. Wakin, "An introduction to compressive sensing," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.
- [8] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk. (2009). "Distributed compressed sensing." [Online]. Available: <https://arxiv.org/pdf/0901.3403.pdf>
- [9] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2005, pp. 1537–1541.
- [10] C. Caione, D. Brunelli, and L. Benini, "Compressive sensing optimization for signal ensembles in WSNs," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 382–392, Feb. 2014.
- [11] W. Chen, R. D. Rodrigues, and I. J. Wassell, "Distributed compressive sensing reconstruction via common support discovery," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011, pp. 1–5.
- [12] J. Oliver, W.-B. Lee, and H.-N. Lee, "Filters with random transmittance for improving resolution in filter-array-based spectrometers," *Opt. Exp.*, vol. 21, no. 4, pp. 3969–3989, Feb. 2013.
- [13] S. Gogineni and A. Nehorai, "Target estimation using sparse modeling for distributed MIMO radar," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5315–5325, Nov. 2011.
- [14] Y. Wu *et al.*, "Accelerated MR diffusion tensor imaging using distributed compressed sensing," *Magn. Reson. Med.*, vol. 71, no. 2, pp. 763–772, Feb. 2014.
- [15] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer, 2010.
- [16] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, May 2012.
- [17] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2015.
- [18] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.
- [19] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.
- [20] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.
- [21] M. Akçakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.
- [22] J. Scarlett, J. S. Evans, and S. Dey, "Compressed sensing with prior information: Information-theoretic limits and practical decoders," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 427–439, Jan. 2013.
- [23] J. Scarlett and V. Cevher, "Limits on support recovery with probabilistic models: An information-theoretic framework," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 593–620, Jan. 2017.
- [24] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.
- [25] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 4634–4643, Dec. 2006.
- [26] G. Tang and A. Nehorai, "Performance analysis for sparse support recovery," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1383–1399, Mar. 2010.
- [27] Y. Jin and B. D. Rao, "Support recovery of sparse signals in the presence of multiple measurement Vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3139–3157, May 2013.
- [28] M. F. Duarte, M. B. Wakin, D. Braon, S. Sarvotham, and R. G. Baraniuk, "Measurement bounds for sparse signal ensembles via graphical models," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4280–4289, Jul. 2013.
- [29] J. Kim, O. Lee, and J. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 278–301, Jan. 2012.
- [30] J. D. Blanchard and M. E. Davies, "Recovery guarantees for rank aware pursuits," *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 427–430, Jul. 2012.
- [31] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York, NY, USA: Wiley, 2001.
- [32] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1303–1338, Oct. 2000.
- [33] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Reading, MA, USA: Addison-Wesley, 1991.
- [34] R. Rajagopalan and P. K. Varsheny, "Data-aggregation techniques in sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 8, no. 4, pp. 48–63, 4th Quart., 2006.

Sangjun Park received the B.S. degree in computer engineering from the Chungnam National University, Daejeon, Korea, in 2009. He is pursuing a Ph.D degree at the School of Electrical Engineering and Computer Science in the Gwangju Institute of Science and Technology (GIST), Gwangju, Korea. His research interests include information theory, numerical optimization and compressed sensing.

Nam Yul Yu (M'07) received the B.S. degree in electronics engineering from the Seoul National University, Seoul, Korea, in 1995 and the M.S. degree in electronic and electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2000. He received the Ph. D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, Ontario, Canada, in 2007. From 2000 to 2003, he was with Telecommunication Research and Development Center, Samsung Electronics, Korea, where he worked on channel coding schemes for wireless communication systems. In 2007, he was a senior research engineer in LG Electronics, Korea, working on the standardization of the 3GPP-LTE. From 2008 to 2014, he has been working as an Assistant/Associate Professor in the Department of Electrical Engineering at the Lakehead University, Thunder Bay, Ontario, Canada. In 2014, he joined the Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, and is currently working as an Associate Professor in the School of Electrical Engineering and Computer Science. From 2009 to 2011, he served as an Associate Editor for Sequences in the IEEE Transactions on Information Theory. His research interests include sequence design, compressed sensing, and security for wireless communications.

Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He then worked at HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member from 1999 to 2002. From 2002 to 2008, he worked as an Assistant Professor at the University of Pittsburgh, PA, USA. In 2009, he then moved to the School of Electrical Engineering and Computer Science, GIST, Korea, where he is currently affiliated. His areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He has received several prestigious national awards, including the Top 100 National Research and Development Award in 2012, the Top 50 Achievements of Fundamental Researches Award in 2013, and the Science/Engineer of the Month (January 2014). He has written more than fifty international journal publications and a hundred international conferences and workshop papers. He was the Director of Electrical Engineering and Computer Science track within GIST College in 2014. In March 2015, he was appointed as the Dean of Research at GIST.

A Versatile Coexistence Decision-Making System for Efficient TV Whitespace Sharing among Whitespace Objects

M. Asif Raza, Zafar Iqbal, Sang-Seon Byun, Hyunduk Kang, Heung-No Lee*

Abstract—In this paper, a coexistence decision making (CDM) system for efficient TV whitespace (TVWS) sharing among whitespace objects (WSOs), registered in coexistence managers in IEEE 802.19.1 system, is introduced. The proposed system is considered versatile in functionality as it jointly takes care of three distinct channel allocation features; a) optimizing system quality of service (QoS) performance metrics, b) improving TVWS utility and c) satisfying WSO channel demands. Regarding system QoS performance metrics, the TVWS sharing problem is defined as an optimization problem with an aim to maximize the system throughput and minimizing unfairness in allocation. Supporting the WSOs channel demands in a TVWS sharing problem is a multifold task which requires elaborate consideration in different aspects of the system performance. To this end, the variations of the SNR of wireless frequency channels which result in variable throughput gain of the WSOs are also taken care of the proposed CDM system. A fast channel allocation algorithm is then designed that implements the TVWS sharing mechanism in a reasonable amount of time. Additionally, the proposed algorithm improves the TVWS utility by promoting a novel frequency reuse method by exploiting the inter-WSO interference information. Simulation results show the superiority of the proposed algorithm over existing TVWS sharing algorithms.

Index Terms— Frequency Reuse, Lagrangian Relaxation, Linear Approximation, Proportional Fairness, TV Whitespace

I. INTRODUCTION

AN unprecedented increase in the deployment of content delivery networks (CDNs) has resulted in the rapid growth of IP traffic. It is reported that by the end of 2016, global IP traffic exceeded 1 zettabytes (10^{21} bytes) per year, of which 62% is attributed to CDNs [1]. It is also anticipated that by 2019,

nearly two-thirds of global IP traffic will originate from non-PC devices, mainly portable and mobile devices [1]. On the other hand, currently available wireless spectrum is considered insufficient for accommodating such large volumes of data. Fortunately, the digitization of TV transmission has partially relinquished VHF and UHF spectrum [2]. Owing to its low loss and excellent propagation characteristics, the TV spectrum is considered a promising candidate for supporting the growing traffic over wireless channels. Considering the growing demand of the wireless spectrum, the regulatory bodies worldwide [3], [4], [5], have permitted unlicensed use of the TV spectrum under certain limits to protect the incumbents. However, the problem of coexistence of secondary devices operating in the same TV band was not dealt by the regulatory bodies.

The coexistence among secondary devices operating in TV spectrum is considered a challenging task due to signal propagation characteristics of TV channels, spatiotemporal variation of TV spectrum and disparity in network technologies of devices operating in the TV spectrum [6]. These diversities may cause coexistence issues, such as an unresolvable interference, spectrum congestion, diversity in network size, etc., as explained in [6], [7], [8], [9]. To address coexistence issues and regulate access to TV spectrum, IEEE has proposed an 802.19.1 standard [10]. The standard provides a set of procedures to enable coexistence among secondary networks operating in heterogeneous network technologies in TVWS, namely WSOs.

A set of procedures that ensures peaceful coexistence among a set of WSOs operating in the same spectrum is referred to as CDM [11]. In this paper, we define an 802.19.1 compliant CDM system that performs TVWS sharing among a set of WSOs, operating in dissimilar MAC/PHY layer technologies and registered in the coexistence manager (CM); an entity in 802.19.1 coexistence system as shall be defined in section III-A. Note that the TVWS refers to the TV spectrum not in use by licensed operators in a spatio-temporal region [10]. The TVWS sharing problem is modeled as an optimization problem with an aim to maximize the system performance metrics like system throughput and fairness in TVWS allocation. The optimization problem is constrained that the channel demands of the WSOs registered in the neighboring CMs are satisfied. In this perspective, variations of the SNR of wireless frequency channels which result in variable throughput gain of the WSOs are taken care of. Note that the neighboring CMs refer to the set of CMs whose WSOs create interference to each other and such WSOs are neighboring WSOs. Thus, the proposed CDM

- M. Asif Raza is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. E-mail: raza@gist.ac.kr.
- Zafar Iqbal is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. E-mail: zafar@gist.ac.kr.
- Sang-Seon Byun is with the Computer Engineering Department, Catholic University of Pusan, Busan 609-757, South Korea. E-mail: ssbyun@gmail.com.
- Hyunduk Kang is with Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, South Korea. Email: henry@etri.re.kr.
- Heung-No Lee* is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. E-mail: heungno@gist.ac.kr. (*the corresponding author.)

system differs from the notion of traditional node-based, link-based or base-station based channel allocation as reported in the TVWS sharing literature. Moreover, the proposed system also improves the TVWS utility by implementing a frequency reuse (FR) method to spatially reuse the available TV spectrum in a joint time-frequency domain in an *ad hoc* coexisting environment. In this paper, the *ad hoc* coexisting environment refers to the coexistence of both, infrastructure based WSOs like WLAN, and *ad hoc* WSOs like personal area network. An *ad hoc* WSO accounts for a local area network that is built spontaneously, as devices connect with each other. The CDM system proposed in this paper is unique, to the best of our knowledge, in the sense that it jointly focuses three distinct TVWS sharing objectives; a) optimizing system performance metrics during TVWS sharing among WSOs registered in neighboring CMs in 802.19.1 system b) improving the TVWS utility by implementing the FR in a joint time-frequency domain, c) taking care of the channel demands of the *heterogeneous*-WSOs. Such a joint focus to implement multiple distinct channel allocation features makes the proposed system a versatile CDM system.

The remainder of the paper is organized as follows. Section II reviews some related studies. Section III summarizes technical background required to establish the baseline for the techniques used in the paper. The system description and problem formulation are defined in Section IV. Section V discusses the solution method and the proposed algorithm. Section VI presents the simulation results and compares the proposed algorithm with existing algorithms. Finally, Section VII concludes the paper.

II. PREVIOUS WORK

In this section, we summarize some standards and algorithms developed for achieving coexistence among secondary users.

IEEE 802.15.2 [12] and 802.15.4 [13] have partially addressed the coexistence issue among devices operating on wireless local area networks and low power wireless personal area networks, respectively. However, these networks operate on industrial, scientific, and medical bands. On the other hand, IEEE 802.22 has recently defined PHY and MAC layer extensions for TVWS. Similarly, IEEE 802.11af [14] has adopted new cognitive radio features to protect incumbents and achieve efficient spectrum utilization among unlicensed devices. IEEE 802.22.1 has also defined methods for peaceful coexistence when a low-power licensed device such as a microphone broadcaster and an unlicensed device both coexist and share the same channel [15]. The European Computer Manufacturers Association (ECMA) has also defined a specification (ECMA 392) for personal/portable cognitive wireless networks operating in TVWS [16]. However, all these standards define self-coexistence in TVWS operations. Non-availability of cross-platform coexistence mechanisms shall cause issues such as an inability to diagnose interference among networks with dissimilar network technologies and may lead to inefficient utilization of the scarce wireless spectrum [11]. Perceiving the need for cross-platform coexistence mechanisms, IEEE has defined an 802.19.1 standard. This standard provides coexistence

protocols and policies for efficient utilization of TVWS across platforms [10].

On algorithmic perspective, a CDM algorithm that results in fair TVWS sharing among neighboring CMs is presented in [10]. The algorithm applies max-min fairness technique to establish fair share distribution during the TVWS sharing process. The issue with the algorithm in [10] is that it focuses fairness in allocation while no consideration to effective utilization of the available TVWS is taken care. Considering the scarcity of the TV spectrum, especially in highly congested spectrum environments, the effective utilization of the available TV spectrum is also an important factor to be considered. Hesar and Roy [17] have discussed the TVWS sharing formulations in secondary cellular networks. The authors adopt heuristic based approaches to defining greedy algorithms to tackle the identified TVWS sharing problems. However, the proposed greedy algorithm for throughput maximization sub-problem searches the entire network to find an optimal solution. For such an algorithm, search over the space of a possibly very large number of network and channel collocation combinations leads to a high runtime complexity to find an optimal solution. An algorithm for opportunistic whitespace sharing among secondary networks has been presented as a graph coloring problem in [18]. The channel sharing algorithm in [18] solves the sharing problem by classifying the sharing process as network wide channel sharing and its localized version. This scheme, however, has performance issue when interference among neighboring access points is relatively high. This situation is quite common in highly congested areas where many collocated WSOs are deployed. Bahrak and Park [10] proposed an algorithm for CDM among heterogeneous networks. The sharing problem in [10] is modeled as a weighted-sum multi-objective optimization problem (MOOP) that is solved using a modified Boltzmann machine. However, an issue in the weighted-sum approach is that it does not find Pareto optimal points in non-convex regions of the solution space boundary [19]. Thus, some of the potential Pareto optimal points are possibly missed by the weighted-sum method.

Khalil et al., have also performed TVWS sharing among heterogeneous networks by defining an interference graph of the networks [20]. A two-stage algorithm is then designed to achieve spectrum sharing among graph nodes. The algorithm maximizes fairness by maximizing the frequency reuse. However, the channel sharing algorithm in [20] has polynomial runtime complexity $\mathcal{O}(N^3)$, for the number of networks (N). This complexity shows that in areas with a high number of deployed networks, the algorithm shall require substantial channel allocation time. Zhang et al., [21] adapt ecology based species competition model to develop a coexistence mechanism called ecological Species Competition based HETerogeneous networks coexistence MEchanism (SCHEME). The SCHEME enables each coexisting network to adjust achieved bandwidth per its QoS requirements dynamically. However, the SCHEME requires the number of channels to be larger than the number of coexisting networks. Such condition cannot be fulfilled in highly congested urban areas where a limited number of TV channels is available for unlicensed use. We have addressed this issue in the

channel allocation mechanism defined in this paper.

On the other hand, some of the existing TVWS sharing algorithms have implemented the concept of FR. For example, in [22], Bian et al., have implemented the concept of FR in sharing a single TV channel among Cognitive Radios (CR). The CR networks operating in orthogonal frequency division multiple access apply the uplink soft FR concept [23]. Again, the proposed method is defined for CR systems deployed in cellular infrastructure. Similarly, Hessar and Roy [17] have presented an FR method in cellular networks operating in TVWS. Moreover, the algorithm proposed in [17] orthogonalizes WSOs in frequency domain only. None of the existing TVWS sharing algorithms reuses TVWS in a joint, time-frequency domain for WSOs operating in an *ad hoc* coexisting environment. Spectrum reuse in both time and frequency domains shall result in even a better utilization of the available TVWS, as discussed in Section VI-C.

Some genetic algorithms (GA), defined for implementing the channel sharing problem, also exist in the literature. For example, the authors in [24] use a GA-based reliability model to assign channels to mobile hosts based on the reliability of the base station and the channels to enhance the overall reliability of the mobile network system. The results show that this method requires higher number of iterations and generally higher number of available channels than the number of mobile hosts in order to achieve higher reliability. Similarly, Shrestha et. al., proposes a GA-based joint out-of-band spectrum sensing and channel allocation scheme for cognitive radio networks [25]. The joint sensing and resource allocation optimization problem has been formulated using fitness functions of sensing utility and the data transmission utility. Jao and Joe consider a new cognitive radio network model with heterogeneous primary users operating simultaneously via multi-radio access technology [26]. It focuses on energy efficient resource allocation and use a GA-based scheme to obtain an optimal solution in terms of power and bandwidth. The authors in [27] proposed solutions for the problem of efficient resource allocation (radio spectrum and power) in the OFDMA-based multicast wireless system that balances the tradeoff between maximizing the total throughput and ensuring a flexible and controllable spectrum sharing among multicast groups. It proposes two separate optimization methods for subcarriers and power and a GA-based joint optimization scheme is used. Results show that the proposed schemes can attain a high total sum-rate and more flexible and fair distribution of the available bandwidth among multicast groups.

The GA in these and such literature work [28], [29] are well suited for multi-objective optimization problems that require searching over a large space under several constraints. However, GA-based methods are computationally expensive and therefore not suitable for the optimization problem with single objective function and a small search space, like the one defined in this paper. Therefore, GA suffers from the drawbacks of slow convergence speed, and low stability. The channel allocation in highly dynamic spectrum environments requires an algorithm that can do allocation process in a quick runtime. Therefore, rather than applying the GA method, the

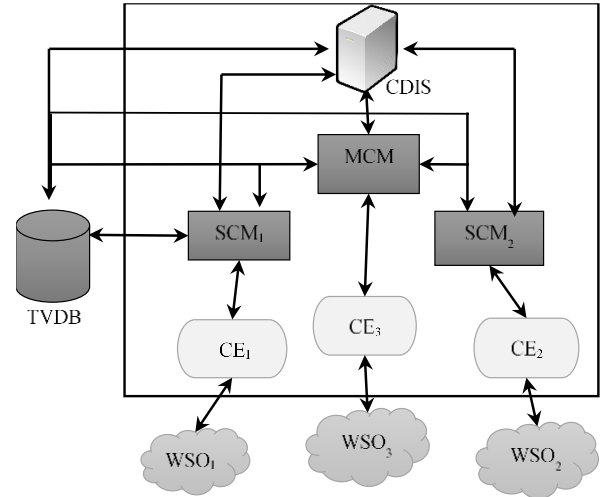


Fig. 1. IEEE 802.19.1 TVWS system architecture. The TVWS database and WSOs interact with the 802.19.1 architecture externally.

nonlinear, binary constrained optimization problem, defined in this paper is transformed into linear optimization problem. Such formulation helps us to apply linear programming solvers to solve the optimization problem and complete the allocation process in a quick, linear runtime.

III. TECHNICAL TERMS AND RESEARCH FOCUS

A. Technical Terms

In this section, we define technical terms that form baseline of the proposed TVWS sharing system, defined in the next section. The proposed system is based on the coexistence system architecture as described in [10] and shown in Fig. 1. The coexistence system in [10] has three logical components: coexistence manager (CM), coexistence enabler (CE), and a coexistence discovery and information server (CDIS).

- The CE registers a WSO to the CM and acts as a communication bridge by translating messages between the WSO and the CM serving the WSO.
- The CM makes coexistence decisions for WSOs registered in it. Moreover, it is required to interact with other CMs, called as neighboring CMs in [10] to resolve coexistence issues among WSOs served by neighboring CMs. In general, it sends configuration commands and control information to the CE.
- The CDIS provides coexistence discovery services like coexistence set information to CMs for registered WSOs.
- The *TVWS database* (TVDB), as shown in Fig. 1, is not part of the coexistence system architecture. It contains information about channels available in the geographic region of each WSO registered with the 802.19.1 system. The TVWS database provides information about the set of TV channels free for whitespace activity to the CMs.

A WSO may register with the IEEE 802.19.1 system before operating in the TV spectrum. In the registration process, a general principle for a WSO to acquire a TV channel is defined in IEEE 802.19.1, summarized as follows. A WSO may perform spectrum sensing to identify and select an available free TV channel or alternatively, it may send a channel

allocation request to its serving CM. If no free channel is available in the geographic region of the WSO, the CM may perform channel sharing among the requesting WSO and the WSOs pre-allocated a TV channel. If such WSOs are registered with other CMs, the CM serving the channel requesting WSO interacts with the other CMs to perform channel sharing. These CMs are called as neighboring CMs to the requesting CM. In this channel sharing procedure, two types of topologies are defined in the 802.19.1 [10]. A distributed CDM topology where neighboring CMs mutually interact to perform channel sharing among WSOs registered within them. A centralized CDM topology where multiple CMs agree to select one of them a master CM (MCM) and rest of the CMs become slave CM (SCM) [10], as shown in Fig. 1. Each SCM provides essential information about operating parameters, including the channel characteristics of each WSO registered within it and its channel demands to the MCM. The MCM performs coexistence services like radio resource allocation to WSOs registered in the SCMs. Some other terms used in the paper are defined as follows.

- A WSO is an entity in 802.19.1 system that represents a TVWS device or network of devices.
- The *channel occupancy* is the duty cycle in a percentage that a network (WSO) occupies a channel [10].
- The *window time* is a slot duration of a scheduling repetition period that satisfies the essential system QoS performance [10].
- The *Coexistence Set (CS)* of a w^{th} WSO is a set of WSOs that are registered in the neighboring CMs that may affect the performance of the w^{th} WSO. In other words, it is a set of WSOs which create interference to the w^{th} WSO.

B. Research Focus

The TVWS sharing problem is defined as, Given a set of available TV channels, a set of CMs with each CM having at least one WSO registered in it and WSOs channel demands, share the TV channels among WSOs such that the following objectives are achieved.

- 1) Maximize the system throughput,
- 2) Minimize unfairness in allocation among WSOs registered in neighboring CMs, and
- 3) Fulfill desired channel demands of the allocated WSOs.

These objectives contradict each other. For example, maximizing the system throughput shall decrease fairness in allocation. Note that from a spectrum allocation perspective, fairness is regarded as equity in access to the resource, the TV spectrum. In other words, being free to use, each network should have an equal opportunity to an access to the given TV spectrum.

Similarly, fulfilling the second and third objectives in conjunction, under the scarcity of the available TVWS, restricts the system accommodating as many as WSOs in the TVWS. Thus, maximizing the fairness while satisfying the channel demands of each allocated WSO is quite complicated in highly congested spectrum environments [30]. Therefore, the fairness in allocation is measured at CM level. The fairness among CMs is deemed at minimum if at least a single WSO in each CM gets the channel.

Considering the above conditions, we design a CDM system, as will be defined in Section IV-A. The system is designed to implement at the MCM in the centralized topology in 802.19.1, as shown in Fig. 1. The system makes use of the information from information messages defined in the 802.19.1 [10] to apply various procedures for defining the proposed TVWS sharing problem as an optimization problem. For example, the WSO registration clause in [10] defines different information acquiring messages that permit a CM to collect desired channel demands, channel statistics, coexistence set elements, available TV channels and related information from WSOs registered within it or with neighboring CMs. Moreover, the inter-CM information sharing messages are also defined in [10]. We assume that using such message templates, the neighboring CMs exchange respective WSOs information with MCM. In order to solve the TVWS sharing problem, the CDM system in MCM then implements a channel allocation process, as will be defined in section V-C. The algorithm makes use of such information available at MCM to implement the subgradient method to solve the TVWS sharing dual problem, Section V-B, to identify a set of WSOs to allocate the TV channels.

The channel allocation process also implements a novel spectrum reuse in Table 3 to have an efficient use of the available TVWS. The spectrum reuse step is also made in compliant with the 802.19.1 by repeated channel allocation using an interference matrix. The CDM defines the interference matrix using the WSOs' CS information available at MCM, as shall be discussed in Section V-D. Note that the CS information is provided by the coexistence discovery algorithm as defined in [10]. The channel allocation process is then executed repeatedly to spatially reuse the TV spectrum to the unallocated WSOs that should not cause interference to pre-allocated WSOs. The proposed channel allocation solution is thus made smoothly integrable to the 802.19.1 system.

IV. SYSTEM DESCRIPTION AND PROBLEM FORMULATION

In the following section, a centralized CDM system is designed that implements a channel allocation process, as shall be discussed in Section V, to implement the TVWS sharing problem defined in Section III-B.

A. System Model

The CDM system is defined as follows,

$$\mathbf{X} = TVWS(\mathcal{C}, \mathcal{J}, \mathcal{Z}, \mathcal{T}, \mathcal{D}). \quad (1)$$

The system parameters are defined as follows. Let c be an index to a set of C neighboring CMs in the system, denoted as \mathcal{C} in Table 1. Let $\mathcal{W}^c, \forall c \in \mathcal{C}$ be a set of network IDs of WSOs registered in the c^{th} CM, as shown in Table 1. Let the network ID, $NID_w \in \mathcal{W}^c$ represents an identifier of the network the w^{th} WSO, registered in c^{th} CM, represents. For example, in the case of IEEE 802.11 type WSO, the NID contains the basic service set identifier used by the WSO.

Let j be an index to the set of all permissible TV whitespace channels, $\mathcal{J} = \{1, 2, \dots, J\}$, where each set element corresponds to a TV channel number, defined on the basis of

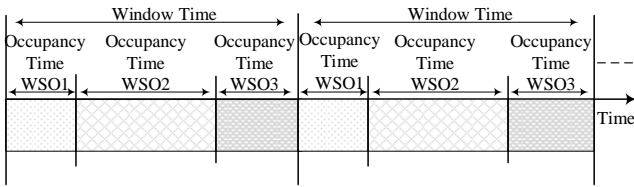


Fig. 2. Scheduling transmission periods for three WSOs on a TV channel.

the regulatory authority rulings. For example, in USA where FCC defines each TV channel to be 6 MHz bandwidth in V/UHF band, therefore, $\mathcal{J} = \{2, 3, \dots, 36, 38, \dots, 51\}$ in the USA. Since, the availability of a TV channel to a w^{th} WSO is a function of geographic location of the WSO and the primary user activity in the region. Therefore, the availability of a TV channel for the secondary use varies spatiotemporally and needs to be determined. We assume that a channel sensing mechanism, as defined in [10] is implemented such that the TVDB contains the set of TV whitespace channels available in the geographic region of each WSO registered in the CMs in the system. Let j be an index to the set \mathcal{J} , then, j^{th} channel availability status to the w^{th} WSO, registered in c^{th} CM, is represented by an indicator function defined as,

$$z_{w,j}^c := \begin{cases} 1, & \text{if } j^{\text{th}} \text{ channel in } \mathcal{J} \text{ is available to } w^{\text{th}} \text{ WSO} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The availability of J channels to the w^{th} WSO, registered in c^{th} CM, are thus represented by a vector of indicator functions defined as,

$$\mathbf{z}_w^c = (z_{w,1}^c, \dots, z_{w,J}^c).$$

The set of channels available to W WSOs registered in c^{th} CM is defined as,

$$\mathbf{Z}^c = (\mathbf{z}_1^c, \mathbf{z}_2^c, \dots, \mathbf{z}_W^c)^T, \forall c \in \mathcal{C}.$$

The system parameter \mathcal{Z} is then defined as follows,

$$\mathcal{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^C\}. \quad (3)$$

The parameter \mathcal{T} in the system in (1) represents the set of window times for the channels in the set \mathcal{J} . In 802.19.1, an algorithm is provided that enables CMs to define the slot duration of the window time. We assume the CMs implement such an algorithm to define the window time, $T_j, \forall j \in \mathcal{J}$, which is then used to define system parameter as,

$$\mathcal{T} = \{T_1, \dots, T_J\}. \quad (4)$$

The system parameter \mathcal{D} in (1) encodes channel demands of CMs, defined as follows. In 802.19.1 [10], a Discovery Information abstraction is provided that allows WSOs to send channel statistics and channel demands like SINR, desired channel occupancy, desired bandwidth etc., to their serving CM [10]. Such information of *heterogeneous*-WSOs is used to define a set of channel demands of w^{th} WSO as follows.

Let $SINR_{w,j}^c$ represents the quality of j^{th} channels to w^{th}

TABLE I
DEFINED PARAMETERS

Input Variables		
Symbol	Description	Value
\mathcal{C}	A set of C CMs in the system.	$\mathcal{C} = \{1, 2, \dots, C\}$
\mathcal{W}^c	A set of NID of W WSOs registered in the c^{th} CM.	$\mathcal{W}^c = \{NID_1, NID_2, \dots, NID_W\}$
\mathcal{J}	A set of permissible TV channels in the system.	$\mathcal{J} = \{1, 2, \dots, J\}$
\mathcal{D}	Channel demands of WSOs, as defined in the system in (1).	-
$O_{w,j}^c$	COT that translates desired occupancy demand of w^{th} WSO on a j^{th} channel.	$\mathbf{O}^c = [O_{w,j}^c]_{1 \times J}, O_{w,j}^c \in \mathbb{R}_{[0, T_j]}$
$I_{w,m}(j)$	Indicator variable encoding m^{th} WSO interference to w^{th} WSO on a j^{th} channel.	$I_{w,m}(j) := \begin{cases} 1 & \text{if } m \text{ interferes } w \\ 0 & \text{otherwise} \end{cases}$
$\mathcal{S}_{w,j}$	Set of WSOs $m \in \mathcal{W}$ such that m^{th} WSO transmission interferes w^{th} WSO transmission on j^{th} channel	$\mathcal{S}_{w,j} = \{m \in \mathcal{W}\}$
$y_{w,j}$	A variable indicating whether m^{th} WSO interferes w^{th} WSO on the j^{th} channel?	$y_{w,j} := \begin{cases} 1 & \exists w \in \mathcal{W} : m \in \mathcal{S}_{w,j} \\ 0 & \text{else} \end{cases}$
$z_{w,j}^c$	An element of the matrix \mathbf{Z} defining accessibility of j^{th} channel to w^{th} WSO.	$\begin{cases} 1 & \text{if } j \text{ accessible to } w^{\text{th}} \text{ WSO} \\ 0 & \text{else} \end{cases}$
Output Variables		
$x_{w,j}^c$	Element of matrix \mathbf{X} defining allocation status of w^{th} WSO on j^{th} channel	$x_{w,j}^c := \begin{cases} 1 & \text{if channel allocated} \\ 0 & \text{otherwise} \end{cases}$

WSOs registered in c^{th} CM. The channel quality is measured in terms of signal to interference and noise ratio (SINR) which depends on interference from primary-to-secondary users and noise floor due to environmental factors. We assume that an interference discovery mechanism is in place that enables each WSO to measure SINR value on each of the channels in \mathcal{J} , as will be further discuss in Section V-D. The quality of all J channels to w^{th} WSO is then defined as,

$$\mathbf{s}_w^c = (SINR_{w,1}^c, SINR_{w,2}^c, \dots, SINR_{w,J}^c), \forall w \in \mathcal{W}^c.$$

Let $p_{w,j}^c$ be the allowed transmission power to w^{th} WSO in the j^{th} channel. The allowed transmission power to w^{th} WSO on J channels is then defined as,

$$\mathbf{p}_w^c = (p_{w,1}^c, \dots, p_{w,J}^c), \forall w \in \mathcal{W}^c.$$

Let B_w^c be the bandwidth demand of w^{th} WSO. The number of channels required by w^{th} WSO is then calculated as,

$$n_w^c = \frac{B_w^c}{b_j}, \forall w \in \mathcal{W}^c, \forall c \in \mathcal{C}$$

where b_j represents the channel bandwidth. Let $O_{w,j}^c$ translates to a timeslot, here called as channel occupancy time (COT) in a window time, such that the w^{th} WSO registered in c^{th} CM can achieve its desired channel occupancy in the allocated j^{th} channel. The relation of COT to a channel window time is shown in Fig. 2 where three WSOs are scheduled in the window time in a single TV channel. The COTs of w^{th} WSO in J TV channels are then represented as,

$$\mathbf{o}_w^{rc} = (O_{w,1}^c, \dots, O_{w,J}^c).$$

The channel demand set of w^{th} WSO is then defined as follows,

$$\{\mathbf{s}_w^{rc}, \mathbf{p}_w^{rc}, n_w^c, \mathbf{o}_w^{rc}\}, \forall c \in \mathcal{C}, \forall w \in \mathcal{W}^c \quad (5)$$

The channel demand set of c^{th} CM is then defined using channel demands of its registered WSOs as follows,

$$\mathcal{D}^c = \{\mathbf{s}^c, \mathbf{p}^c, N^c, \mathbf{o}^c\}, \forall c \in \mathcal{C}, \quad (6)$$

where $\mathbf{s}^c = (\mathbf{s}_1^c, \dots, \mathbf{s}_W^c)$, $\mathbf{p}^c = (\mathbf{p}_1^c, \dots, \mathbf{p}_W^c)$, $N^c = (n_1^c, \dots, n_W^c)$ and $\mathbf{o}^c = (\mathbf{o}_1^c, \dots, \mathbf{o}_W^c)$. Let $\mathbf{S} = (\mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^C)^T$, $\mathbf{P} = (\mathbf{p}^1, \mathbf{p}^2, \dots, \mathbf{p}^C)^T$, $\mathbf{N} = (N^1, N^2, \dots, N^C)^T$, $\mathbf{O} = (\mathbf{o}^1, \mathbf{o}^2, \dots, \mathbf{o}^C)^T$, the system parameter \mathcal{D} is then defined using the channel demands of all neighboring CMs as follows,

$$\mathcal{D} = \{\mathbf{S}, \mathbf{P}, \mathbf{N}, \mathbf{O}\}. \quad (7)$$

The system in (1) then executes the channel allocation algorithm, as will be discussed in Section V, to allocate TV channels to the WSOs registered in the neighboring CMs such that the allocation satisfies the required system QoS performance. The system QoS performance is preserved if the following allocation condition is satisfied,

$$\sum_{c \in \mathcal{C}} \sum_{w \in \mathcal{W}^c} O_{w,j}^c \leq T_j, \forall j \in \mathcal{J} \quad (8)$$

where T_j refers to the window time in a j^{th} channel. The algorithm proposed in Section V solves the TVWS sharing optimization problem, as will be defined in (14) and outputs a channel allocation matrix \mathbf{X} , defined as follows. Let $x_{w,j}^c \in \{0,1\}$, be a binary decision variable such that if $x_{w,j}^c = 1$, the j^{th} channel is allocated to the w^{th} WSO registered in c^{th} CM; otherwise $x_{w,j}^c = 0$. The allocation status of WSOs registered in the neighboring CMs is then represented by a matrix \mathbf{X} as,

$$\mathbf{X} := \begin{bmatrix} x_{1,1}^1 & x_{1,2}^1 & \dots & x_{1,J}^1 \\ & \vdots & & \\ x_{W^1,1}^1 & x_{W^1,2}^1 & \dots & x_{W^1,J}^1 \\ x_{1,1}^2 & x_{1,2}^2 & \dots & x_{1,J}^2 \\ & \vdots & & \\ x_{W^c,1}^c & x_{W^c,2}^c & \dots & x_{W^c,J}^c \end{bmatrix}, \quad (9)$$

where $W^c = |\mathcal{W}^c|$, $\forall c \in \mathcal{C}$, i.e., the number of WSOs registered in the c^{th} CM. The w^{th} row in the \mathbf{X} represents the channels allocation status, in the set \mathcal{J} , to the w^{th} WSO registered in c^{th} CM. The j^{th} column in the \mathbf{X} represents the channels allocation status of all the WSOs, from all the CMs in the set \mathcal{C} . The

allocation matrix \mathbf{X} thus orthogonalizes WSOs, registered in the neighboring CMs, in a joint frequency-time domain. The WSOs scheduled on different channels can transmit at the same time using their respective allotted channel (frequency slot) while WSOs scheduled on the same channel can transmit in their respective time slot (here COT).

The system in (1) thus, implements the TVWS sharing problem, defined in Section III-B, as an optimization problem, as discussed in the following section.

B. Problem Formulation

In this section, the proposed TVWS sharing problem is formulated as an optimization problem using well-established proportional fairness method. It is because the proportional fairness is considered one of the most suitable methods to achieve a trade-off between two competing interests [31], [32], [33]. Originally, Kelly defined the proportional fairness as an adjustment process which adjusts the rates of users according to the charges they pay. The proportional fairness method thus was defined for elastic traffic in computer network services [34]. Similarly, in the channel sharing literature, a proportionally fair allocation mostly has been achieved by adjusting the rates of the users based upon some performance criteria like maximizing the resource utilization, etc. [35], [36]. However, applying the proportional fairness in its original to model the TVWS sharing problem proposed in this paper is not suitable. It is because, the third objective in the problem defined in Section III-B makes the resource allocation as binary decision allocation, i.e., a channel is either allocated to a WSO, $x_{w,j}^c = 1$ or not $x_{w,j}^c = 0$. Therefore, WSO allocation (here COT) adjustment is not possible. Consequently, we rewrite the proportional fairness in a binary decision allocation perspective as follows.

Let the maximum data rate the w^{th} WSO can achieve on j^{th} channel be defined by using Shannon channel capacity formula,

$$r_{w,j}^c = b_j \log(1 + SINR_{w,j}^c). \quad (10)$$

The maximum rate $r_{w,j}^c$, $\forall w \in \mathcal{W}^c$ is then used to defined a utility function as a normalized rate achieved by c^{th} CM in j^{th} channel as follows,

$$\mathcal{U}_{c,j} = \sum_{w \in \mathcal{W}^c} \frac{x_{w,j}^c r_{w,j}^c}{O_{w,j}^c + \delta_{O_{w,j}^c, 0}} \quad (11)$$

where $\delta_{O_{w,j}^c, 0}$ defines Kronecker delta function as:

$$\delta_{O_{w,j}^c, 0} := \begin{cases} 1, & \text{if } O_{w,j}^c = 0, \\ 0, & \text{otherwise.} \end{cases}$$

This function prevents denominator term in (11) from becoming zero. The utility function in (11) measures the worth of the resource (channel) to c^{th} CM, i.e., given a channel is allocated to the WSOs in the c^{th} CM for the duration of $\sum_{w \in \mathcal{W}^c} O_{w,j}^c$, how does it translate for the CM in terms of the achieved throughput. In other words, maximizing the function in

(11) shall prefer a CM with WSOs achieving high data rate and lower channel occupancy demand over a CM with WSOs achieving low data rate and high channel occupancy demand. Such preference based allocation shall lead to an efficient use of the resources (TVWS). The distribution $\mathbf{U} = [\mathcal{U}_{c,j}]_{\mathcal{C} \times \mathcal{J}}$ is then said to be proportionally fair if it is feasible and for all other feasible solutions $\mathbf{V} = [v_{c,j}]_{\mathcal{C} \times \mathcal{J}}$, the following holds [34],

$$\sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} \frac{v_{c,j} - \mathcal{U}_{c,j}}{\mathcal{U}_{c,j}} \leq 0. \quad (12)$$

It has been shown in [34], [37] that the rates achieved by users become proportionally fair if the sum of logarithmic rates obtained is optimized. Moreover, it is shown in [38] that if all rates are proportionally fair, they maximize the throughput over all other feasible throughputs. Therefore, if the logarithmic sum of the utility function in (11) is maximized, the normalized rate achieved by neighboring CMs shall become proportionally fair. Let a j^{th} channel is said to be allocated to the c^{th} CM if at least one of its registered WSO is scheduled on the channel. The allocation status of the channels in the \mathcal{J} , to the c^{th} CM, is then defined as follows,

$$\mathbf{x}^c := \begin{bmatrix} x_{1,1}^c & x_{1,2}^c & \cdots & x_{1,J}^c \\ \vdots & \vdots & \ddots & \vdots \\ x_{W^c,1}^c & x_{W^c,2}^c & \cdots & x_{W^c,J}^c \end{bmatrix}. \quad (13)$$

Let $\mathbf{1} = (1, 1, \dots, 1)_{1 \times J}$. Let $\mathbf{O}_j \in \mathbf{O}$ be the j^{th} column vector in COT demand matrix in the system parameter \mathcal{D} , defined as, $\mathbf{O}_j = (O_{1,j}^1, O_{2,j}^1, \dots, O_{W^1,j}^1, O_{1,j}^2, \dots, O_{W^c,j}^c)^T$ where $W^c = |\mathcal{W}^c|, \forall c \in \mathcal{C}$. Let $\mathbf{X}_j \in \mathbf{X}$ represents the j^{th} column vector of the allocation matrix \mathbf{X} . The TVWS sharing problem is then defined as follows,

$$\max. \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} \log(\mathcal{U}_{c,j} + 1) \quad (14a)$$

$$\text{subject to } \mathbf{x}^c \leq \mathbf{Z}^c, \quad \forall c \in \mathcal{C}, \quad (14b)$$

$$\mathbf{X}_j^T \mathbf{O}_j \leq T_j, \quad \forall j \in \mathcal{J}, \forall c \in \mathcal{C}, \quad (14c)$$

$$\mathbf{x}^c \mathbf{1}^T \leq (N^c)^T, \quad \forall c \in \mathcal{C}, \quad (14d)$$

$$\mathbf{x}^c \in \{0, 1\}, \quad \forall c \in \mathcal{C}. \quad (14e)$$

The constraint in (14b) ensures that a channel can be allocated to the WSOs registered in c^{th} CM only if the channel is available in their respective region, i.e., $x_{w,j}^c \in \mathbf{x}^c = 1$ iff $z_{w,j}^c \in \mathbf{Z}^c = 1$. The constraint in (14c) ensures that the WSOs scheduled in a j^{th} channel preserve the system QoS performance, as defined in (8), i.e., the total allocated channel occupancy time of coexisting WSOs must preserve the channel window time. The constraint in (14d) ensures that the number of channels allocated to the c^{th} CM is restricted by the number of channels desired by its WSOs. Finally, (14e) forces the decision variable to be binary valued.

The constraints in (14e) and (14c) helps the system in (1) to satisfy the third objective of TVWS sharing problem in Section III-B. The optimization problem in (14) seeks to optimize a concave objective function over a convex set. The problem in (14) has a unique solution, as from the optimization theory [39], maximizing a concave function over a convex set has a unique solution. A solution approach to the problem in (14) is presented in the following section.

V. SOLUTION METHOD

The nonlinear objective function (14a) and binary-valued constraint (14b) makes the problem in (14) a nonlinear combinatorial optimization problem. Determining the optimal solution of such a problem is a challenging task as the problem becomes intractable as the number of discrete variables increases [40]. Therefore, to ease the solution approach, the problem in (14) is transformed into a linear programming problem with relaxed binary constraint.

A. Linearization

The objective function (14a) is linearized using a piecewise linear approximation. In this process, tangent line approximation is used to approximate the objective function in (14a), denoted as, F . The detailed description of linear approximation is provided in Appendix A. Using this function, the problem in (14) is linearized as,

$$\max. \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} F(\mathcal{U}_{c,j}) \quad (15a)$$

$$\text{subject to } \mathbf{x}^c \leq \mathbf{Z}^c, \quad \forall c \in \mathcal{C} \quad (15b)$$

$$\mathbf{X}_j^T \mathbf{O}_j \leq T_j, \quad \forall j \in \mathcal{J}, \forall c \in \mathcal{C} \quad (15c)$$

$$\mathbf{x}^c \mathbf{1}^T \leq (N^c)^T, \quad \forall c \in \mathcal{C} \quad (15d)$$

$$\mathbf{x}^c \in \{0, 1\}, \quad \forall c \in \mathcal{C} \quad (15e)$$

To tackle the binary-valued constraint (15b), we apply Lagrangian relaxation as explained followings.

B. Lagrangian Relaxation

Lagrangian relaxation [41] relaxes a subset of constraints by adding them to the objective function with a penalty term called the Lagrangian multiplier. Let $\boldsymbol{\lambda} := [\lambda_{w,j}]_{W \times J}$ be the Lagrangian multipliers matrix. Then, the relaxed problem can be defined as,

$$\max_{\mathbf{X}}. P(\mathbf{X}, \boldsymbol{\lambda}) = \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} F(\mathcal{U}_{c,j}) + \boldsymbol{\lambda}^T (\mathbf{Z}^c - \mathbf{x}^c) \quad (16a)$$

$$\text{subject to: } \mathbf{X}_j^T \mathbf{O}_j \leq T_j, \quad \forall j \in \mathcal{J}, \forall c \in \mathcal{C} \quad (16b)$$

$$\mathbf{x}^c \mathbf{1}^T \leq (N^c)^T, \quad \forall c \in \mathcal{C} \quad (16c)$$

$$\mathbf{x}^c \in \{0, 1\}, \quad \forall c \in \mathcal{C} \quad (16d)$$

For a given $\boldsymbol{\lambda}$, the Lagrangian relaxation can be defined as,

$$h(\boldsymbol{\lambda}) = \max_{\mathbf{X}} \{P(\mathbf{X}, \boldsymbol{\lambda}) : \text{constraints (16b), (16c), (16d)}\} \quad (17)$$

Then the generalized dual problem of the relaxed problem is defined as followings,

TABLE 2
ALGORITHM: DUAL PROBLEM BASED ON LAGRANGIAN RELAXATION

Step 0:	a) Choose initial values of λ^0 .
	b) Set parameters, for example, $\rho = 2.0$, $\rho^{\min} = 0.001$, $\rho_{iter} = 0$, $\rho^{\max iter} = 5$, $k = 0$, $k^{\max} = 10$, $F^{\text{best}} = 0$, $h^{\text{best}} = -\infty$, $h^{\text{upper}} = 0$, $\mathbf{X}'_k = [0]_{W \times J}$.
Step 1:	a) Increment as $k = k + 1$, $\rho_{iter} = \rho_{iter} + 1$
	b) Given λ^k , solve the relaxed problem using any linear programming technique and obtain \mathbf{X}_k .
Step 2:	Validate \mathbf{X}_k as: set $x_{w,j}^c := 0$ if $z_{w,j}^c = 0$.
Step 3:	Perform frequency reuse as in Table 3 and get \mathbf{X}'_k .
Step 4:	Use \mathbf{X}'_k to compute the value of the function in (16a), called as F , and fairness index value H in (20). If $F > F^{\text{best}}$: $F^{\text{best}} = F$, $h^{\text{upper}} = F^{\text{best}}$ and $\mathbf{X} = \mathbf{X}'_k$.
Step 5:	a) Use \mathbf{X}'_k to compute: <ul style="list-style-type: none"> - Subgradient vector as, $\nabla h(\lambda^k) = \left[\frac{\partial h}{\partial \lambda_{w,j}^k}, \forall w \right]$, - Dual objective in (18), - Step size as, $t_k = \frac{\rho(h^{\text{upper}} - h(\lambda^k))}{\ \nabla h(\lambda^k)\ ^2}$.
	b) Update the dual variable as, $\lambda^{k+1} = \max\{\lambda^k + t_k \nabla h(\lambda^k), 0\}$
Step 6:	If $h^{\text{best}} < h(\lambda^k)$ then $h^{\text{best}} = h(\lambda^k)$ else if $\rho_{iter} > \rho^{\max iter}$ then $\rho = \max\{\frac{\rho}{2}, \rho^{\min}\}$ and $\rho_{iter} = 0$.
Step 7:	If $t_k < 0.001$ or $k > k^{\max}$ stop; otherwise, go to Step 1.

$$L^* = \min_{\lambda} \{h(\lambda) : \lambda \geq 0\}. \quad (18)$$

The solution to (17) is the upper bound of the solution to the original problem (16). Note that (17) is a concave function. For a concave function, a gradient-based approach is generally used to compute a value as close as desired to the optimal value. Thus, if h would have been differentiable, we can use a gradient descent method to have a convergence toward the optimal value. The proposed problem, however, cannot be solved using a gradient descent method. It is because the objective function is piecewise linear which is non-differentiable at the intersection point of adjacent linear pieces, but sub-differentiable at this point. The subdifferential of $h(\lambda)$ at such a point is the set of all subgradients at that point. Thus, we need to compute a sequence of $\{\lambda^k\}_{k \in \mathbb{N}}$ such that either $h(\lambda^k)$ converges to the optimal solution using the subgradient method, which is given in the following dual algorithm. The convergence property of the subgradient algorithm is presented in Appendix B.

C. Subgradient Algorithm for Lagrangian Relaxation based TVWS Sharing Problem

The algorithm defined in Table 2, can be described as follows. In Step 0, the input parameters to the algorithm are defined as follows. The initial values of λ^0 are defined randomly. The parameter ρ is used in defining step size t^k ,

defined in the range $\rho^{\min} < \rho \leq 2$ [41]. The ρ_{iter} with upper limit of $\rho^{\max iter}$ counts the number of iterations after which the parameter ρ is updated. The k^{\max} is defined as stopping criteria for the algorithm.

The algorithm uses variables initialized in Step 0 to apply a linear programming (LP) solver to solve the dual problem and obtain the k^{th} iteration allocation matrix \mathbf{X}_k . LP solvers are available on both the commercial and freeware basis. The entries in \mathbf{X}_k are then adjusted based upon the corresponding entries in \mathbf{Z}^c such that $x_{w,j}^c \in \mathbf{x}_k^c$, $\forall \mathbf{x}_k^c \in \mathbf{X}_k$ are set equal to zero if the corresponding element, $z_{w,j}^c \in \mathbf{z}_w^c$, $\forall \mathbf{z}_w^c \in \mathbf{Z}^c$ is zero. This validation ensures the constraint in (14b).

The algorithm then applies the FR process in Step 3 in Table 2. In this process, the algorithm makes use of the current allocation vector, \mathbf{X}_k and interference matrix, as shall be discussed in Section IV-D, to identify a set of WSOs which do not get the channel. The algorithm then repeatedly applies LP solver to performs channel allocation to the unallocated WSOs such that they do not cause interference to the allocated WSOs of neighboring CMs. The FR process is detailed in Section V-E. The outcome of FR process is an updated allocation matrix \mathbf{X}'_k which is then used to compute the function values in (16a) and the fairness in allocation among neighboring CMs.

Several fairness measures or metrics are used in the literature to determine whether networks are receiving a fair share of spectrum or not. For example, max-min fairness, Jain's fairness index, fairly shared spectrum efficiency, worst-case fairness. In this paper, we adopt Jain's fairness index [42] to measure fairness in allocation among neighboring CMs. The reason is that it satisfies the desired properties of fairness measure like population size independence, continuity etc., as listed in [43]. These properties are important to be considered in measuring the fairness in allocation. For example, the continuity property shows any slight change in the allocation of individual WSO. Thus, an inefficient use of the TVWS is identified by the fairness index as a WSO with bad channel characteristics gets a high proportion of the spectrum. It is ensured through the use of the continuous allocation metric like fraction of throughput demand, as defined in (19). Such an allocation metric is suitable to measure the fairness in allocation for the case where WSOs demand unequal channel bandwidth [43]. Therefore, based on the fraction of throughput demand of CMs, an allocation metric is defined as follows,

$$T^c = \frac{d^c}{d'^c}, \forall c \in \mathcal{C}, \quad (19)$$

where d^c and d'^c represents the maximum data the c^{th} CM desire to transmit and it can transmit using its allocated channels, respectively. These terms are defined as follows. Let the maximum data the c^{th} CM can transmit using its allocated channels is defined in terms of the data the WSOs registered in it can transmit, defined as follows.

$$d^c = \sum_{j \in \mathcal{J}} \sum_{w \in \mathcal{W}^c} x_{w,j}^c O_{w,j}^c r_{w,j}^c, \forall c \in \mathcal{C}. \quad (20)$$

Note that channels are considered as additive white Gaussian noise (AWGN). The data the CM desires to transmit is defined as,

$$d^{lc} = \sum_{j \in \mathcal{J}} \sum_{w \in \mathcal{W}} O_{w,j}^c \bar{r}_{w,j}^c, \forall c \in \mathcal{C}.$$

The normalized throughput vector (T^1, \dots, T^C) is then adopted to measure fairness in allocation using Jain's fairness index [42] as,

$$H(T^1, T^2, \dots, T^C) = \frac{\left(\sum_{c \in \mathcal{C}} T^c \right)^2}{C \sum_{c \in \mathcal{C}} (T^c)^2}. \quad (21)$$

Function H in (21) outputs a value in the range of $[0, 1]$; when the value is closer to 1, the allocation is deemed fairer.

If the current iteration value of the objective function, F , is optimal, then F^{best} is updated with F and \mathbf{X} with \mathbf{X}'_k . As the iteration progresses, the feasible primal F^{best} and lower bound h^{best} approach gradually to the integer optimal by adjusting λ^k using the subgradient method as defined in Step 5. In Step 5, the sub-gradient vector of the objective function and the Lagrangian multiplier vector λ^k for the k^{th} iteration are calculated. The step size t_k is used to calculate the multiplier vector for the next iteration. The Lagrange multipliers are thus adjusted iteratively. The convergence property of the subgradient algorithm is discussed under Appendix B. The algorithm terminates as one of the termination conditions satisfied:

- Dual step size becomes less than a set threshold or,
- the number of iterations exceeds the maximum number of iterations.

After the overall iteration ends, we regard the final value of F^{best} as the approximated optimal solution and the corresponding allocation matrix \mathbf{X} is the algorithm output.

The interference matrix, \mathbf{Y} , that is used to implement the FR step in Table 3 is defined in the following section.

D. Interference Matrix

The WSOs registered in the neighboring CMs and interfering on the available TV channels is represented using an interfering matrix called as Y-matrix in this paper. Note that the Y-matrix does not model the interference among coexisting WSOs. Rather, it represents the set of WSOs which cannot transmit simultaneously on the available TVWS due to interfering transmission regions. In fact, in IEEE 802.19.1 [10], a coexistence discovery algorithm is presented that the CDIS and CM run to perform the statistical analysis of the expected interference among coexisting WSOs. Briefly, the algorithm in [10] takes the WSOs' geographic location, transmitter and receiver characteristics, antenna height and directivity, height above average terrain and other related parameters to execute interference discovery process. In this process, a cumulative distribution function of the potential interference from m^{th} WSO to w^{th} WSO is estimated. Both of these, m^{th} and w^{th} WSOs, could register to the same CM or different CMs in the system. The minimum interference level, experienced by 90%

devices of the w^{th} WSO, is then taken as the potential interference value from an m^{th} WSO to w^{th} WSO. The measured interference value is then compared to a threshold. If the value is greater than the threshold, the m^{th} WSO is considered potential interferer to the w^{th} WSO and is included in its CS. A similar rule is applied for interference discovery of the w^{th} WSO into the m^{th} WSO. Thus, the outcome of the interference analysis process is a CS of each WSO registered in the CMs in the system. The system in (1) then makes use of the CS of each WSO to generate a Y-matrix as follows.

Let a set $\mathcal{S}_{w,j} = \{I_{w,m}(j)\}, \forall m \in \mathcal{W}$, be an encoded CS of w^{th} WSO on a j^{th} channel such that an indicator variable $I_{w,m}(j) = 1$ if m^{th} WSO interferes w^{th} WSO transmission on the j^{th} channel, as defined in Table 1; otherwise $I_{w,m}(j) = 0$. The encoded CS of all the WSOs coexisting on j^{th} channel are then used to define a j^{th} channel interference matrix $\mathbf{y}(j)$ as follows,

$$\mathbf{y}(j) := \begin{bmatrix} \times & I_{1,2}(j) & \cdots & I_{1,W}(j) \\ & & & \vdots \\ I_{W,1}(j) & I_{W,2}(j) & \cdots & \times \end{bmatrix} \quad (22)$$

where \times in diagonal vector in $\mathbf{y}(j)$ represents don't care condition. This condition translate a self-interference indicator variable, $I_{w,w}(j)$, having no meaning. The w^{th} row in $\mathbf{y}(j)$ matrix represents encoded CS of w^{th} WSO. The interference matrices for all channels in the system are then used to define an interference matrix \mathbf{Y} as follows,

$$\mathbf{Y} = [\mathbf{y}(1) \ \mathbf{y}(2) \ \cdots \ \mathbf{y}(J)] \quad (23)$$

The TVWS sharing algorithm, in Table 2 makes use of the interference matrix \mathbf{Y} to implement FR in sharing TVWS among *heterogeneous* WSOs, as discussed in the following subsection.

E. Frequency Reuse

The frequency reuse (FR) subroutine in Table 3 performs spatial reuse of the TV spectrum to enhance its effective utilization. The FR process is implemented to the WSOs do not getting channel in the initial allocation phase in Step 1, Table 2. This requires to identify a set of unallocated WSOs eligible for the FR. In this process, an encoded CS $\mathcal{S}_{w,j}, \forall m \in \mathcal{W}$ and an interference matrix \mathbf{Y} are used to define the set of unallocated WSOs, \mathcal{W}' . To generate encoded CS and Y-matrix, we make use of the CS of each WSO available at MCM. Note that the 802.19.1 defines different message clauses that enable CMs to exchange their WSO related information [10]. Let us assume the CS of WSOs are available to CDM at MCM. Given such information available, an encoded CS of WSOs, $\mathcal{S}_{w,j}, \forall m \in \mathcal{W}$ and an interference matrix \mathbf{Y} , are generated, as defined in Section V-D. Initially the Y-matrix is filled with all ones. Let \mathbf{X}_k be an initial allocation matrix available from Step 2, Table 2. The Y-matrix is then updated based on the \mathbf{X}_k and $\mathcal{S}_{w,j}, \forall m \in \mathcal{W}$ in

TABLE 3
SUBROUTINE: FREQUENCY REUSE

Input:	$\lambda^k, \mathbf{X}'_k = \mathbf{X}_k, \mathcal{Z}, \text{CS}$.
Output:	\mathbf{X}'_k
Step 0:	Given CS generate encoded CS, i.e., $\mathcal{S}_{w,j}, \forall w \in \mathcal{W}, \forall j \in \mathcal{J}$ and interference matrix \mathbf{Y} , as defined in Section V-D.
Step 1:	Given \mathbf{X}'_k , update $\mathbf{y}(j) \in \mathbf{Y}, \forall j \in \mathcal{J}$ as: For each w^{th} WSO do: if $x_{w,j} = 1: I_{w,m}(j) = 0, \forall m \in \mathcal{W}, \forall j \in \mathcal{J}$ or if $(x_{m,j} = 1 \text{ and } w \in \mathcal{S}_{m,j}): I_{w,m}(j) = 0, \forall m \in \mathcal{W}$.
Step 2:	Define unallocated WSO set in the system as, $\mathcal{W}' = \left\{ \forall w \in \mathcal{W} : \exists j \in \mathcal{J} \mid \sum_{m \in \mathcal{W}} I_{w,m}(j) > 0 \right\}$.
Step 3:	While $\sum_{w \in \mathcal{W}} \sum_{m \in \mathcal{W}} I_{w,m}(j) > 0, \forall j \in \mathcal{J}$ and $\mathcal{W}' \neq \{\}$ do a) Given λ^k , and \mathcal{W}' ; solve the relaxed problem using any linear programming solver and obtain \mathbf{X}_k . b) Perform following updates: 1) Update \mathbf{X}_k as, $x_{w,j}^c := 0$ if $z_{w,j}^c = 0$. 2) Update \mathbf{X}'_k as, $\mathbf{X}'_k = \mathbf{X}'_k + \mathbf{X}_k$. 3) Update \mathcal{W}' as, $\mathcal{W}' \leftarrow \mathcal{W}' \setminus \{ \forall w \in \mathcal{W}' \mid \exists j \in \mathcal{J} : x_{w,j}^c = 1 \}$. 4) Update \mathbf{Y} as in Step 1.

Step 1, Table 3, as follows. For each j^{th} channel in the system, update interference matrix $\mathbf{y}(j) \in \mathbf{Y}$ as,

- 1) If j^{th} channel is allocated to w^{th} WSO, set all w^{th} row elements in $\mathbf{y}, \forall \mathbf{y} \in \mathbf{Y}$ equal to zero, or
- 2) If j^{th} channel is allocated to m^{th} WSO and w^{th} WSO is in the CS of m^{th} WSO, set all w^{th} row elements in the matrix \mathbf{y} equal to zero.

The above two steps identify the eligibility of the WSOs for implementing the FR process. For example, if the w^{th} WSO is already allocated a channel, we aim to restrict it in taking part the FR process. Therefore, the w^{th} row entries in the entire Y-matrix are flipped zero in the first step above. Similarly, if a j^{th} channel is already allocated to m^{th} WSO and if w^{th} WSO transmission in the j^{th} channel shall create harmful interference to the m^{th} WSO transmission, the j^{th} channel cannot be spatially reused at unallocated w^{th} WSO. Therefore, Y-matrix entries corresponding to w^{th} row are also flipped zero. The updated Y-matrix thus defines a set of unallocated WSOs. These are the WSOs for which at least one nonzero entry exists in the corresponding row in the Y-matrix, as defined, in Step 2, Table 3.

The subroutine in Step 3, Table 3 then repeatedly allocates the available TV channels to the WSOs in the set \mathcal{W}' as follows. The relaxed problem in (17) is solved using any LP solver for the WSOs in the set \mathcal{W}' and an allocation matrix \mathbf{X}_k is obtained. The \mathbf{X}_k is then used to update $\mathbf{X}'_k, \mathcal{W}'$, and Y-matrix, as defined in Step 3-b)2), 3-b)3), and 3-b)4), respectively. This repetitive update and allocation process continues until all WSOs in the set \mathcal{W}' get the channel or no more FR is possible.

Let us apply the FR implementation in the coexisting scenario shown in Fig 3. In this figure, four WSOs operating in three network technologies, an IEEE 802.22 regional area

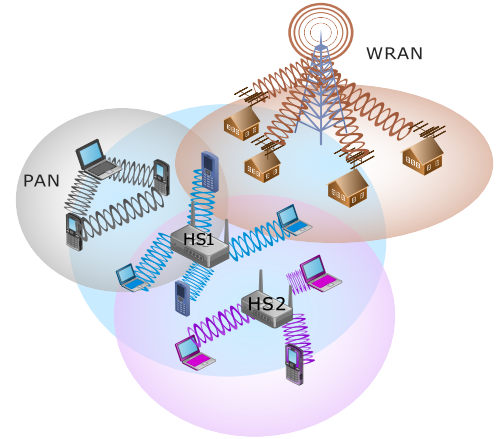


Fig. 3. IEEE 802.22 wireless regional area network (WRAN), IEEE 802.11 hotspots (HS1, HS2), and IEEE 802.15.4 personal area network (PAN) coexisting in some geographic region.

network, IEEE 802.11 local area networks and IEEE 802.15.4 personal area network are deployed in some geographic region. The shaded area around each transmitter denotes its transmission radius. The circular links between a transmitter and receivers show wireless connectivity between them. The receiver nodes in some networks receive interfering signals from other collocated transmitters as shown in the figure. Let WRAN, HS1, HS2, and PAN are labelled as, WSO 1, 2, 3 and 4, respectively. Let us assume each of the WSO is registered in a dedicated CM, i.e., four neighboring CMs are available in the CDM system. Let us suppose that a single TV channel is available in the region for secondary use. Then, based on coexisting scenario shown in the figure, the encoded CS of each WSO can be defined as follows.

$$\mathcal{S}_{1,1} = \{0, 1, 0, 0\}, \mathcal{S}_{2,1} = \{1, 0, 1, 1\}, \mathcal{S}_{3,1} = \{0, 1, 0, 0\}, \mathcal{S}_{4,1} = \{0, 1, 0, 0\}.$$

The Y-matrix is then populated from the bitwise OR operation on the CS of the WSOs. The generated Y-matrix is $\mathbf{Y} = [1 \ 1 \ 1 \ 1]$. Let for some given input parameters, as listed in Table 1, the algorithm in Table 2 finds an initial allocation vector, $\mathbf{X} = [1 \ 0 \ 1 \ 0]$. The allocation vector shows WSO 1 and WSO 3 are allocated the channel. The FR process is then invoked. The Y-matrix is updated to identify WSOs eligible for spatially reusing the channel, as follows. The XOR operation is performed as, $(\mathbf{Y} = \mathbf{X} \oplus \mathbf{Y})$. This operation turns the entries in Y-matrix equal to zero where the corresponding entries in X-matrix are ones. The Y-matrix at this stage looks like, $\mathbf{Y} = [0 \ 1 \ 0 \ 1]$. It is then updated using the CS of allotted WSOs as previously defined in the second rule of Y-matrix update. The second entry in Y-matrix is thus flipped zero as WSO 2 is in the CS of allotted WSO 1. The updated Y-matrix then looks like, $\mathbf{Y} = [0 \ 0 \ 0 \ 1]$. The algorithm then solves the dual problem again and allocates the channel to WSO 4. The final allocation matrix then looks like $\mathbf{X} = [1 \ 0 \ 1 \ 1]$. The final allocation shows that the available TV channel is reused at WSO 4 without causing harmful interference to allotted WSO 1 and WSO 3.

F. Scheduling Map

Once the allocation process in Table 2 and frequency reuse in Table 3 terminates, the CDM system generates a scheduling map to send it to the CMs in the system. The scheduling map (SM) is a map showing the WSOs' scheduling periods arranged in window time in the allocated channels. In this paper the

scheduling period of a w^{th} WSO refers to its channel timeslot, i.e., COT. For example, SM of three WSOs scheduled in an allocated TV channel is shown in terms of their COT defined in the window time in Fig. 2. Thus, given the COT of WSOs and the allocation matrix \mathbf{X} , from the algorithm in Table 2, the SM is a simple procedure of defining two timing parameters; transmission start time and transmission end time. The CDM system defines the timing parameters for WSOs registered in the CMs in the system as follows.

Let a pair of transmission variables, $(t_{w,j}^{\text{start}}, t_{w,j}^{\text{stop}})$, precisely define the time instance the w^{th} WSO, registered in c^{th} CM, may start and stop its transmission on an allotted j^{th} channel, respectively. The $t_{w,j}^{\text{start}}$ and $t_{w,j}^{\text{stop}}$ are calculated as follows. Let a variable $C_{w,m(w)}$ be defined as the cost of sharing a channel between two WSOs, $w, m \in \mathcal{W}$, where $m(w)$ represents a WSO m sharing a channel with WSO w . Let τ_w represents the control overhead associated with MAC technology of the w^{th} WSO. The control overhead is defined as the amount of time required to perform control signaling while operating in the TVWS. This value is fixed and predetermined based upon the underlying network technology of the WSO. For example, if an 802.22 WSO employs OFDMA, one OFDM symbol is used for both the frame preamble and the frame header; except for the first frame in the superframe which consumes two additional symbols (1/4 cyclic prefix mode). If we consider two OFDM symbols per frame as a control region then using a symbol duration, $T_{\text{Sym}}=0.3733$ ms [44], the control overhead per frame is computed as, 0.7466 ms. Other settings may generate different overhead. Similarly, if a WSO m operates in a different network technology than that of the WSO w , its control overhead will be different from that of WSO w . The total overhead in a channel varies as the channel is shared among heterogeneous WSOs. The value of the parameter $C_{w,m(w)}$ is then defined simply by adding the control overhead of all WSOs sharing a channel as follows:

$$C_{w,m(w)} := \begin{cases} \tau_w + \tau_m & \text{if } MAC_w \neq MAC_m, \forall (w, m) \in \mathcal{W}^c \\ 0 & \text{otherwise} \end{cases}, \quad (24)$$

where $\mathcal{W}^c \subset \mathcal{W}^x$ refers to the set of WSOs with NID listed before NID of w^{th} WSO in \mathcal{W}^c . The timing parameters are computed as,

$$t_w^{\text{start}} = \sum_{m \in \mathcal{W}^c} O_{m,j}^c x_{m,j}^c + C_{w,m(w)} \quad \text{and} \quad t_w^{\text{stop}} = t_w^{\text{start}} + O_{w,j}^c. \quad (25)$$

Thus, the $t_{w,j}^{\text{start}}$ refers to the time instance in the scheduling window that all the WSOs m have utilized the channel for the duration of their respective COT. Note that in defining the scheduling map we make a simplifying assumption that the timers of WSOs in the system are pre-synchronized and WSOs sharing a j^{th} channel have agreed on the reference time (the time instance the window time starts) as defined in [10]. Timer synchronization may be done by having agreements between service providers managing the WSOs which is outside the scope of this paper.

The CDM defines SM and send it to the SCMs. The SCMs send the SM to the registered WSOs. Such implementation

shall reduce the control signaling between the WSOs and the pertinent CM. The control signaling is otherwise inevitable while performing context switching among WSOs scheduled in the TV channel. Once the spectrum has been allocated, the SM remains unchanged unless i) an incumbent appears in one of the assigned channels ii) a change in a WSO's channel occupancy demand or some other coexisting WSO's demand requires readjusting the WSO's allocation.

VI. SIMULATIONS AND ANALYSIS

The performance of the proposed channel sharing algorithm is compared with two other channel allocation algorithms, proposed in [18] and [17].

A. Comparative Channel Allocation Schemes

In this section, we summarize the allocation mechanism of the comparative TVWS allocation schemes. In [17], two TVWS sharing problems are defined; one for maximizing the number of channels allocated to the networks and the second for maximizing the total throughput under the minimum fairness constraint of allocating at least a single channel to each network. In this simulation setup, we implement the second problem as it closely matches with the channel sharing scheme proposed in this paper. The TVWS sharing algorithm proposed in [17] then selects a node (WSO) having a minimum of the assigned channels and the minimum number of the available channels to it. The algorithm assigns a TV channel to the selected WSO and calculates the total throughput. It keeps assigning the channel to other WSOs as long as the total throughput is increasing. This procedure is repeated for every channel. The algorithm terminates as no more increase in the throughput is observed.

The TVWS sharing problem in [18] is modeled as a lexicographic ordering of throughputs of access points of coexisting networks. The proposed problem is then transformed into a graph coloring problem. An algorithm called as, Share, is then proposed to solve the graph coloring problem. The Share algorithm operates in three phases. In the first phase of allocation, it orthogonalizes the WSOs in the available TV channels (frequency slots). In the second phase, a mutual channel sharing is performed among allotted WSOs of the first phase under the condition that their first phase throughputs do not decrease. The fairness is improved in the third phase by sharing the channel with unallocated WSOs such that lexicographically ordered throughputs do not decrease.

We select the algorithms in [18] and [17] due to the close resemblance of their TVWS sharing problems to the proposed channel sharing mechanism. For example, both considers optimizing throughput under minimum fairness in allocation. However, there exist some fundamental differences as well. For example, both the allocation schemes orthogonalize the WSOs in frequency domain by allocating a dedicated channel to each allocated WSO while the proposed scheme orthogonalize WSOs in a joint time-frequency domain by slicing the available TVWS in the frequency bands and further slicing each channel (frequency band) into a number of COTs in the channel window time, as discussed in section IV. Moreover, the algorithm in [17] is intended for TVWS channel allocation to cellular networks

while the proposed scheme is intended for TVWS sharing in an *ad hoc* coexisting environment, as discussed in Section I. Similarly, the TVWS sharing algorithm in [18] does not implement the FR concept. Therefore, we implement the proposed algorithm without FR process as well to have a fair comparison with the scheme in [18]. This is achieved by omitting Step 3 in Table 2 during the implementation of the proposed algorithm.

Finally, the performance of the proposed allocation scheme with and without FR implementation is judged in comparison with the Scheme in [17] and the Scheme in [18], respectively.

B. Simulation Setup

Simulation setup consists of 32 WSOs deployed in some geographic region and connected to an 802.19.1 coexistence system. The system has 32 CMs, each serving a single WSO. We select a dedicated CM for each WSO as the schemes in [18] and [17] performs TVWS sharing at network (WSO) level. The number of available TV channels in the region varies from 2 to 16. The WSO types and transmission powers are modeled using FCC regulations [2]. For this purpose, the specifications for fixed, mode 1 and mode 2 WSO types are used. The fixed, mode 1 and mode 2 type WSOs are allowed to have maximum antenna gain of 4 watts (W) effective isotropic radiated power (EIRP), 100 mW EIRP, and 100 mWatt EIRP respectively. The WSO access technologies are IEEE 802.22 and IEEE 802.11af. In this simulation setup, we implement the compulsory channel requirement of each WSO where the standard definition of the above technologies mandates a single TV channel of regulatory defined bandwidth as a requirement of a device to operate in the TVWS. Note that the bandwidth of a TV channel is set equal to 6 MHz.

Two parameters; WSO channel occupancy demand, $O_{w,j}^c$ and WSO density in the region, $K_{w,j}^c$ are varied to observe their effect on allocation behavior of the three allocation schemes as follows. Let T_j represents the window time on the j^{th} channel. Note that the 802.19.1 [10] does not define MAC layer frame structure for operations in TVWS. Therefore, the channel window time is not defined in an absolute time domain in 802.19.1. In this simulation setup, we define the channel window time as a unit length, without loss of generality, i.e., $T_j = 1, \forall j \in \mathcal{J}$. Then, three allocation subdomains are defined on the T_j as follows; low subdomain consists of up to 33 percent of the channel window time, defined as, $O^L := (0, 0.33]T_j$, A medium subdomain consisting of 34 to 67 percent of the channel window time, defined as, $O^M := [0.34, 0.67]T_j$ and a high subdomain consists of 67 to 100 percent of the channel window time, defined as, $O^H := [0.67, 1]T_j$. The channel occupancy demand of each WSO is then randomly defined on these subdomains.

The WSO density in the region is reflected using the number of WSOs in the CS of each WSO as follows. Let W be the number of WSOs registered in all CMs in the system then, we define three WSO density subdomains as; low $K^L := (0, 0.33]W$, medium $K^M := [0.34, 0.67]W$, and high

$K^H := [0.67, 1]W$. The CS of each WSO is randomly defined on these subdomains. Let $K_{w,j}^c$ represents the number of WSOs in the CS of w^{th} WSO on the j^{th} channel, registered in c^{th} CM. Then, the effect of the variability in the translated channel occupancy demand and WSO density is measured using a pair of parameters $(O_{w,j}^c, K_{w,j}^c)$. Note that varying each of these parameters on three respective subdomains results in $2^3 = 27$ possible allocation combinations. Out of 27, we select three cases to study the performance metrics defined in Section VI-C, as follows.

- Low: low COT, low WSO density, i.e., $O_{w,j}^c \in O^L$ and $K_{w,j}^c \in K^L$,
- Medium: medium COT, medium WSO density, i.e., $O_{w,j}^c \in O^M$ and $K_{w,j}^c \in K^M$, and
- High: high COT, high WSO density, i.e., $O_{w,j}^c \in O^H$ and $K_{w,j}^c \in K^H$.

Next, we apply the `intlinprog` routine of MATLAB® to solve the proposed TVWS sharing problem. The routine applies the mixed-integer linear programming technique. Since we need binary valued vector \mathbf{X} , therefore, we set all the decision variables, $x_{w,j}^c \in \mathbf{X}, \forall \mathbf{X}^c \in \mathbf{X}$, to be integer variables in the `intlinprog` routine. The binary decision may lead to the situation where the COT of allocated WSOs may not fit the channel window time. For example, let us assume the WSO 1, 2, 3 and 4 in Fig. 3 coexist in a TV channel. Let their COT demand is defined as, 0.25, 0.33, 0.37 and 0.15, respectively. Let us assume the `intlinprog` routine outcome as $\mathbf{X} = [1 \ 0 \ 1 \ 1]$, i.e., the WSO 1, 3 and 4 gets the channel. This results in total COT of allocated WSOs equal to 0.77 which is less than the channel window time; 1. The second WSO cannot be accommodated in the channel considering the constraint (16b). In this simulation, the solution \mathbf{X} is engineered such that the second WSO is partially allocated the desired COT so as to maximize the channel utilization while maintaining constraint (16b). The purpose of such engineering the solution is to reduce the channel waste. In order to have a fair comparison, the same engineering principle is applied to the allocation matrix generated by the comparative allocation schemes. The comparative analysis of the three allocation schemes is then performed as discussed in the following section.

C. Comparative Analysis

The relative performance of the three allocation schemes is evaluated using the following metrics: system throughput, fairness in allocation among CMs and WSO satisfaction from the allocation. These performance metrics are selected to analyze how well the three allocation schemes achieve the TVWS sharing objectives, as defined in Section III-B. The simulation results of the performance metrics are presented in Fig. 4 to Fig. 6, respectively. Subplots (a), (b), and (c) in these figures show the effect of varying the $(O_{w,j}^c, K_{w,j}^c)$ pair in low, medium and high subdomains, respectively. The study results

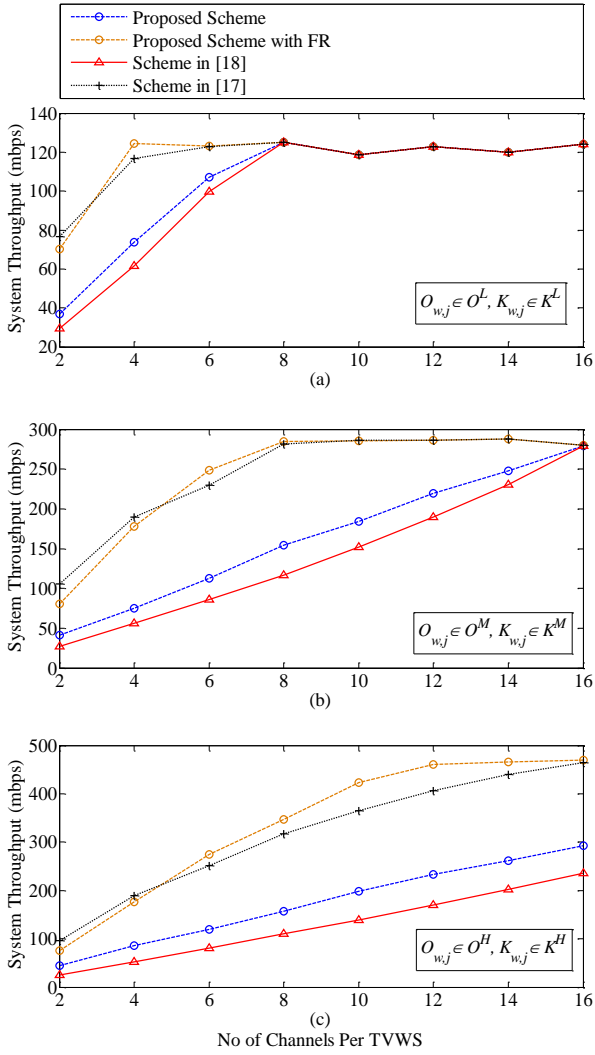


Fig. 4. System throughput for 32 WSOs registered in all CMs for a varying number of TV channels in the system.

are discussed as follows.

1) System Throughput

Fig. 4 shows the system throughput (ST) achieved by the three allocation schemes. Given the allocation matrix \mathbf{X} , and the SINR values, the ST is defined using Shannon capacity formula [45] as,

$$ST = \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} \sum_{w \in \mathcal{W}} x_{w,j}^c O_{w,j}^c b_j \log_2(1 + SINR_{w,j}^c). \quad (26)$$

It is shown in Fig. 4 that, for most of the channels in the system, the proposed scheme achieves higher ST gain than the comparative TVWS sharing schemes. However, the proposed scheme with FR implementation achieves slightly lower ST than the Scheme in [17] for the case when the number of channels in the system is two. This is because the Scheme in [17] focuses on maximizing the throughput in the TVWS allocation process while the proposed scheme focuses on making a balance among the contradicting QoS metrics; ST and fairness in allocation. Consequently, the WSOs with lower channel quality (here lower SINR value) also get a proportion of the available TVWS which reduces the total ST gain in the proposed scheme. However, as the number of channels in the system reaches to four and above,

the proposed scheme achieves higher ST gain and remains so until both the schemes converge to the maximum achievable ST. The reason for such improvement is that the proposed scheme applies a joint time-frequency FR concept which accommodates a higher number of WSOs in the available TV channels while the Scheme in [17] applies FR concept in frequency domain only. Note that the ST gain in this study is defined as maximum if all of the WSOs in all the CMs get their desired channel demands.

The effect of variability in the $(O_{w,j}^c, K_{w,j}^c)$ pair values on the ST gain of the three allocation schemes is shown in Fig. 4(a), 4(b) and 4(c), respectively. The three allocation schemes converge to the maximum ST, as the number of channels in the system reaches 8 and 16, as shown in Fig. 4(a) and 4(b), respectively. However, in high subdomain case (Fig. 4(c)), none of the allocation scheme achieves the maximum ST. The reason is that the high channel occupancy demand of WSOs results in a few WSOs to saturate the available TVWS while leaving no channel share for rest of the WSOs.

Another notable property of the ST study is that, as the $(O_{w,j}^c, K_{w,j}^c)$ pair values increases from low to high subdomains, the ST gain of the proposed scheme improves over ST gains in the comparative scheme, as shown in Fig 4(a) through Fig. 4(c), respectively. This improvement is attributed to the combined effect of the use of the proportional fairness in the allocation and implementing FR in a joint time-frequency domain in the proposed scheme, as defined in the Section IV and V respectively.

2) Fairness

The fairness in allocation among CMs in the system is measured using equation (21) where the variability in CMs'

normalized throughput vector, $\mathbf{T} = (T^1, T^2, \dots, T^C)'$ is used as a fairness metric to compute the fairness index (FI) value. The FI result, as shown in Fig. 5, confirms that the proposed scheme achieves the highest FI value due to the combined use of the proportional fairness method and the FR implementation in the joint time-frequency domain. On the other hand, though both, the Scheme in [17] and the Scheme in [18], optimize the fairness in allocation. However, both the schemes make an orthogonal TV channel allocation thus, resulting in lesser number of WSOs to get the channel which reduces FI value. Moreover, the constraint of maintaining the lexicographically ordered throughputs of the WSOs in the Scheme in [18] further reduces the degree of the fairness in allocation.

The effect of varying the values of the $(O_{w,j}^c, K_{w,j}^c)$ pair in low, medium and high subdomains is shown in Fig. 5(a), 5(b) and 5(c), respectively. It is shown in Fig. 5(a) and Fig. 5(b) that the FI values of all the comparative allocation schemes converge to the maximum FI value, i.e., 1, as the number of channels in the system reaches 8 and 16, respectively. However, in the high subdomain case (Fig. 5(c)), none of the comparative allocation schemes converge to the maximum FI value except for the proposed scheme with the FR implementation. It is because, in all other schemes, their orthogonal channel allocation policy result in a few WSOs to saturate the available TVWS while in the proposed scheme, the spatial reuse of the TVWS in a joint time-frequency domain

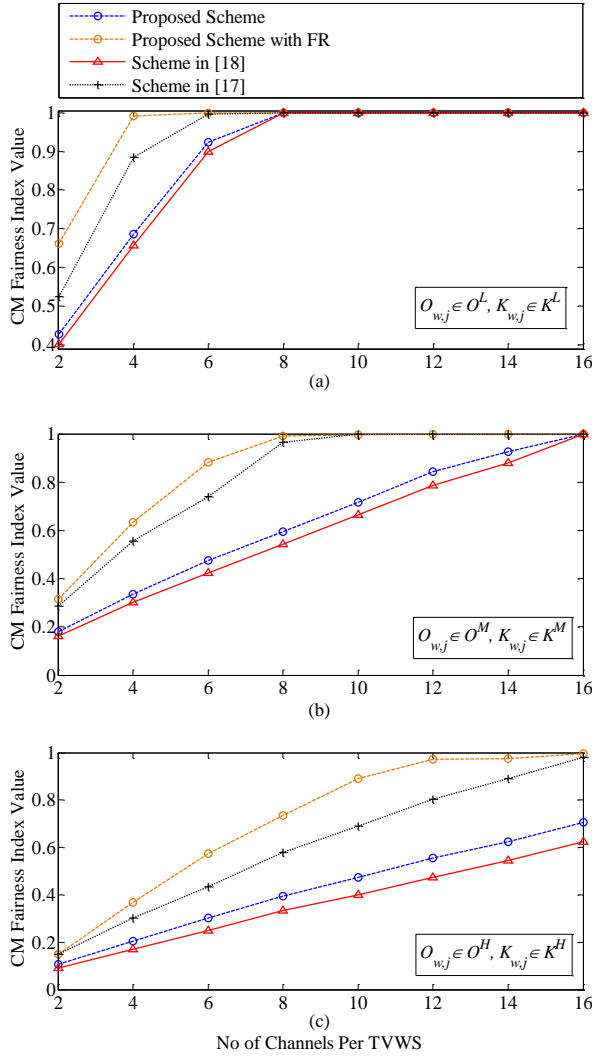


Fig. 5. Fairness index value calculated using normalized throughput vector of CMs for a varying number of TV channels in the system.

accommodates as many as WSOs, registered in the CMs which improves fairness in allocation.

3) WSO Satisfaction

In this study, we analyze the performance of the three allocation schemes the third objective of the TVWS sharing problem defined in Section III-B. In this study, a WSO is considered satisfied from allocation if it gets its desired channel demand for the duration of desired channel occupancy. The system-wide WSO satisfaction percentage (S) is then calculated using percentage of the mean satisfaction as,

$$S = 100 \sum_{c \in \mathcal{C}} \frac{\sum_{w \in \mathcal{W}^c} \frac{\sum_{j \in \mathcal{J}} x_{w,j}^c}{n_w}}{\mathcal{W}^c} \quad (27)$$

Fig. 6 shows the simulation result of the satisfaction study of the three allocation schemes. This figure shows that the proposed scheme and the Scheme in [18] achieves similar satisfaction result as their lines overlap each other. However, the proposed scheme with FR implementation achieves better satisfaction result than that of the Scheme in [17]. It is because, the TVWS allocation in a joint time-frequency domain enables the proposed

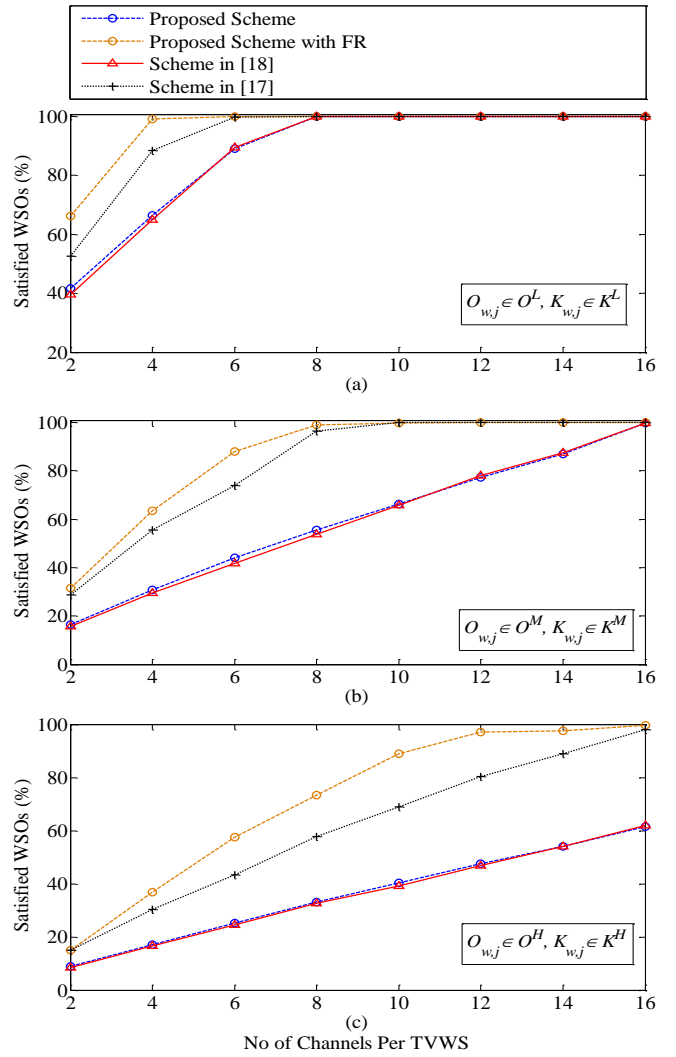


Fig. 6. Percentage of total 32 WSOs satisfied from the allocation.

scheme to accommodate as many as WSOs in the available TVWS while the third objective in the TVWS sharing problem, in Section III-B, requires the proposed scheme to satisfy the channel demand of each allotted WSO. Such an allocation strategy improves the satisfaction result of the proposed scheme.

From the results in Fig. 4 to Fig. 6, it is clear that none of the comparative schemes results in better performance than the proposed scheme in any of the performance metric. The proposed scheme, however, gives fairer channel allocation among all comparative allocation schemes. The proposed scheme with the FR implementation, however, outperforms the comparative schemes, in most of the TV channels in the system, in all the three performance metrics as shown in Fig. 4 to Fig. 6.

D. Increasing WSO Density

In this section, the effect of increasing the number of coexisting WSOs in the performance of the proposed allocation scheme is evaluated. The performance is measured using the metric like system throughput and WSO satisfaction, for the three subdomain cases, i.e., low, medium and high. The number of WSOs registered in each CM in the system varies in a set,

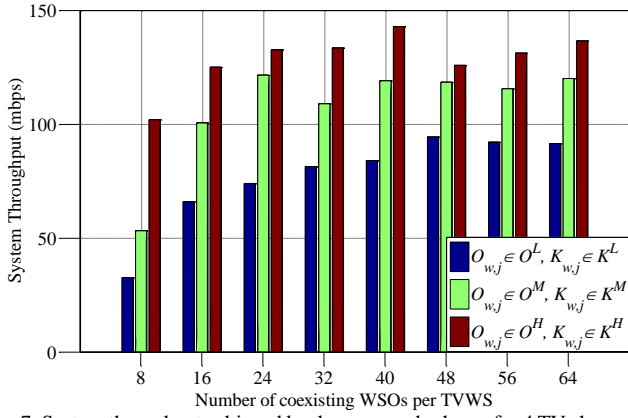


Fig. 7. System throughput achieved by the proposed scheme for 4 TV channels and a varying number of WSOs.

$W \in \{8, 16, 24, \dots, 64\}$. The number of available TV channels remains constant at 4, and the other simulation parameters are same as defined in Section VI-B. The results of the performance study are shown in Fig. 7 and Fig. 8.

Fig. 7 shows that the highest throughput gain is achieved in the high subdomain case, i.e., when $(O_{w,j} \in O^M, K_{w,j} \in K^M)$. The reason is that the proportional fairness method in the proposed scheme selects the WSOs with high throughput gain to share the available TVWS. While spatially reusing the frequency further helps the proposed scheme to accommodate as many as WSOs in the available TVWS. Consequently, the ST increases in high subdomain case. On the other hand, the achieved throughput is the least in low subdomain case, i.e., when $(O_{w,j} \in O^L, K_{w,j} \in K^L)$. It is because; the low channel occupancy demand of the WSOs could not saturate the available whitespace.

Fig. 8 shows the percentage of the number of WSOs satisfied from the allocation, calculated using (27). This figure shows that the satisfaction is the highest in the low subdomain, followed by the medium subdomain, especially in the case when $W=8$, for each CM. The reason is that a relatively greater number of WSOs can be satisfied per TVWS when $W=8$. The WSP value then sharply declines as the number of WSOs in the system increases, especially for the medium and high subdomain cases.

The results in Fig. 7 and Fig. 8 shall facilitate the modeling of a channel sharing system such that given the statistics of channel quality, the WSOs channel demands and the WSO density in the system, one can estimate an optimal number of WSOs that can be accommodated on the available TVWS such that the resource utilization is maximized.

E. Algorithm Scalability Test

The scalability of the proposed algorithm in terms of time taken to complete the allocation process is evaluated. In this experiment, the total number of coexisting WSOs registered in all the CMs in the system varies geometrically as, 2^W where $W \in \{3, 4, 5, 6, 7\}$. The number of TV channels in the system increases at a constant interval of 4 as, $J \in \{4, 8, 12, \dots, 48\}$. The remaining simulation parameters are same as defined in Section VI-B. The specifications of the computer system used for the scalability test is listed in Table 4. Using the above parameters,

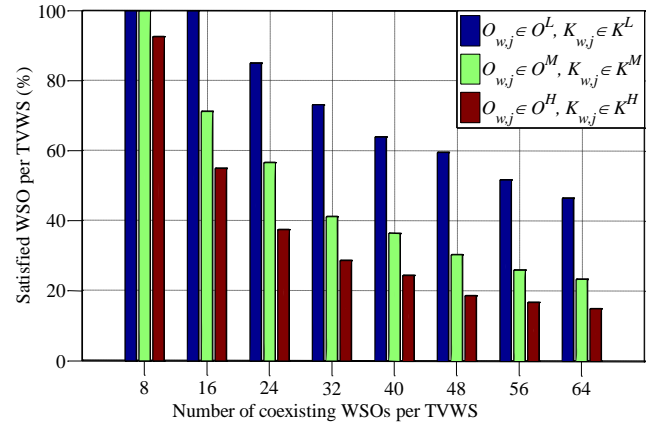


Fig. 8. WSOs satisfied from allocation with varying WSO density in the region. The number of TV channels in the system is 4.

the `intlinprog` routine solves the proposed TVWS sharing problem. The routine uses the branch and bound method to find an optimal solution point \mathbf{X} . The branch and bound split the problem into sub-problems, and each sub-problem is expanded until a solution is found as long as its cost does not exceed the set upper bound. The exact computational complexity of any branching algorithm is hard to find as time complexity of such a branching algorithm is usually analyzed by the method of branching vector. However, it has been mentioned in [46] that when the best-first search branch and bound technique is used, the upper bound to generate an expected solution is $\sum_{i=0}^n T(i) \leq \sum_{i=0}^n n-i+1 \leq (n+1)^2$ where n is the number of nodes visited. Thus, the complexity of such an algorithm is $\mathcal{O}(n^2)$.

In this experiment, we measure the simulation time taken using the MATLAB® tic-toc stopwatch timer. The time recorded for the high domain channel assignment is shown in Fig. 9. The result in this figure is generated using the average time required to complete allocation for the high subdomain case, i.e., $O_{w,j} \in O^H$ and $K_{w,j} \in K^H$. In this study, we perform the population engineering step, as defined in Section VI-B, using the `intlinprog` routine of the MATLAB. The figure indicates that for defined simulation parameters, the channel sharing process took a few hundreds of milliseconds to complete the allocation process which is quite acceptable for real-time implementation of the algorithm. The Fig. 9 shows that the algorithm execution time does not grow geometrically as the number of WSOs in the system increases. Rather, the algorithm has linear time allocation behavior as shown in Fig. 9.

VII. CONCLUSION

In this study, we investigated the channel sharing problem in a TVWS sharing domain with the objective of maximizing the resource utilization. The defined TVWS sharing problem optimizes the system throughput under a minimum fairness in allocation while constrained to satisfy the WSO channel occupancy demand on each allocated channel. To solve the defined problem, we proposed a channel allocation algorithm that shares the available TVWS among coexisting WSOs operating on incompatible network technologies. In order to

TABLE 4: COMPUTER SYSTEM CONFIGURATION

Symbol	Description	Quantity
Processor	Intel quad core i5-2500k	CPU = 3.30 GHz
Onboard memory	-	8555 MB
Memory used by MATLAB®	-	1289 MB

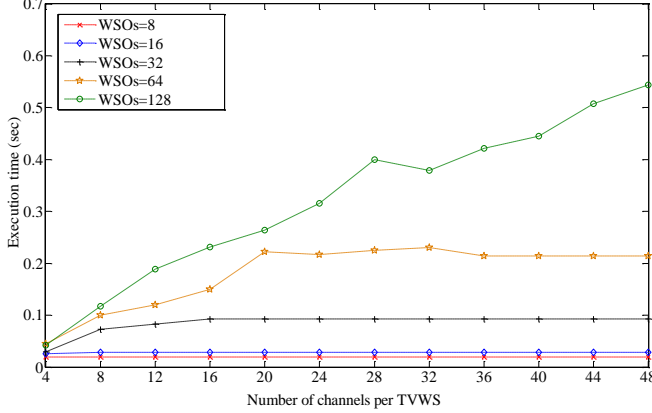


Fig. 9. Algorithm execution time for varying number of WSOs and varying number of TV channels in the system.

improve the TVWS utilization and to accommodate as many as WSOs in the available TVWS, the proposed algorithm spatially reuse the available TV spectrum. The simulation results show that the frequency reuse property of the proposed algorithm results in comparatively higher WSO satisfaction from the allocation, better fairness in allocation and higher system throughput gain. Moreover, the fast allocation process of the proposed algorithm makes it a promising candidate for implementation in 802.19.1 based coexistence system. The proposed algorithm can be implemented by a centralized decision-making entity, i.e., the master CM in the IEEE 802.19.1 system without requiring any major change in the baseline architecture of 802.19.1 TVWS sharing system.

Appendix A

In this section, we apply tangent plane approximation to linearize the objective function in (14a).

Let for some given points on the graph, $(q_1 = x_{1,j}^c, q_2 = x_{2,j}^c)$,

and $F = \log(q+1)$, where $q = \frac{q_1 r_{1,j}^c}{O_{1,j}^c + \delta_{O_{1,j}^c, 0}} + \frac{q_2 r_{2,j}^c}{O_{2,j}^c + \delta_{O_{2,j}^c, 0}}$. If

$\log(\mathcal{U}_{c,j} + 1)$ is differentiable at (q_1, q_2) , then the surface has tangent plane at (q_1, q_2, F) . The equation of the tangent plane at (q_1, q_2, F) is given by,

$$\frac{\partial y}{\partial x_{1,j}^c}(q_1, q_2)(x_{1,j}^c - q_1) + \frac{\partial y}{\partial x_{2,j}^c}(q_1, q_2)(x_{2,j}^c - q_2) - (F - F) = 0$$

where y denotes multivariate objective function $\log(\mathcal{U}_{c,j} + 1)$

and $F = \log(\mathcal{U}_{c,j} + 1)$.

The tangent plane equation is rearranged as,

$$F = F + \frac{\partial y}{\partial x_{1,j}^c}(q_1, q_2)(x_{1,j}^c - q_1) + \frac{\partial y}{\partial x_{2,j}^c}(q_1, q_2)(x_{2,j}^c - q_2)$$

where $\frac{\partial y}{\partial x_{1,j}^c} = \frac{r_{1,j}^c}{(\mathcal{U}_{c,j} + 1)(O_{1,j}^c + \delta_{O_{1,j}^c, 0})}$ denotes partial

derivative of log function at $x_{1,j}^c$. Thus, if F is differentiable at (q_1, q_2) , then the tangent plane to the surface at (q_1, q_2) provides a good approximation to F near (q_1, q_2) ,

$$F \approx F + \frac{\partial y}{\partial x_{1,j}^c}(q_1, q_2)(x_{1,j}^c - q_1) + \frac{\partial y}{\partial x_{2,j}^c}(q_1, q_2)(x_{2,j}^c - q_2)$$

which is called as linear approximation of y near (q_1, q_2) .

For a general case with $|\mathcal{W}^c| = n$, and near to some given point,

$\mathbf{q} = q_1 = x_{1,j}^c, \dots, q_n = x_{n,j}^c$, we define linear approximation of y as,

$$F \approx F + \frac{\partial y}{\partial x_{1,j}^c}(\mathbf{q})(x_{1,j}^c - q_1) + \dots + \frac{\partial y}{\partial x_{n,j}^c}(\mathbf{q})(x_{n,j}^c - q_n).$$

Appendix B

In this section, we aim to discuss the convergence property of the algorithm in Table 2. Note that our discussion here closely follows the discussion on the convergence of sub-gradient algorithm defined in [47]. Interested readers are referred to [47] for seeking knowledge beyond what is presented in this short discussion.

Given $\lambda^0 \in E^W$ and the sequence $\{t_k\}$ of positive scalars, called *step sizes*, in Table 2, define the sequence $\{\lambda^k\}$ as defined in Step 5-b) in Table 2,

$$\lambda^{k+1} = \max\{\lambda^k + t_k \nabla h(\lambda^k), 0\}.$$

For any λ , the maximum of (17) is assumed for at least one value of the index k . Since (17) is piecewise linear, there then exists at least one point λ^* such that $h(\lambda^*) = h^* = \max P(\mathbf{X}, \lambda^*)$. Then, $h(\lambda^k)$ will converge to its optimum h^* under the conditions,

$$\lim_{k \rightarrow \infty} t^k \rightarrow 0, \quad \sum_{k=0}^{\infty} t_k = \infty.$$

For the proof of the convergence of subgradient algorithm, the interested readers are encouraged to consult [47].

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIP) under Grant NRF-2015R1A2A1A05001826.

REFERENCES

- [1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021" San Jose, CA, USA, Tech. Rep., Feb. 2017.
- [2] Wikipedia, "White spaces (Radio)," [https://en.wikipedia.org/wiki/White_spaces_\(radio\)](https://en.wikipedia.org/wiki/White_spaces_(radio)), 2011.
- [3] FCC, "Third memorandum opinion and order in the matter of unlicensed operation in the TV broadcast bands, additional spectrum for unlicensed devices below 900 MHz and in the 3 GHz band," docket 02-380, Apr. 2012.
- [4] Ofcom, "Implementing TV White Spaces," Feb. 2015.
- [5] Industry Canada, "Consultation on a Policy and Technical Framework for the Use of Non-Broadcasting Applications in the Television Broadcasting Bands Below 698 MHz," Aug. 2011.
- [6] T. Baykas, M. Kasslin, M. Cummings, H. Kang, J. Kwak, R. Paine, A. Reznik, R. Saeed, S. J. Shellhammer, "Developing a Standard for TV White Space Coexistence: Technical Challenges and Solution Approaches," *IEEE Wireless Comm.*, vol. 9, no. 1, pp. 10-22, Feb. 2012.
- [7] C. Ghosh, S. Roy, and D. Cavalcanti, "Coexistence Challenges for Heterogeneous Cognitive Wireless Networks in TV Whitespaces," *IEEE Wireless Comm.*, vol. 18, no. 4, pp. 22-31, Aug. 2011.
- [8] L. Akter and B. Natarajan, "Modeling Fairness in Resource Allocation for Secondary Users in a Competitive Cognitive Radio Network," *Proc. IEEE WTS*, pp. 1-6, 2010.
- [9] G. Villardi, Y. D. Alemseged, C. Sun, C. S. Sum, T. H. Nguyen, T. Baykas, and H. Harada, "Enabling Coexistence of Multiple Cognitive Networks in TV White Space," *IEEE Wireless Comm.*, Aug. 2011.
- [10] IEEE Standard 802.19.1, "TV white space coexistence methods," May, 2014.
- [11] B. Bahrak and J. Park, "Coexistence Decision Making for Spectrum Sharing Among Heterogeneous Wireless Systems," *IEEE TWC*, vol. 31, no. 3, pp. 1298-1307, Mar. 2014.
- [12] IEEE Standard 802.15.2, "coexistence of wireless personal area networks with other wireless devices operating in unlicensed frequency bands," 2003.
- [13] S. Pollin, I. Tan, B. Hodge, C. Chun, and A. Bahai, "Harmful Coexistence between 802.15.4 and 802.11: a Measurement-Based Study," *Proc. 2008 Int. Conf. Cognitive Radio Oriented Wireless Netw.*
- [14] IEEE standard 802.11af, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications— Television white spaces (TVWS) operation," 2014.
- [15] IEEE Standard 802.22.1, "Standard to Enhance Harmful Interference Protection for Low-Power Licensed Devices Operating in TV Broadcast Bands," 2010.
- [16] *MAC and PHY for Operation in TV White Space*, std. ECMA-392, Jun. 2012.
- [17] F. Hessar and S. Roy, "Resource allocation techniques for cellular networks in TV white space spectrum," in *Dynamic Spectrum Access Networks (DYSPAN)*, 2014 IEEE International Symposium on, April 2014, pp. 72-81.
- [18] T. Bansal, D. Li, and P. Sinha, "Opportunistic Channel Sharing in Cognitive Radio Networks," *IEEE Trans. on Mobile Computing*, vol. 13, no. 4, pp. 852-865, Apr. 2014, doi: 10.1109/TMC.2013.59.
- [19] T. Marler and J. Arora, "Survey of Multi-objective Optimization Methods for Engineering," *Structural Multidisciplinary Optimization*, vol. 26, no. 6, pp. 369-395, 2004.
- [20] K. Khalil, G. Farhadi, and A. Ito, "Coexistence Management for Heterogeneous Networks in White Spaces," *Proc. International Conference on Computing, Networking, and Communications, Wireless Network Symposium*, pp. 691-697, Feb. 2014, doi:10.1109/ICCNC.2014.6785420.
- [21] D. Zhang, Q. Liu, L. Chen, W. Xu, "Ecology-based Coexistence Mechanism in Heterogeneous Cognitive Radio Networks," IEEE GLOBECOM, 2015.
- [22] K. Bian, J. M. Park and B. Gao, "Frequency reuse over a single TV white space channel," USA, Elsevier, pp 77-100, 2014.
- [23] B. Gao, J-M J. Park, and Y. Tang, "Uplink soft frequency reuse for self-coexistence of cognitive radio networks," *IEEE Trans. on Mob. Comp.*, Vol.13, Issue 6, June 2014. Pp. 1536-1233.
- [24] L. M. O. Khanbary, and D. P. Vidyarthi, "Reliability-based channel allocation using genetic algorithm in mobile computing," *IEEE Trans. Vehicular Tech.*, vol. 58, no. 8, pp. 4248-4256, Oct. 2009.
- [25] A. P. Shrestha, J. Won, S.-J. Yoo, M. Seo, and H.-W. Cho, "Genetic algorithm based sensing and channel allocation in cognitive ad-hoc networks," *Int. Conf. on ICT Convergence (ICTC)*, pp. 109-111, 2016.
- [26] Y. Jao, and I. Joe, "Energy-efficient resource allocation for heterogeneous cognitive radio network based on two-tier crossover genetic algorithm," *Jr. Commun. Networks*, vol. 18, no. 1, pp. 112-122, Feb. 2016.
- [27] D. T. Ngo, C. Tellambura, and H. H. Nguyen, "Efficient resource allocation for OFDMA multicast systems with spectrum-sharing control," *IEEE Trans. Vehicular Tech.*, vol. 58, no. 9, pp. 4878-4889, Nov. 2009.
- [28] M. R. Sherif, I. W. Habib, M. Nagshineh, and P. Kermani, "Adaptive allocation of resources and call admission control for wireless ATM using genetic algorithms," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 2, pp. 268-282, Feb. 2000.
- [29] M. Alias, S. Chen, and L. Hanzo, "Multiple-antenna-aided OFDM employing genetic-algorithm-assisted minimum bit error rate multiuser detection," *IEEE Trans. Veh. Technol.*, vol. 54, no. 5, pp. 1713-1721, Sep. 2005.
- [30] F. Hessar and S. Roy, "Capacity considerations for secondary networks in tv white space," *IEEE Trans. on Mob. Comp.*, Vol.14, Issue 9, Sep. 2015. Pp. 1780-1793.
- [31] L. E. Li, M. Pal, Y. R. Yang, "Proportional Fairness in Multi-rate Wireless LANs", *IEEE INFOCOMM.*, 2008, pp. 1004-1012.
- [32] T. Bonald, L. Massouli, A. Prouti, and J. Virtamo, "A Queuing Analysis of Max-min Fairness, Proportional Fairness, and Balanced Fairness," *Queueing Systems: Theory and Applications*, vol. 53, no. 1-2, 2006.
- [33] L. B. Jiang and S. C. Liew, "Proportional Fairness in Wireless LANs and Ad-hoc Networks," *IEEE WCNC*, 2005.
- [34] F. Kelly, "Charging and Rate Control for Elastic Traffic," *European Trans. on Telecomm.*, vol. 8, no. 1, pp. 33-37, 1997.
- [35] T. Bu ; L. Li ; R. Ramjee, "Generalized Proportional Fair Scheduling in Third Generation Wireless Data Networks," in *Proceedings IEEE INFOCOM 2006*, April 2006.
- [36] C. Cano, D. J. Leith, "Coexistence of WiFi and LTE in Unlicensed Bands: A Proportional Fair Allocation Scheme," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 2288 - 2293, June 2015.
- [37] FP Kelly, AK Maulloo and DKH Tan, "Rate control for communication networks: shadow prices, proportional fairness, and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237-252, Mar. 1998.
- [38] S. S. Byun and C. Yoo, "Minimum DVS Gateway Deployment in DVS-based Overlay Streaming," *Computer Comm.*, vol. 62, no. 3, pp. 537-550, 2008.
- [39] L. Daniel and K. Narayanan, "Congestion Control 2: Utility, Fairness, and Optimization in Resource Allocation," *Lecture Notes*, March 2013.
- [40] W. Murray and K. M. Ng, "An Algorithm for Nonlinear Optimization Problems with Binary Variables," *Computational Optimization and Applications*, vol. 47, no. 2, pp. 257-288, Oct. 2010.
- [41] M. Pioro and D. Medhi, "General Optimization Methods for Network Design," *Routing, Flow, and Capacity Design in Communication and Computer Networks*, USA, Elsevier, pp. 178-184, 2004.
- [42] R. Jain, "Selection of Techniques and Metrics," *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design Measurements Simulation and Modeling*, Wiley Computer Publishing, pp. 30-40, 1991.
- [43] R. Jain, D.M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Technical Report, Eastern Research Lab., Dig. Equip. Corp., Hudson, MA, available at: <http://www1.cse.wustl.edu/jain/papers/ftp/fairness.pdf>, 1984.
- [44] IEEE Standard for information technology, Part 22: Cognitive wireless RAN medium access control (MAC) and physical layer (PHY) specifications: Policies and procedures for operation in the TV Bands, IEEE Std 802.22-2011, Jul. 2011.
- [45] T. Robert, "Data Communications Fundamentals," *Data Communications: An Introduction to Concepts and Design*, USA, Elsevier, pp. 68-69, 2013.
- [46] A. Kugel, "Proofs for the paper: Average Case Complexity of Branch-and-Bound Algorithms on Random b-ary Trees," Proofs Report, Faculty of Engineering and Computer Sciences, Ulm University, 89069 Ulm, available at https://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.190/Mitarbeiter/kuegel/proofs.pdf.
- [47] M. Held, P. Wolfe, and H. P. Crowder, "Validation of Subgradient Optimization," *Math Programming*. Vol. 6, 1974, pp. 62 - 88, IBM Systems Research Institute, New York, U.S.A.,

A Versatile Coexistence Decision-Making System for Efficient TV Whitespace Sharing among Whitespace Objects

M. Asif Raza, Zafar Iqbal, Sang-Seon Byun, Hyunduk Kang, Heung-No Lee*

Abstract—In this paper, a coexistence decision making (CDM) system for efficient TV whitespace (TVWS) sharing among whitespace objects (WSOs), registered in coexistence managers in IEEE 802.19.1 system, is introduced. The proposed system is considered versatile in functionality as it jointly takes care of three distinct channel allocation features; a) optimizing system quality of service (QoS) performance metrics, b) improving TVWS utility and c) satisfying WSO channel demands. Regarding system QoS performance metrics, the TVWS sharing problem is defined as an optimization problem with an aim to maximize the system throughput and minimizing unfairness in allocation. Supporting the WSOs channel demands in a TVWS sharing problem is a multifold task which requires elaborate consideration in different aspects of the system performance. To this end, the variations of the SNR of wireless frequency channels which result in variable throughput gain of the WSOs are also taken care of the proposed CDM system. A fast channel allocation algorithm is then designed that implements the TVWS sharing mechanism in a reasonable amount of time. Additionally, the proposed algorithm improves the TVWS utility by promoting a novel frequency reuse method by exploiting the inter-WSO interference information. Simulation results show the superiority of the proposed algorithm over existing TVWS sharing algorithms.

Index Terms— Frequency Reuse, Lagrangian Relaxation, Linear Approximation, Proportional Fairness, TV Whitespace

I. INTRODUCTION

AN unprecedented increase in the deployment of content delivery networks (CDNs) has resulted in the rapid growth of IP traffic. It is reported that by the end of 2016, global IP traffic exceeded 1 zettabytes (10^{21} bytes) per year, of which 62% is attributed to CDNs [1]. It is also anticipated that by 2019,

nearly two-thirds of global IP traffic will originate from non-PC devices, mainly portable and mobile devices [1]. On the other hand, currently available wireless spectrum is considered insufficient for accommodating such large volumes of data. Fortunately, the digitization of TV transmission has partially relinquished VHF and UHF spectrum [2]. Owing to its low loss and excellent propagation characteristics, the TV spectrum is considered a promising candidate for supporting the growing traffic over wireless channels. Considering the growing demand of the wireless spectrum, the regulatory bodies worldwide [3], [4], [5], have permitted unlicensed use of the TV spectrum under certain limits to protect the incumbents. However, the problem of coexistence of secondary devices operating in the same TV band was not dealt by the regulatory bodies.

The coexistence among secondary devices operating in TV spectrum is considered a challenging task due to signal propagation characteristics of TV channels, spatiotemporal variation of TV spectrum and disparity in network technologies of devices operating in the TV spectrum [6]. These diversities may cause coexistence issues, such as an unresolvable interference, spectrum congestion, diversity in network size, etc., as explained in [6], [7], [8], [9]. To address coexistence issues and regulate access to TV spectrum, IEEE has proposed an 802.19.1 standard [10]. The standard provides a set of procedures to enable coexistence among secondary networks operating in heterogeneous network technologies in TVWS, namely WSOs.

A set of procedures that ensures peaceful coexistence among a set of WSOs operating in the same spectrum is referred to as CDM [11]. In this paper, we define an 802.19.1 compliant CDM system that performs TVWS sharing among a set of WSOs, operating in dissimilar MAC/PHY layer technologies and registered in the coexistence manager (CM); an entity in 802.19.1 coexistence system as shall be defined in section III-A. Note that the TVWS refers to the TV spectrum not in use by licensed operators in a spatio-temporal region [10]. The TVWS sharing problem is modeled as an optimization problem with an aim to maximize the system performance metrics like system throughput and fairness in TVWS allocation. The optimization problem is constrained that the channel demands of the WSOs registered in the neighboring CMs are satisfied. In this perspective, variations of the SNR of wireless frequency channels which result in variable throughput gain of the WSOs are taken care of. Note that the neighboring CMs refer to the set of CMs whose WSOs create interference to each other and such WSOs are neighboring WSOs. Thus, the proposed CDM

- M. Asif Raza is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. E-mail: raza@gist.ac.kr.
- Zafar Iqbal is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. E-mail: zafar@gist.ac.kr.
- Sang-Seon Byun is with the Computer Engineering Department, Catholic University of Pusan, Busan 609-757, South Korea. E-mail: ssbyun@gmail.com.
- Hyunduk Kang is with Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, South Korea. Email: henry@etri.re.kr.
- Heung-No Lee* is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. E-mail: heungno@gist.ac.kr. (*the corresponding author.)

system differs from the notion of traditional node-based, link-based or base-station based channel allocation as reported in the TVWS sharing literature. Moreover, the proposed system also improves the TVWS utility by implementing a frequency reuse (FR) method to spatially reuse the available TV spectrum in a joint time-frequency domain in an *ad hoc* coexisting environment. In this paper, the *ad hoc* coexisting environment refers to the coexistence of both, infrastructure based WSOs like WLAN, and *ad hoc* WSOs like personal area network. An *ad hoc* WSO accounts for a local area network that is built spontaneously, as devices connect with each other. The CDM system proposed in this paper is unique, to the best of our knowledge, in the sense that it jointly focuses three distinct TVWS sharing objectives; a) optimizing system performance metrics during TVWS sharing among WSOs registered in neighboring CMs in 802.19.1 system b) improving the TVWS utility by implementing the FR in a joint time-frequency domain, c) taking care of the channel demands of the *heterogeneous*-WSOs. Such a joint focus to implement multiple distinct channel allocation features makes the proposed system a versatile CDM system.

The remainder of the paper is organized as follows. Section II reviews some related studies. Section III summarizes technical background required to establish the baseline for the techniques used in the paper. The system description and problem formulation are defined in Section IV. Section V discusses the solution method and the proposed algorithm. Section VI presents the simulation results and compares the proposed algorithm with existing algorithms. Finally, Section VII concludes the paper.

II. PREVIOUS WORK

In this section, we summarize some standards and algorithms developed for achieving coexistence among secondary users.

IEEE 802.15.2 [12] and 802.15.4 [13] have partially addressed the coexistence issue among devices operating on wireless local area networks and low power wireless personal area networks, respectively. However, these networks operate on industrial, scientific, and medical bands. On the other hand, IEEE 802.22 has recently defined PHY and MAC layer extensions for TVWS. Similarly, IEEE 802.11af [14] has adopted new cognitive radio features to protect incumbents and achieve efficient spectrum utilization among unlicensed devices. IEEE 802.22.1 has also defined methods for peaceful coexistence when a low-power licensed device such as a microphone broadcaster and an unlicensed device both coexist and share the same channel [15]. The European Computer Manufacturers Association (ECMA) has also defined a specification (ECMA 392) for personal/portable cognitive wireless networks operating in TVWS [16]. However, all these standards define self-coexistence in TVWS operations. Non-availability of cross-platform coexistence mechanisms shall cause issues such as an inability to diagnose interference among networks with dissimilar network technologies and may lead to inefficient utilization of the scarce wireless spectrum [11]. Perceiving the need for cross-platform coexistence mechanisms, IEEE has defined an 802.19.1 standard. This standard provides coexistence

protocols and policies for efficient utilization of TVWS across platforms [10].

On algorithmic perspective, a CDM algorithm that results in fair TVWS sharing among neighboring CMs is presented in [10]. The algorithm applies max-min fairness technique to establish fair share distribution during the TVWS sharing process. The issue with the algorithm in [10] is that it focuses fairness in allocation while no consideration to effective utilization of the available TVWS is taken care. Considering the scarcity of the TV spectrum, especially in highly congested spectrum environments, the effective utilization of the available TV spectrum is also an important factor to be considered. Hesar and Roy [17] have discussed the TVWS sharing formulations in secondary cellular networks. The authors adopt heuristic based approaches to defining greedy algorithms to tackle the identified TVWS sharing problems. However, the proposed greedy algorithm for throughput maximization sub-problem searches the entire network to find an optimal solution. For such an algorithm, search over the space of a possibly very large number of network and channel collocation combinations leads to a high runtime complexity to find an optimal solution. An algorithm for opportunistic whitespace sharing among secondary networks has been presented as a graph coloring problem in [18]. The channel sharing algorithm in [18] solves the sharing problem by classifying the sharing process as network wide channel sharing and its localized version. This scheme, however, has performance issue when interference among neighboring access points is relatively high. This situation is quite common in highly congested areas where many collocated WSOs are deployed. Bahrak and Park [10] proposed an algorithm for CDM among heterogeneous networks. The sharing problem in [10] is modeled as a weighted-sum multi-objective optimization problem (MOOP) that is solved using a modified Boltzmann machine. However, an issue in the weighted-sum approach is that it does not find Pareto optimal points in non-convex regions of the solution space boundary [19]. Thus, some of the potential Pareto optimal points are possibly missed by the weighted-sum method.

Khalil et al., have also performed TVWS sharing among heterogeneous networks by defining an interference graph of the networks [20]. A two-stage algorithm is then designed to achieve spectrum sharing among graph nodes. The algorithm maximizes fairness by maximizing the frequency reuse. However, the channel sharing algorithm in [20] has polynomial runtime complexity $\mathcal{O}(N^3)$, for the number of networks (N). This complexity shows that in areas with a high number of deployed networks, the algorithm shall require substantial channel allocation time. Zhang et al., [21] adapt ecology based species competition model to develop a coexistence mechanism called ecological Species Competition based HETerogeneous networks coexistence MEchanism (SCHEME). The SCHEME enables each coexisting network to adjust achieved bandwidth per its QoS requirements dynamically. However, the SCHEME requires the number of channels to be larger than the number of coexisting networks. Such condition cannot be fulfilled in highly congested urban areas where a limited number of TV channels is available for unlicensed use. We have addressed this issue in the

channel allocation mechanism defined in this paper.

On the other hand, some of the existing TVWS sharing algorithms have implemented the concept of FR. For example, in [22], Bian et al., have implemented the concept of FR in sharing a single TV channel among Cognitive Radios (CR). The CR networks operating in orthogonal frequency division multiple access apply the uplink soft FR concept [23]. Again, the proposed method is defined for CR systems deployed in cellular infrastructure. Similarly, Hessar and Roy [17] have presented an FR method in cellular networks operating in TVWS. Moreover, the algorithm proposed in [17] orthogonalizes WSOs in frequency domain only. None of the existing TVWS sharing algorithms reuses TVWS in a joint, time-frequency domain for WSOs operating in an *ad hoc* coexisting environment. Spectrum reuse in both time and frequency domains shall result in even a better utilization of the available TVWS, as discussed in Section VI-C.

Some genetic algorithms (GA), defined for implementing the channel sharing problem, also exist in the literature. For example, the authors in [24] use a GA-based reliability model to assign channels to mobile hosts based on the reliability of the base station and the channels to enhance the overall reliability of the mobile network system. The results show that this method requires higher number of iterations and generally higher number of available channels than the number of mobile hosts in order to achieve higher reliability. Similarly, Shrestha et. al., proposes a GA-based joint out-of-band spectrum sensing and channel allocation scheme for cognitive radio networks [25]. The joint sensing and resource allocation optimization problem has been formulated using fitness functions of sensing utility and the data transmission utility. Jao and Joe consider a new cognitive radio network model with heterogeneous primary users operating simultaneously via multi-radio access technology [26]. It focuses on energy efficient resource allocation and use a GA-based scheme to obtain an optimal solution in terms of power and bandwidth. The authors in [27] proposed solutions for the problem of efficient resource allocation (radio spectrum and power) in the OFDMA-based multicast wireless system that balances the tradeoff between maximizing the total throughput and ensuring a flexible and controllable spectrum sharing among multicast groups. It proposes two separate optimization methods for subcarriers and power and a GA-based joint optimization scheme is used. Results show that the proposed schemes can attain a high total sum-rate and more flexible and fair distribution of the available bandwidth among multicast groups.

The GA in these and such literature work [28], [29] are well suited for multi-objective optimization problems that require searching over a large space under several constraints. However, GA-based methods are computationally expensive and therefore not suitable for the optimization problem with single objective function and a small search space, like the one defined in this paper. Therefore, GA suffers from the drawbacks of slow convergence speed, and low stability. The channel allocation in highly dynamic spectrum environments requires an algorithm that can do allocation process in a quick runtime. Therefore, rather than applying the GA method, the

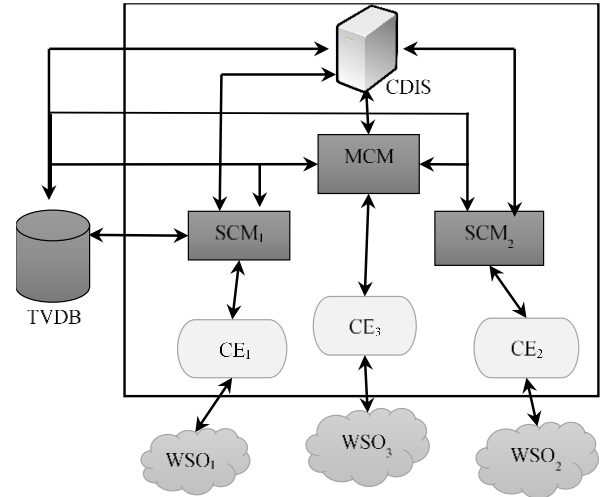


Fig. 1. IEEE 802.19.1 TVWS system architecture. The TVWS database and WSOs interact with the 802.19.1 architecture externally.

nonlinear, binary constrained optimization problem, defined in this paper is transformed into linear optimization problem. Such formulation helps us to apply linear programming solvers to solve the optimization problem and complete the allocation process in a quick, linear runtime.

III. TECHNICAL TERMS AND RESEARCH FOCUS

A. Technical Terms

In this section, we define technical terms that form baseline of the proposed TVWS sharing system, defined in the next section. The proposed system is based on the coexistence system architecture as described in [10] and shown in Fig. 1. The coexistence system in [10] has three logical components: coexistence manager (CM), coexistence enabler (CE), and a coexistence discovery and information server (CDIS).

- The CE registers a WSO to the CM and acts as a communication bridge by translating messages between the WSO and the CM serving the WSO.
- The CM makes coexistence decisions for WSOs registered in it. Moreover, it is required to interact with other CMs, called as neighboring CMs in [10] to resolve coexistence issues among WSOs served by neighboring CMs. In general, it sends configuration commands and control information to the CE.
- The CDIS provides coexistence discovery services like coexistence set information to CMs for registered WSOs.
- The *TVWS database* (TVDB), as shown in Fig. 1, is not part of the coexistence system architecture. It contains information about channels available in the geographic region of each WSO registered with the 802.19.1 system. The TVWS database provides information about the set of TV channels free for whitespace activity to the CMs.

A WSO may register with the IEEE 802.19.1 system before operating in the TV spectrum. In the registration process, a general principle for a WSO to acquire a TV channel is defined in IEEE 802.19.1, summarized as follows. A WSO may perform spectrum sensing to identify and select an available free TV channel or alternatively, it may send a channel

allocation request to its serving CM. If no free channel is available in the geographic region of the WSO, the CM may perform channel sharing among the requesting WSO and the WSOs pre-allocated a TV channel. If such WSOs are registered with other CMs, the CM serving the channel requesting WSO interacts with the other CMs to perform channel sharing. These CMs are called as neighboring CMs to the requesting CM. In this channel sharing procedure, two types of topologies are defined in the 802.19.1 [10]. A distributed CDM topology where neighboring CMs mutually interact to perform channel sharing among WSOs registered within them. A centralized CDM topology where multiple CMs agree to select one of them a master CM (MCM) and rest of the CMs become slave CM (SCM) [10], as shown in Fig. 1. Each SCM provides essential information about operating parameters, including the channel characteristics of each WSO registered within it and its channel demands to the MCM. The MCM performs coexistence services like radio resource allocation to WSOs registered in the SCMs. Some other terms used in the paper are defined as follows.

- A WSO is an entity in 802.19.1 system that represents a TVWS device or network of devices.
- The *channel occupancy* is the duty cycle in a percentage that a network (WSO) occupies a channel [10].
- The *window time* is a slot duration of a scheduling repetition period that satisfies the essential system QoS performance [10].
- The *Coexistence Set* (CS) of a w^{th} WSO is a set of WSOs that are registered in the neighboring CMs that may affect the performance of the w^{th} WSO. In other words, it is a set of WSOs which create interference to the w^{th} WSO.

B. Research Focus

The TVWS sharing problem is defined as,

Given a set of available TV channels, a set of CMs with each CM having at least one WSO registered in it and WSOs channel demands, share the TV channels among WSOs such that the following objectives are achieved.

- 1) Maximize the system throughput,
- 2) Minimize unfairness in allocation among WSOs registered in neighboring CMs, and
- 3) Fulfill desired channel demands of the allocated WSOs.

These objectives contradict each other. For example, maximizing the system throughput shall decrease fairness in allocation. Note that from a spectrum allocation perspective, fairness is regarded as equity in access to the resource, the TV spectrum. In other words, being free to use, each network should have an equal opportunity to an access to the given TV spectrum.

Similarly, fulfilling the second and third objectives in conjunction, under the scarcity of the available TVWS, restricts the system accommodating as many as WSOs in the TVWS. Thus, maximizing the fairness while satisfying the channel demands of each allocated WSO is quite complicated in highly congested spectrum environments [30]. Therefore, the fairness in allocation is measured at CM level. The fairness among CMs is deemed at minimum if at least a single WSO in each CM gets the channel.

Considering the above conditions, we design a CDM system, as will be defined in Section IV-A. The system is designed to implement at the MCM in the centralized topology in 802.19.1, as shown in Fig. 1. The system makes use of the information from information messages defined in the 802.19.1 [10] to apply various procedures for defining the proposed TVWS sharing problem as an optimization problem. For example, the WSO registration clause in [10] defines different information acquiring messages that permit a CM to collect desired channel demands, channel statistics, coexistence set elements, available TV channels and related information from WSOs registered within it or with neighboring CMs. Moreover, the inter-CM information sharing messages are also defined in [10]. We assume that using such message templates, the neighboring CMs exchange respective WSOs information with MCM. In order to solve the TVWS sharing problem, the CDM system in MCM then implements a channel allocation process, as will be defined in section V-C. The algorithm makes use of such information available at MCM to implement the subgradient method to solve the TVWS sharing dual problem, Section V-B, to identify a set of WSOs to allocate the TV channels.

The channel allocation process also implements a novel spectrum reuse in Table 3 to have an efficient use of the available TVWS. The spectrum reuse step is also made in compliant with the 802.19.1 by repeated channel allocation using an interference matrix. The CDM defines the interference matrix using the WSOs' CS information available at MCM, as shall be discussed in Section V-D. Note that the CS information is provided by the coexistence discovery algorithm as defined in [10]. The channel allocation process is then executed repeatedly to spatially reuse the TV spectrum to the unallocated WSOs that should not cause interference to pre-allocated WSOs. The proposed channel allocation solution is thus made smoothly integrable to the 802.19.1 system.

IV. SYSTEM DESCRIPTION AND PROBLEM FORMULATION

In the following section, a centralized CDM system is designed that implements a channel allocation process, as shall be discussed in Section V, to implement the TVWS sharing problem defined in Section III-B.

A. System Model

The CDM system is defined as follows,

$$\mathbf{X} = TVWS(\mathcal{C}, \mathcal{J}, \mathcal{Z}, \mathcal{T}, \mathcal{D}). \quad (1)$$

The system parameters are defined as follows. Let c be an index to a set of C neighboring CMs in the system, denoted as \mathcal{C} in Table 1. Let $\mathcal{W}^c, \forall c \in \mathcal{C}$ be a set of network IDs of WSOs registered in the c^{th} CM, as shown in Table 1. Let the network ID, $NID_w \in \mathcal{W}^c$ represents an identifier of the network the w^{th} WSO, registered in c^{th} CM, represents. For example, in the case of IEEE 802.11 type WSO, the NID contains the basic service set identifier used by the WSO.

Let j be an index to the set of all permissible TV whitespace channels, $\mathcal{J} = \{1, 2, \dots, J\}$, where each set element corresponds to a TV channel number, defined on the basis of

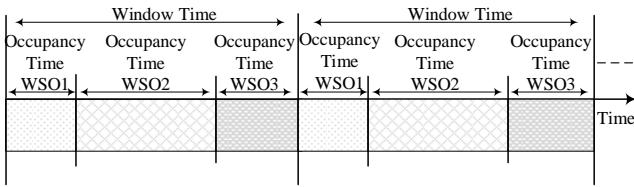


Fig. 2. Scheduling transmission periods for three WSOs on a TV channel.

the regulatory authority rulings. For example, in USA where FCC defines each TV channel to be 6 MHz bandwidth in V/UHF band, therefore, $\mathcal{J} = \{2, 3, \dots, 36, 38, \dots, 51\}$ in the USA. Since, the availability of a TV channel to a w^{th} WSO is a function of geographic location of the WSO and the primary user activity in the region. Therefore, the availability of a TV channel for the secondary use varies spatiotemporally and needs to be determined. We assume that a channel sensing mechanism, as defined in [10] is implemented such that the TVDB contains the set of TV whitespace channels available in the geographic region of each WSO registered in the CMs in the system. Let j be an index to the set \mathcal{J} , then, j^{th} channel availability status to the w^{th} WSO, registered in c^{th} CM, is represented by an indicator function defined as,

$$z_{w,j}^c := \begin{cases} 1, & \text{if } j^{\text{th}} \text{ channel in } \mathcal{J} \text{ is available to } w^{\text{th}} \text{ WSO} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The availability of J channels to the w^{th} WSO, registered in c^{th} CM, are thus represented by a vector of indicator functions defined as,

$$\mathbf{z}_w^c = (z_{w,1}^c, \dots, z_{w,J}^c).$$

The set of channels available to W WSOs registered in c^{th} CM is defined as,

$$\mathbf{Z}^c = (\mathbf{z}_1^c, \mathbf{z}_2^c, \dots, \mathbf{z}_W^c)^T, \forall c \in \mathcal{C}.$$

The system parameter \mathcal{Z} is then defined as follows,

$$\mathcal{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^C\}. \quad (3)$$

The parameter \mathcal{T} in the system in (1) represents the set of window times for the channels in the set \mathcal{J} . In 802.19.1, an algorithm is provided that enables CMs to define the slot duration of the window time. We assume the CMs implement such an algorithm to define the window time, $T_j, \forall j \in \mathcal{J}$, which is then used to define system parameter as,

$$\mathcal{T} = \{T_1, \dots, T_J\}. \quad (4)$$

The system parameter \mathcal{D} in (1) encodes channel demands of CMs, defined as follows. In 802.19.1 [10], a Discovery Information abstraction is provided that allows WSOs to send channel statistics and channel demands like SINR, desired channel occupancy, desired bandwidth etc., to their serving CM [10]. Such information of *heterogeneous*-WSOs is used to define a set of channel demands of w^{th} WSO as follows.

Let $SINR_{w,j}^c$ represents the quality of j^{th} channels to w^{th}

TABLE I
DEFINED PARAMETERS

Input Variables		
Symbol	Description	Value
\mathcal{C}	A set of C CMs in the system.	$\mathcal{C} = \{1, 2, \dots, C\}$
\mathcal{W}^c	A set of NID of W WSOs registered in the c^{th} CM.	$\mathcal{W}^c = \{NID_1, NID_2, \dots, NID_W\}$
\mathcal{J}	A set of permissible TV channels in the system.	$\mathcal{J} = \{1, 2, \dots, J\}$
\mathcal{D}	Channel demands of WSOs, as defined in the system in (1).	-
$O_{w,j}^c$	COT that translates desired occupancy demand of w^{th} WSO on a j^{th} channel.	$\mathbf{O}^c = [O_{w,j}^c]_{1 \times J}, O_{w,j}^c \in \mathbb{R}_{[0, T_j]}$
$I_{w,m}(j)$	Indicator variable encoding m^{th} WSO interference to w^{th} WSO on a j^{th} channel.	$I_{w,m}(j) := \begin{cases} 1 & \text{if } m \text{ interferes } w \\ 0 & \text{otherwise} \end{cases}$
$\mathcal{S}_{w,j}$	Set of WSOs $m \in \mathcal{W}$ such that m^{th} WSO transmission interferes w^{th} WSO transmission on j^{th} channel	$\mathcal{S}_{w,j} = \{m \in \mathcal{W}\}$
$y_{w,j}$	A variable indicating whether m^{th} WSO interferes w^{th} WSO on the j^{th} channel?	$y_{w,j} := \begin{cases} 1 & \exists w \in \mathcal{W} : m \in \mathcal{S}_{w,j} \\ 0 & \text{else} \end{cases}$
$z_{w,j}^c$	An element of the matrix \mathbf{Z} defining accessibility of j^{th} channel to w^{th} WSO.	$\begin{cases} 1 & \text{if } j \text{ accessible to } w^{\text{th}} \text{ WSO} \\ 0 & \text{else} \end{cases}$
Output Variables		
$x_{w,j}^c$	Element of matrix \mathbf{X} defining allocation status of w^{th} WSO on j^{th} channel	$x_{w,j}^c := \begin{cases} 1 & \text{if channel allocated} \\ 0 & \text{otherwise} \end{cases}$

WSOs registered in c^{th} CM. The channel quality is measured in terms of signal to interference and noise ratio (SINR) which depends on interference from primary-to-secondary users and noise floor due to environmental factors. We assume that an interference discovery mechanism is in place that enables each WSO to measure SINR value on each of the channels in \mathcal{J} , as will be further discuss in Section V-D. The quality of all J channels to w^{th} WSO is then defined as,

$$\mathbf{s}_w^c = (SINR_{w,1}^c, SINR_{w,2}^c, \dots, SINR_{w,J}^c), \forall w \in \mathcal{W}^c.$$

Let $p_{w,j}^c$ be the allowed transmission power to w^{th} WSO in the j^{th} channel. The allowed transmission power to w^{th} WSO on J channels is then defined as,

$$\mathbf{p}_w^c = (p_{w,1}^c, \dots, p_{w,J}^c), \forall w \in \mathcal{W}^c.$$

Let B_w^c be the bandwidth demand of w^{th} WSO. The number of channels required by w^{th} WSO is then calculated as,

$$n_w^c = \frac{B_w^c}{b_j}, \forall w \in \mathcal{W}^c, \forall c \in \mathcal{C}$$

where b_j represents the channel bandwidth. Let $O_{w,j}^c$ translates to a timeslot, here called as channel occupancy time (COT) in a window time, such that the w^{th} WSO registered in c^{th} CM can achieve its desired channel occupancy in the allocated j^{th} channel. The relation of COT to a channel window time is shown in Fig. 2 where three WSOs are scheduled in the window time in a single TV channel. The COTs of w^{th} WSO in J TV channels are then represented as,

$$\mathbf{o}_w^{rc} = (O_{w,1}^c, \dots, O_{w,J}^c).$$

The channel demand set of w^{th} WSO is then defined as follows,

$$\{\mathbf{s}_w^{rc}, \mathbf{p}_w^{rc}, n_w^c, \mathbf{o}_w^{rc}\}, \forall c \in \mathcal{C}, \forall w \in \mathcal{W}^c \quad (5)$$

The channel demand set of c^{th} CM is then defined using channel demands of its registered WSOs as follows,

$$\mathcal{D}^c = \{\mathbf{s}^c, \mathbf{p}^c, N^c, \mathbf{o}^c\}, \forall c \in \mathcal{C}, \quad (6)$$

where $\mathbf{s}^c = (\mathbf{s}_1^c, \dots, \mathbf{s}_W^c)$, $\mathbf{p}^c = (\mathbf{p}_1^c, \dots, \mathbf{p}_W^c)$, $N^c = (n_1^c, \dots, n_W^c)$ and $\mathbf{o}^c = (\mathbf{o}_1^c, \dots, \mathbf{o}_W^c)$. Let $\mathbf{S} = (\mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^C)^T$, $\mathbf{P} = (\mathbf{p}^1, \mathbf{p}^2, \dots, \mathbf{p}^C)^T$, $\mathbf{N} = (N^1, N^2, \dots, N^C)^T$, $\mathbf{O} = (\mathbf{o}^1, \mathbf{o}^2, \dots, \mathbf{o}^C)^T$, the system parameter \mathcal{D} is then defined using the channel demands of all neighboring CMs as follows,

$$\mathcal{D} = \{\mathbf{S}, \mathbf{P}, \mathbf{N}, \mathbf{O}\}. \quad (7)$$

The system in (1) then executes the channel allocation algorithm, as will be discussed in Section V, to allocate TV channels to the WSOs registered in the neighboring CMs such that the allocation satisfies the required system QoS performance. The system QoS performance is preserved if the following allocation condition is satisfied,

$$\sum_{c \in \mathcal{C}} \sum_{w \in \mathcal{W}^c} O_{w,j}^c \leq T_j, \forall j \in \mathcal{J} \quad (8)$$

where T_j refers to the window time in a j^{th} channel. The algorithm proposed in Section V solves the TVWS sharing optimization problem, as will be defined in (14) and outputs a channel allocation matrix \mathbf{X} , defined as follows. Let $x_{w,j}^c \in \{0,1\}$, be a binary decision variable such that if $x_{w,j}^c = 1$, the j^{th} channel is allocated to the w^{th} WSO registered in c^{th} CM; otherwise $x_{w,j}^c = 0$. The allocation status of WSOs registered in the neighboring CMs is then represented by a matrix \mathbf{X} as,

$$\mathbf{X} := \begin{bmatrix} x_{1,1}^1 & x_{1,2}^1 & \dots & x_{1,J}^1 \\ \vdots & \vdots & \ddots & \vdots \\ x_{W^1,1}^1 & x_{W^1,2}^1 & \dots & x_{W^1,J}^1 \\ x_{1,1}^2 & x_{1,2}^2 & \dots & x_{1,J}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{W^c,1}^c & x_{W^c,2}^c & \dots & x_{W^c,J}^c \end{bmatrix}, \quad (9)$$

where $W^c = |\mathcal{W}^c|$, $\forall c \in \mathcal{C}$, i.e., the number of WSOs registered in the c^{th} CM. The w^{th} row in the \mathbf{X} represents the channels allocation status, in the set \mathcal{J} , to the w^{th} WSO registered in c^{th} CM. The j^{th} column in the \mathbf{X} represents the channels allocation status of all the WSOs, from all the CMs in the set \mathcal{C} . The

allocation matrix \mathbf{X} thus orthogonalizes WSOs, registered in the neighboring CMs, in a joint frequency-time domain. The WSOs scheduled on different channels can transmit at the same time using their respective allotted channel (frequency slot) while WSOs scheduled on the same channel can transmit in their respective time slot (here COT).

The system in (1) thus, implements the TVWS sharing problem, defined in Section III-B, as an optimization problem, as discussed in the following section.

B. Problem Formulation

In this section, the proposed TVWS sharing problem is formulated as an optimization problem using well-established proportional fairness method. It is because the proportional fairness is considered one of the most suitable methods to achieve a trade-off between two competing interests [31], [32], [33]. Originally, Kelly defined the proportional fairness as an adjustment process which adjusts the rates of users according to the charges they pay. The proportional fairness method thus was defined for elastic traffic in computer network services [34]. Similarly, in the channel sharing literature, a proportionally fair allocation mostly has been achieved by adjusting the rates of the users based upon some performance criteria like maximizing the resource utilization, etc. [35], [36]. However, applying the proportional fairness in its original to model the TVWS sharing problem proposed in this paper is not suitable. It is because, the third objective in the problem defined in Section III-B makes the resource allocation as binary decision allocation, i.e., a channel is either allocated to a WSO, $x_{w,j}^c = 1$ or not $x_{w,j}^c = 0$. Therefore, WSO allocation (here COT) adjustment is not possible. Consequently, we rewrite the proportional fairness in a binary decision allocation perspective as follows.

Let the maximum data rate the w^{th} WSO can achieve on j^{th} channel be defined by using Shannon channel capacity formula,

$$r_{w,j}^c = b_j \log(1 + SINR_{w,j}^c). \quad (10)$$

The maximum rate $r_{w,j}^c$, $\forall w \in \mathcal{W}^c$ is then used to defined a utility function as a normalized rate achieved by c^{th} CM in j^{th} channel as follows,

$$\mathcal{U}_{c,j} = \sum_{w \in \mathcal{W}^c} \frac{x_{w,j}^c r_{w,j}^c}{O_{w,j}^c + \delta_{O_{w,j}^c, 0}} \quad (11)$$

where $\delta_{O_{w,j}^c, 0}$ defines Kronecker delta function as:

$$\delta_{O_{w,j}^c, 0} := \begin{cases} 1, & \text{if } O_{w,j}^c = 0, \\ 0, & \text{otherwise.} \end{cases}$$

This function prevents denominator term in (11) from becoming zero. The utility function in (11) measures the worth of the resource (channel) to c^{th} CM, i.e., given a channel is allocated to the WSOs in the c^{th} CM for the duration of $\sum_{w \in \mathcal{W}^c} O_{w,j}^c$, how does it translate for the CM in terms of the achieved throughput. In other words, maximizing the function in

(11) shall prefer a CM with WSOs achieving high data rate and lower channel occupancy demand over a CM with WSOs achieving low data rate and high channel occupancy demand. Such preference based allocation shall lead to an efficient use of the resources (TVWS). The distribution $\mathbf{U} = [\mathcal{U}_{c,j}]_{c \times J}$ is then said to be proportionally fair if it is feasible and for all other feasible solutions $\mathbf{V} = [v_{c,j}]_{c \times J}$, the following holds [34],

$$\sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} \frac{v_{c,j} - \mathcal{U}_{c,j}}{\mathcal{U}_{c,j}} \leq 0. \quad (12)$$

It has been shown in [34], [37] that the rates achieved by users become proportionally fair if the sum of logarithmic rates obtained is optimized. Moreover, it is shown in [38] that if all rates are proportionally fair, they maximize the throughput over all other feasible throughputs. Therefore, if the logarithmic sum of the utility function in (11) is maximized, the normalized rate achieved by neighboring CMs shall become proportionally fair. Let a j^{th} channel is said to be allocated to the c^{th} CM if at least one of its registered WSO is scheduled on the channel. The allocation status of the channels in the \mathcal{J} , to the c^{th} CM, is then defined as follows,

$$\mathbf{x}^c := \begin{bmatrix} x_{1,1}^c & x_{1,2}^c & \cdots & x_{1,J}^c \\ \vdots & \vdots & \ddots & \vdots \\ x_{W^c,1}^c & x_{W^c,2}^c & \cdots & x_{W^c,J}^c \end{bmatrix}. \quad (13)$$

Let $\mathbf{1} = (1, 1, \dots, 1)_{1 \times J}$. Let $\mathbf{O}_j \in \mathbf{O}$ be the j^{th} column vector in COT demand matrix in the system parameter \mathcal{D} , defined as, $\mathbf{O}_j = (O_{1,j}^1, O_{2,j}^1, \dots, O_{W^1,j}^1, O_{1,j}^2, \dots, O_{W^c,j}^c)^T$ where $W^c = |\mathcal{W}^c|, \forall c \in \mathcal{C}$. Let $\mathbf{X}_j \in \mathbf{X}$ represents the j^{th} column vector of the allocation matrix \mathbf{X} . The TVWS sharing problem is then defined as follows,

$$\max. \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} \log(\mathcal{U}_{c,j} + 1) \quad (14a)$$

$$\text{subject to } \mathbf{x}^c \leq \mathbf{Z}^c, \quad \forall c \in \mathcal{C}, \quad (14b)$$

$$\mathbf{X}_j^T \mathbf{O}_j \leq T_j, \quad \forall j \in \mathcal{J}, \forall c \in \mathcal{C}, \quad (14c)$$

$$\mathbf{x}^c \mathbf{1}^T \leq (N^c)^T, \quad \forall c \in \mathcal{C}, \quad (14d)$$

$$\mathbf{x}^c \in \{0, 1\}, \quad \forall c \in \mathcal{C}. \quad (14e)$$

The constraint in (14b) ensures that a channel can be allocated to the WSOs registered in c^{th} CM only if the channel is available in their respective region, i.e., $x_{w,j}^c \in \mathbf{x}^c = 1$ iff $z_{w,j}^c \in \mathbf{Z}^c = 1$. The constraint in (14c) ensures that the WSOs scheduled in a j^{th} channel preserve the system QoS performance, as defined in (8), i.e., the total allocated channel occupancy time of coexisting WSOs must preserve the channel window time. The constraint in (14d) ensures that the number of channels allocated to the c^{th} CM is restricted by the number of channels desired by its WSOs. Finally, (14e) forces the decision variable to be binary valued.

The constraints in (14e) and (14c) helps the system in (1) to satisfy the third objective of TVWS sharing problem in Section III-B. The optimization problem in (14) seeks to optimize a concave objective function over a convex set. The problem in (14) has a unique solution, as from the optimization theory [39], maximizing a concave function over a convex set has a unique solution. A solution approach to the problem in (14) is presented in the following section.

V. SOLUTION METHOD

The nonlinear objective function (14a) and binary-valued constraint (14b) makes the problem in (14) a nonlinear combinatorial optimization problem. Determining the optimal solution of such a problem is a challenging task as the problem becomes intractable as the number of discrete variables increases [40]. Therefore, to ease the solution approach, the problem in (14) is transformed into a linear programming problem with relaxed binary constraint.

A. Linearization

The objective function (14a) is linearized using a piecewise linear approximation. In this process, tangent line approximation is used to approximate the objective function in (14a), denoted as, F . The detailed description of linear approximation is provided in Appendix A. Using this function, the problem in (14) is linearized as,

$$\max. \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} F(\mathcal{U}_{c,j}) \quad (15a)$$

$$\text{subject to } \mathbf{x}^c \leq \mathbf{Z}^c, \quad \forall c \in \mathcal{C} \quad (15b)$$

$$\mathbf{X}_j^T \mathbf{O}_j \leq T_j, \quad \forall j \in \mathcal{J}, \forall c \in \mathcal{C} \quad (15c)$$

$$\mathbf{x}^c \mathbf{1}^T \leq (N^c)^T, \quad \forall c \in \mathcal{C} \quad (15d)$$

$$\mathbf{x}^c \in \{0, 1\}, \quad \forall c \in \mathcal{C} \quad (15e)$$

To tackle the binary-valued constraint (15b), we apply Lagrangian relaxation as explained followings.

B. Lagrangian Relaxation

Lagrangian relaxation [41] relaxes a subset of constraints by adding them to the objective function with a penalty term called the Lagrangian multiplier. Let $\boldsymbol{\lambda} := [\lambda_{w,j}]_{W \times J}$ be the Lagrangian multipliers matrix. Then, the relaxed problem can be defined as,

$$\max_{\mathbf{X}}. P(\mathbf{X}, \boldsymbol{\lambda}) = \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} F(\mathcal{U}_{c,j}) + \boldsymbol{\lambda}^T (\mathbf{Z}^c - \mathbf{x}^c) \quad (16a)$$

$$\text{subject to: } \mathbf{X}_j^T \mathbf{O}_j \leq T_j, \quad \forall j \in \mathcal{J}, \forall c \in \mathcal{C} \quad (16b)$$

$$\mathbf{x}^c \mathbf{1}^T \leq (N^c)^T, \quad \forall c \in \mathcal{C} \quad (16c)$$

$$\mathbf{x}^c \in \{0, 1\}, \quad \forall c \in \mathcal{C} \quad (16d)$$

For a given $\boldsymbol{\lambda}$, the Lagrangian relaxation can be defined as,

$$h(\boldsymbol{\lambda}) = \max_{\mathbf{X}} \{P(\mathbf{X}, \boldsymbol{\lambda}) : \text{constraints (16b), (16c), (16d)}\} \quad (17)$$

Then the generalized dual problem of the relaxed problem is defined as followings,

TABLE 2
ALGORITHM: DUAL PROBLEM BASED ON LAGRANGIAN RELAXATION

Step 0:	a) Choose initial values of λ^0 . b) Set parameters, for example, $\rho = 2.0, \rho^{\min} = 0.001, \rho_{iter} = 0, \rho^{\max iter} = 5, k = 0,$ $k^{\max} = 10, F^{\text{best}} = 0, h^{\text{best}} = -\infty, h^{\text{upper}} = 0, \mathbf{X}'_k = [0]_{W \times J}$.
Step 1:	a) Increment as $k = k + 1, \rho_{iter} = \rho_{iter} + 1$ b) Given λ^k , solve the relaxed problem using any linear programming technique and obtain \mathbf{X}_k .
Step 2:	Validate \mathbf{X}_k as: set $x_{w,j}^c := 0$ if $z_{w,j}^c = 0$.
Step 3:	Perform frequency reuse as in Table 3 and get \mathbf{X}'_k .
Step 4:	Use \mathbf{X}'_k to compute the value of the function in (16a), called as F , and fairness index value H in (20). If $F > F^{\text{best}}$: $F^{\text{best}} = F$, $h^{\text{upper}} = F^{\text{best}}$ and $\mathbf{X} = \mathbf{X}'_k$.
Step 5:	a) Use \mathbf{X}'_k to compute: <ul style="list-style-type: none"> - Subgradient vector as, $\nabla h(\lambda^k) = \left[\frac{\partial h}{\partial \lambda_{w,j}^k}, \forall w \right]$, - Dual objective in (18), - Step size as, $t_k = \frac{\rho(h^{\text{upper}} - h(\lambda^k))}{\ \nabla h(\lambda^k)\ ^2}$. b) Update the dual variable as, $\lambda^{k+1} = \max\{\lambda^k + t_k \nabla h(\lambda^k), 0\}$
Step 6:	If $h^{\text{best}} < h(\lambda^k)$ then $h^{\text{best}} = h(\lambda^k)$ else if $\rho_{iter} > \rho^{\max iter}$ then $\rho = \max\{\frac{\rho}{2}, \rho^{\min}\}$ and $\rho_{iter} = 0$.
Step 7:	If $t_k < 0.001$ or $k > k^{\max}$ stop; otherwise, go to Step 1.

$$L^* = \min_{\lambda} \{h(\lambda) : \lambda \geq 0\}. \quad (18)$$

The solution to (17) is the upper bound of the solution to the original problem (16). Note that (17) is a concave function. For a concave function, a gradient-based approach is generally used to compute a value as close as desired to the optimal value. Thus, if h would have been differentiable, we can use a gradient descent method to have a convergence toward the optimal value. The proposed problem, however, cannot be solved using a gradient descent method. It is because the objective function is piecewise linear which is non-differentiable at the intersection point of adjacent linear pieces, but sub-differentiable at this point. The subdifferential of $h(\lambda)$ at such a point is the set of all subgradients at that point. Thus, we need to compute a sequence of $\{\lambda^k\}_{k \in \mathbb{N}}$ such that either $h(\lambda^k)$ converges to the optimal solution using the subgradient method, which is given in the following dual algorithm. The convergence property of the subgradient algorithm is presented in Appendix B.

C. Subgradient Algorithm for Lagrangian Relaxation based TVWS Sharing Problem

The algorithm defined in Table 2, can be described as follows. In Step 0, the input parameters to the algorithm are defined as follows. The initial values of λ^0 are defined randomly. The parameter ρ is used in defining step size t^k ,

defined in the range $\rho^{\min} < \rho \leq 2$ [41]. The ρ_{iter} with upper limit of $\rho^{\max iter}$ counts the number of iterations after which the parameter ρ is updated. The k^{\max} is defined as stopping criteria for the algorithm.

The algorithm uses variables initialized in Step 0 to apply a linear programming (LP) solver to solve the dual problem and obtain the k^{th} iteration allocation matrix \mathbf{X}_k . LP solvers are available on both the commercial and freeware basis. The entries in \mathbf{X}_k are then adjusted based upon the corresponding entries in \mathbf{Z}^c such that $x_{w,j}^c \in \mathbf{x}_k^c, \forall \mathbf{x}_k^c \in \mathbf{X}_k$ are set equal to zero if the corresponding element, $z_{w,j}^c \in \mathbf{z}_w^c, \forall \mathbf{z}_w^c \in \mathbf{Z}^c$ is zero. This validation ensures the constraint in (14b).

The algorithm then applies the FR process in Step 3 in Table 2. In this process, the algorithm makes use of the current allocation vector, \mathbf{X}_k and interference matrix, as shall be discussed in Section IV-D, to identify a set of WSOs which do not get the channel. The algorithm then repeatedly applies LP solver to performs channel allocation to the unallocated WSOs such that they do not cause interference to the allocated WSOs of neighboring CMs. The FR process is detailed in Section V-E. The outcome of FR process is an updated allocation matrix \mathbf{X}'_k which is then used to compute the function values in (16a) and the fairness in allocation among neighboring CMs.

Several fairness measures or metrics are used in the literature to determine whether networks are receiving a fair share of spectrum or not. For example, max-min fairness, Jain's fairness index, fairly shared spectrum efficiency, worst-case fairness. In this paper, we adopt Jain's fairness index [42] to measure fairness in allocation among neighboring CMs. The reason is that it satisfies the desired properties of fairness measure like population size independence, continuity etc., as listed in [43]. These properties are important to be considered in measuring the fairness in allocation. For example, the continuity property shows any slight change in the allocation of individual WSO. Thus, an inefficient use of the TVWS is identified by the fairness index as a WSO with bad channel characteristics gets a high proportion of the spectrum. It is ensured through the use of the continuous allocation metric like fraction of throughput demand, as defined in (19). Such an allocation metric is suitable to measure the fairness in allocation for the case where WSOs demand unequal channel bandwidth [43]. Therefore, based on the fraction of throughput demand of CMs, an allocation metric is defined as follows,

$$T^c = \frac{d^c}{d'^c}, \forall c \in \mathcal{C}, \quad (19)$$

where d^c and d'^c represents the maximum data the c^{th} CM desire to transmit and it can transmit using its allocated channels, respectively. These terms are defined as follows. Let the maximum data the c^{th} CM can transmit using its allocated channels is defined in terms of the data the WSOs registered in it can transmit, defined as follows.

$$d^c = \sum_{j \in \mathcal{J}} \sum_{w \in \mathcal{W}^c} x_{w,j}^c O_{w,j}^c r_{w,j}^c, \forall c \in \mathcal{C}. \quad (20)$$

Note that channels are considered as additive white Gaussian noise (AWGN). The data the CM desires to transmit is defined as,

$$d^{lc} = \sum_{j \in \mathcal{J}} \sum_{w \in \mathcal{W}} O_{w,j}^c \bar{r}_{w,j}^c, \forall c \in \mathcal{C}.$$

The normalized throughput vector (T^1, \dots, T^C) is then adopted to measure fairness in allocation using Jain's fairness index [42] as,

$$H(T^1, T^2, \dots, T^C) = \frac{\left(\sum_{c \in \mathcal{C}} T^c \right)^2}{C \sum_{c \in \mathcal{C}} (T^c)^2}. \quad (21)$$

Function H in (21) outputs a value in the range of $[0, 1]$; when the value is closer to 1, the allocation is deemed fairer.

If the current iteration value of the objective function, F , is optimal, then F^{best} is updated with F and \mathbf{X} with \mathbf{X}'_k . As the iteration progresses, the feasible primal F^{best} and lower bound h^{best} approach gradually to the integer optimal by adjusting λ^k using the subgradient method as defined in Step 5. In Step 5, the sub-gradient vector of the objective function and the Lagrangian multiplier vector λ^k for the k^{th} iteration are calculated. The step size t_k is used to calculate the multiplier vector for the next iteration. The Lagrange multipliers are thus adjusted iteratively. The convergence property of the subgradient algorithm is discussed under Appendix B. The algorithm terminates as one of the termination conditions satisfied:

- Dual step size becomes less than a set threshold or,
- the number of iterations exceeds the maximum number of iterations.

After the overall iteration ends, we regard the final value of F^{best} as the approximated optimal solution and the corresponding allocation matrix \mathbf{X} is the algorithm output.

The interference matrix, \mathbf{Y} , that is used to implement the FR step in Table 3 is defined in the following section.

D. Interference Matrix

The WSOs registered in the neighboring CMs and interfering on the available TV channels is represented using an interfering matrix called as Y-matrix in this paper. Note that the Y-matrix does not model the interference among coexisting WSOs. Rather, it represents the set of WSOs which cannot transmit simultaneously on the available TVWS due to interfering transmission regions. In fact, in IEEE 802.19.1 [10], a coexistence discovery algorithm is presented that the CDIS and CM run to perform the statistical analysis of the expected interference among coexisting WSOs. Briefly, the algorithm in [10] takes the WSOs' geographic location, transmitter and receiver characteristics, antenna height and directivity, height above average terrain and other related parameters to execute interference discovery process. In this process, a cumulative distribution function of the potential interference from m^{th} WSO to w^{th} WSO is estimated. Both of these, m^{th} and w^{th} WSOs, could register to the same CM or different CMs in the system. The minimum interference level, experienced by 90%

devices of the w^{th} WSO, is then taken as the potential interference value from an m^{th} WSO to w^{th} WSO. The measured interference value is then compared to a threshold. If the value is greater than the threshold, the m^{th} WSO is considered potential interferer to the w^{th} WSO and is included in its CS. A similar rule is applied for interference discovery of the w^{th} WSO into the m^{th} WSO. Thus, the outcome of the interference analysis process is a CS of each WSO registered in the CMs in the system. The system in (1) then makes use of the CS of each WSO to generate a Y-matrix as follows.

Let a set $\mathcal{S}_{w,j} = \{I_{w,m}(j)\}, \forall m \in \mathcal{W}$, be an encoded CS of w^{th} WSO on a j^{th} channel such that an indicator variable $I_{w,m}(j) = 1$ if m^{th} WSO interferes w^{th} WSO transmission on the j^{th} channel, as defined in Table 1; otherwise $I_{w,m}(j) = 0$. The encoded CS of all the WSOs coexisting on j^{th} channel are then used to define a j^{th} channel interference matrix $\mathbf{y}(j)$ as follows,

$$\mathbf{y}(j) := \begin{bmatrix} \times & I_{1,2}(j) & \cdots & I_{1,W}(j) \\ & & & \vdots \\ I_{W,1}(j) & I_{W,2}(j) & \cdots & \times \end{bmatrix} \quad (22)$$

where \times in diagonal vector in $\mathbf{y}(j)$ represents don't care condition. This condition translate a self-interference indicator variable, $I_{w,w}(j)$, having no meaning. The w^{th} row in $\mathbf{y}(j)$ matrix represents encoded CS of w^{th} WSO. The interference matrices for all channels in the system are then used to define an interference matrix \mathbf{Y} as follows,

$$\mathbf{Y} = [\mathbf{y}(1) \ \mathbf{y}(2) \ \cdots \ \mathbf{y}(J)] \quad (23)$$

The TVWS sharing algorithm, in Table 2 makes use of the interference matrix \mathbf{Y} to implement FR in sharing TVWS among *heterogeneous* WSOs, as discussed in the following subsection.

E. Frequency Reuse

The frequency reuse (FR) subroutine in Table 3 performs spatial reuse of the TV spectrum to enhance its effective utilization. The FR process is implemented to the WSOs do not getting channel in the initial allocation phase in Step 1, Table 2. This requires to identify a set of unallocated WSOs eligible for the FR. In this process, an encoded CS $\mathcal{S}_{w,j}, \forall m \in \mathcal{W}$ and an interference matrix \mathbf{Y} are used to define the set of unallocated WSOs, \mathcal{W}' . To generate encoded CS and Y-matrix, we make use of the CS of each WSO available at MCM. Note that the 802.19.1 defines different message clauses that enable CMs to exchange their WSO related information [10]. Let us assume the CS of WSOs are available to CDM at MCM. Given such information available, an encoded CS of WSOs, $\mathcal{S}_{w,j}, \forall m \in \mathcal{W}$ and an interference matrix \mathbf{Y} , are generated, as defined in Section V-D. Initially the Y-matrix is filled with all ones. Let \mathbf{X}_k be an initial allocation matrix available from Step 2, Table 2. The Y-matrix is then updated based on the \mathbf{X}_k and $\mathcal{S}_{w,j}, \forall m \in \mathcal{W}$ in

TABLE 3
SUBROUTINE: FREQUENCY REUSE

Input:	$\lambda^k, \mathbf{X}'_k = \mathbf{X}_k, \mathcal{Z}, \text{CS}$.
Output:	\mathbf{X}'_k
Step 0:	Given CS generate encoded CS, i.e., $\mathcal{S}_{w,j}, \forall w \in \mathcal{W}, \forall j \in \mathcal{J}$ and interference matrix \mathbf{Y} , as defined in Section V-D.
Step 1:	Given \mathbf{X}'_k , update $\mathbf{y}(j) \in \mathbf{Y}, \forall j \in \mathcal{J}$ as: For each w^{th} WSO do: if $x_{w,j} = 1: I_{w,m}(j) = 0, \forall m \in \mathcal{W}, \forall j \in \mathcal{J}$ or if $(x_{m,j} = 1 \text{ and } w \in \mathcal{S}_{m,j}): I_{w,m}(j) = 0, \forall m \in \mathcal{W}$.
Step 2:	Define unallocated WSO set in the system as, $\mathcal{W}' = \left\{ \forall w \in \mathcal{W} : \exists j \in \mathcal{J} \mid \sum_{m \in \mathcal{W}} I_{w,m}(j) > 0 \right\}$.
Step 3:	While $\sum_{w \in \mathcal{W}} \sum_{m \in \mathcal{W}} I_{w,m}(j) > 0, \forall j \in \mathcal{J}$ and $\mathcal{W}' \neq \{\}$ do a) Given λ^k , and \mathcal{W}' ; solve the relaxed problem using any linear programming solver and obtain \mathbf{X}'_k . b) Perform following updates: 1) Update \mathbf{X}'_k as, $x_{w,j}^c := 0$ if $z_{w,j}^c = 0$. 2) Update \mathbf{X}'_k as, $\mathbf{X}'_k = \mathbf{X}'_k + \mathbf{X}_k$. 3) Update \mathcal{W}' as, $\mathcal{W}' \leftarrow \mathcal{W}' \setminus \{ \forall w \in \mathcal{W}' \mid \exists j \in \mathcal{J} : x_{w,j}^c = 1 \}$. 4) Update \mathbf{Y} as in Step 1.

Step 1, Table 3, as follows. For each j^{th} channel in the system, update interference matrix $\mathbf{y}(j) \in \mathbf{Y}$ as,

- 1) If j^{th} channel is allocated to w^{th} WSO, set all w^{th} row elements in $\mathbf{y}, \forall \mathbf{y} \in \mathbf{Y}$ equal to zero, or
- 2) If j^{th} channel is allocated to m^{th} WSO and w^{th} WSO is in the CS of m^{th} WSO, set all w^{th} row elements in the matrix \mathbf{y} equal to zero.

The above two steps identify the eligibility of the WSOs for implementing the FR process. For example, if the w^{th} WSO is already allocated a channel, we aim to restrict it in taking part the FR process. Therefore, the w^{th} row entries in the entire Y-matrix are flipped zero in the first step above. Similarly, if a j^{th} channel is already allocated to m^{th} WSO and if w^{th} WSO transmission in the j^{th} channel shall create harmful interference to the m^{th} WSO transmission, the j^{th} channel cannot be spatially reused at unallocated w^{th} WSO. Therefore, Y-matrix entries corresponding to w^{th} row are also flipped zero. The updated Y-matrix thus defines a set of unallocated WSOs. These are the WSOs for which at least one nonzero entry exists in the corresponding row in the Y-matrix, as defined, in Step 2, Table 3.

The subroutine in Step 3, Table 3 then repeatedly allocates the available TV channels to the WSOs in the set \mathcal{W}' as follows. The relaxed problem in (17) is solved using any LP solver for the WSOs in the set \mathcal{W}' and an allocation matrix \mathbf{X}'_k is obtained. The \mathbf{X}'_k is then used to update $\mathbf{X}'_k, \mathcal{W}'$, and Y-matrix, as defined in Step 3-b)2), 3-b)3), and 3-b)4), respectively. This repetitive update and allocation process continues until all WSOs in the set \mathcal{W}' get the channel or no more FR is possible.

Let us apply the FR implementation in the coexisting scenario shown in Fig 3. In this figure, four WSOs operating in three network technologies, an IEEE 802.22 regional area

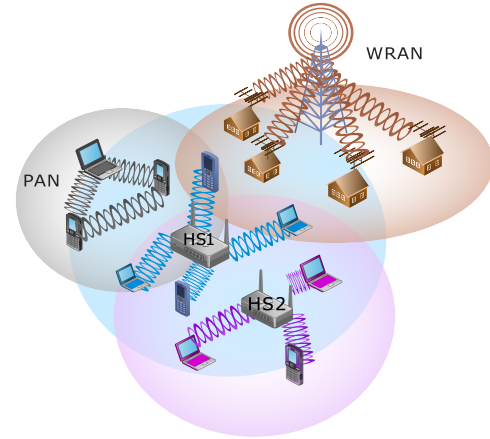


Fig. 3. IEEE 802.22 wireless regional area network (WRAN), IEEE 802.11 hotspots (HS1, HS2), and IEEE 802.15.4 personal area network (PAN) coexisting in some geographic region.

network, IEEE 802.11 local area networks and IEEE 802.15.4 personal area network are deployed in some geographic region. The shaded area around each transmitter denotes its transmission radius. The circular links between a transmitter and receivers show wireless connectivity between them. The receiver nodes in some networks receive interfering signals from other collocated transmitters as shown in the figure. Let WRAN, HS1, HS2, and PAN are labelled as, WSO 1, 2, 3 and 4, respectively. Let us assume each of the WSO is registered in a dedicated CM, i.e., four neighboring CMs are available in the CDM system. Let us suppose that a single TV channel is available in the region for secondary use. Then, based on coexisting scenario shown in the figure, the encoded CS of each WSO can be defined as follows.

$$\mathcal{S}_{1,1} = \{0, 1, 0, 0\}, \mathcal{S}_{2,1} = \{1, 0, 1, 1\}, \mathcal{S}_{3,1} = \{0, 1, 0, 0\}, \mathcal{S}_{4,1} = \{0, 1, 0, 0\}.$$

The Y-matrix is then populated from the bitwise OR operation on the CS of the WSOs. The generated Y-matrix is $\mathbf{Y} = [1 \ 1 \ 1 \ 1]$. Let for some given input parameters, as listed in Table 1, the algorithm in Table 2 finds an initial allocation vector, $\mathbf{X} = [1 \ 0 \ 1 \ 0]$. The allocation vector shows WSO 1 and WSO 3 are allocated the channel. The FR process is then invoked. The Y-matrix is updated to identify WSOs eligible for spatially reusing the channel, as follows. The XOR operation is performed as, $(\mathbf{Y} = \mathbf{X} \oplus \mathbf{Y})$. This operation turns the entries in Y-matrix equal to zero where the corresponding entries in X-matrix are ones. The Y-matrix at this stage looks like, $\mathbf{Y} = [0 \ 1 \ 0 \ 1]$. It is then updated using the CS of allotted WSOs as previously defined in the second rule of Y-matrix update. The second entry in Y-matrix is thus flipped zero as WSO 2 is in the CS of allotted WSO 1. The updated Y-matrix then looks like, $\mathbf{Y} = [0 \ 0 \ 0 \ 1]$. The algorithm then solves the dual problem again and allocates the channel to WSO 4. The final allocation matrix then looks like $\mathbf{X} = [1 \ 0 \ 1 \ 1]$. The final allocation shows that the available TV channel is reused at WSO 4 without causing harmful interference to allotted WSO 1 and WSO 3.

F. Scheduling Map

Once the allocation process in Table 2 and frequency reuse in Table 3 terminates, the CDM system generates a scheduling map to send it to the CMs in the system. The scheduling map (SM) is a map showing the WSOs' scheduling periods arranged in window time in the allocated channels. In this paper the

scheduling period of a w^{th} WSO refers to its channel timeslot, i.e., COT. For example, SM of three WSOs scheduled in an allocated TV channel is shown in terms of their COT defined in the window time in Fig. 2. Thus, given the COT of WSOs and the allocation matrix \mathbf{X} , from the algorithm in Table 2, the SM is a simple procedure of defining two timing parameters; transmission start time and transmission end time. The CDM system defines the timing parameters for WSOs registered in the CMs in the system as follows.

Let a pair of transmission variables, $(t_{w,j}^{\text{start}}, t_{w,j}^{\text{stop}})$, precisely define the time instance the w^{th} WSO, registered in c^{th} CM, may start and stop its transmission on an allotted j^{th} channel, respectively. The $t_{w,j}^{\text{start}}$ and $t_{w,j}^{\text{stop}}$ are calculated as follows. Let a variable $C_{w,m(w)}$ be defined as the cost of sharing a channel between two WSOs, $w, m \in \mathcal{W}$, where $m(w)$ represents a WSO m sharing a channel with WSO w . Let τ_w represents the control overhead associated with MAC technology of the w^{th} WSO. The control overhead is defined as the amount of time required to perform control signaling while operating in the TVWS. This value is fixed and predetermined based upon the underlying network technology of the WSO. For example, if an 802.22 WSO employs OFDMA, one OFDM symbol is used for both the frame preamble and the frame header; except for the first frame in the superframe which consumes two additional symbols (1/4 cyclic prefix mode). If we consider two OFDM symbols per frame as a control region then using a symbol duration, $T_{\text{Sym}}=0.3733$ ms [44], the control overhead per frame is computed as, 0.7466 ms. Other settings may generate different overhead. Similarly, if a WSO m operates in a different network technology than that of the WSO w , its control overhead will be different from that of WSO w . The total overhead in a channel varies as the channel is shared among heterogeneous WSOs. The value of the parameter $C_{w,m(w)}$ is then defined simply by adding the control overhead of all WSOs sharing a channel as follows:

$$C_{w,m(w)} := \begin{cases} \tau_w + \tau_m & \text{if } MAC_w \neq MAC_m, \forall (w, m) \in \mathcal{W}^c \\ 0 & \text{otherwise} \end{cases}, \quad (24)$$

where $\mathcal{W}^c \subset \mathcal{W}^x$ refers to the set of WSOs with NID listed before NID of w^{th} WSO in \mathcal{W}^c . The timing parameters are computed as,

$$t_w^{\text{start}} = \sum_{m \in \mathcal{W}^c} O_{m,j}^c x_{m,j}^c + C_{w,m(w)} \quad \text{and} \quad t_w^{\text{stop}} = t_w^{\text{start}} + O_{w,j}^c. \quad (25)$$

Thus, the $t_{w,j}^{\text{start}}$ refers to the time instance in the scheduling window that all the WSOs m have utilized the channel for the duration of their respective COT. Note that in defining the scheduling map we make a simplifying assumption that the timers of WSOs in the system are pre-synchronized and WSOs sharing a j^{th} channel have agreed on the reference time (the time instance the window time starts) as defined in [10]. Timer synchronization may be done by having agreements between service providers managing the WSOs which is outside the scope of this paper.

The CDM defines SM and send it to the SCMs. The SCMs send the SM to the registered WSOs. Such implementation

shall reduce the control signaling between the WSOs and the pertinent CM. The control signaling is otherwise inevitable while performing context switching among WSOs scheduled in the TV channel. Once the spectrum has been allocated, the SM remains unchanged unless i) an incumbent appears in one of the assigned channels ii) a change in a WSO's channel occupancy demand or some other coexisting WSO's demand requires readjusting the WSO's allocation.

VI. SIMULATIONS AND ANALYSIS

The performance of the proposed channel sharing algorithm is compared with two other channel allocation algorithms, proposed in [18] and [17].

A. Comparative Channel Allocation Schemes

In this section, we summarize the allocation mechanism of the comparative TVWS allocation schemes. In [17], two TVWS sharing problems are defined; one for maximizing the number of channels allocated to the networks and the second for maximizing the total throughput under the minimum fairness constraint of allocating at least a single channel to each network. In this simulation setup, we implement the second problem as it closely matches with the channel sharing scheme proposed in this paper. The TVWS sharing algorithm proposed in [17] then selects a node (WSO) having a minimum of the assigned channels and the minimum number of the available channels to it. The algorithm assigns a TV channel to the selected WSO and calculates the total throughput. It keeps assigning the channel to other WSOs as long as the total throughput is increasing. This procedure is repeated for every channel. The algorithm terminates as no more increase in the throughput is observed.

The TVWS sharing problem in [18] is modeled as a lexicographic ordering of throughputs of access points of coexisting networks. The proposed problem is then transformed into a graph coloring problem. An algorithm called as, Share, is then proposed to solve the graph coloring problem. The Share algorithm operates in three phases. In the first phase of allocation, it orthogonalizes the WSOs in the available TV channels (frequency slots). In the second phase, a mutual channel sharing is performed among allotted WSOs of the first phase under the condition that their first phase throughputs do not decrease. The fairness is improved in the third phase by sharing the channel with unallocated WSOs such that lexicographically ordered throughputs do not decrease.

We select the algorithms in [18] and [17] due to the close resemblance of their TVWS sharing problems to the proposed channel sharing mechanism. For example, both considers optimizing throughput under minimum fairness in allocation. However, there exist some fundamental differences as well. For example, both the allocation schemes orthogonalize the WSOs in frequency domain by allocating a dedicated channel to each allocated WSO while the proposed scheme orthogonalize WSOs in a joint time-frequency domain by slicing the available TVWS in the frequency bands and further slicing each channel (frequency band) into a number of COTs in the channel window time, as discussed in section IV. Moreover, the algorithm in [17] is intended for TVWS channel allocation to cellular networks

while the proposed scheme is intended for TVWS sharing in an *ad hoc* coexisting environment, as discussed in Section I. Similarly, the TVWS sharing algorithm in [18] does not implement the FR concept. Therefore, we implement the proposed algorithm without FR process as well to have a fair comparison with the scheme in [18]. This is achieved by omitting Step 3 in Table 2 during the implementation of the proposed algorithm.

Finally, the performance of the proposed allocation scheme with and without FR implementation is judged in comparison with the Scheme in [17] and the Scheme in [18], respectively.

B. Simulation Setup

Simulation setup consists of 32 WSOs deployed in some geographic region and connected to an 802.19.1 coexistence system. The system has 32 CMs, each serving a single WSO. We select a dedicated CM for each WSO as the schemes in [18] and [17] performs TVWS sharing at network (WSO) level. The number of available TV channels in the region varies from 2 to 16. The WSO types and transmission powers are modeled using FCC regulations [2]. For this purpose, the specifications for fixed, mode 1 and mode 2 WSO types are used. The fixed, mode 1 and mode 2 type WSOs are allowed to have maximum antenna gain of 4 watts (W) effective isotropic radiated power (EIRP), 100 mW EIRP, and 100 mWatt EIRP respectively. The WSO access technologies are IEEE 802.22 and IEEE 802.11af. In this simulation setup, we implement the compulsory channel requirement of each WSO where the standard definition of the above technologies mandates a single TV channel of regulatory defined bandwidth as a requirement of a device to operate in the TVWS. Note that the bandwidth of a TV channel is set equal to 6 MHz.

Two parameters; WSO channel occupancy demand, $O_{w,j}^c$ and WSO density in the region, $K_{w,j}^c$ are varied to observe their effect on allocation behavior of the three allocation schemes as follows. Let T_j represents the window time on the j^{th} channel. Note that the 802.19.1 [10] does not define MAC layer frame structure for operations in TVWS. Therefore, the channel window time is not defined in an absolute time domain in 802.19.1. In this simulation setup, we define the channel window time as a unit length, without loss of generality, i.e., $T_j = 1, \forall j \in \mathcal{J}$. Then, three allocation subdomains are defined on the T_j as follows; low subdomain consists of up to 33 percent of the channel window time, defined as, $O^L := (0, 0.33]T_j$, A medium subdomain consisting of 34 to 67 percent of the channel window time, defined as, $O^M := [0.34, 0.67]T_j$ and a high subdomain consists of 67 to 100 percent of the channel window time, defined as, $O^H := [0.67, 1]T_j$. The channel occupancy demand of each WSO is then randomly defined on these subdomains.

The WSO density in the region is reflected using the number of WSOs in the CS of each WSO as follows. Let W be the number of WSOs registered in all CMs in the system then, we define three WSO density subdomains as; low $K^L := (0, 0.33]W$, medium $K^M := [0.34, 0.67]W$, and high

$K^H := [0.67, 1]W$. The CS of each WSO is randomly defined on these subdomains. Let $K_{w,j}^c$ represents the number of WSOs in the CS of w^{th} WSO on the j^{th} channel, registered in c^{th} CM. Then, the effect of the variability in the translated channel occupancy demand and WSO density is measured using a pair of parameters $(O_{w,j}^c, K_{w,j}^c)$. Note that varying each of these parameters on three respective subdomains results in $2^3 = 27$ possible allocation combinations. Out of 27, we select three cases to study the performance metrics defined in Section VI-C, as follows.

- Low: low COT, low WSO density, i.e., $O_{w,j}^c \in O^L$ and $K_{w,j}^c \in K^L$,
- Medium: medium COT, medium WSO density, i.e., $O_{w,j}^c \in O^M$ and $K_{w,j}^c \in K^M$, and
- High: high COT, high WSO density, i.e., $O_{w,j}^c \in O^H$ and $K_{w,j}^c \in K^H$.

Next, we apply the `intlinprog` routine of MATLAB® to solve the proposed TVWS sharing problem. The routine applies the mixed-integer linear programming technique. Since we need binary valued vector \mathbf{X} , therefore, we set all the decision variables, $x_{w,j}^c \in \mathbf{X}, \forall \mathbf{X}^c \in \mathbf{X}$, to be integer variables in the `intlinprog` routine. The binary decision may lead to the situation where the COT of allocated WSOs may not fit the channel window time. For example, let us assume the WSO 1, 2, 3 and 4 in Fig. 3 coexist in a TV channel. Let their COT demand is defined as, 0.25, 0.33, 0.37 and 0.15, respectively. Let us assume the `intlinprog` routine outcome as $\mathbf{X} = [1 \ 0 \ 1 \ 1]$, i.e., the WSO 1, 3 and 4 gets the channel. This results in total COT of allocated WSOs equal to 0.77 which is less than the channel window time; 1. The second WSO cannot be accommodated in the channel considering the constraint (16b). In this simulation, the solution \mathbf{X} is engineered such that the second WSO is partially allocated the desired COT so as to maximize the channel utilization while maintaining constraint (16b). The purpose of such engineering the solution is to reduce the channel waste. In order to have a fair comparison, the same engineering principle is applied to the allocation matrix generated by the comparative allocation schemes. The comparative analysis of the three allocation schemes is then performed as discussed in the following section.

C. Comparative Analysis

The relative performance of the three allocation schemes is evaluated using the following metrics: system throughput, fairness in allocation among CMs and WSO satisfaction from the allocation. These performance metrics are selected to analyze how well the three allocation schemes achieve the TVWS sharing objectives, as defined in Section III-B. The simulation results of the performance metrics are presented in Fig. 4 to Fig. 6, respectively. Subplots (a), (b), and (c) in these figures show the effect of varying the $(O_{w,j}^c, K_{w,j}^c)$ pair in low, medium and high subdomains, respectively. The study results

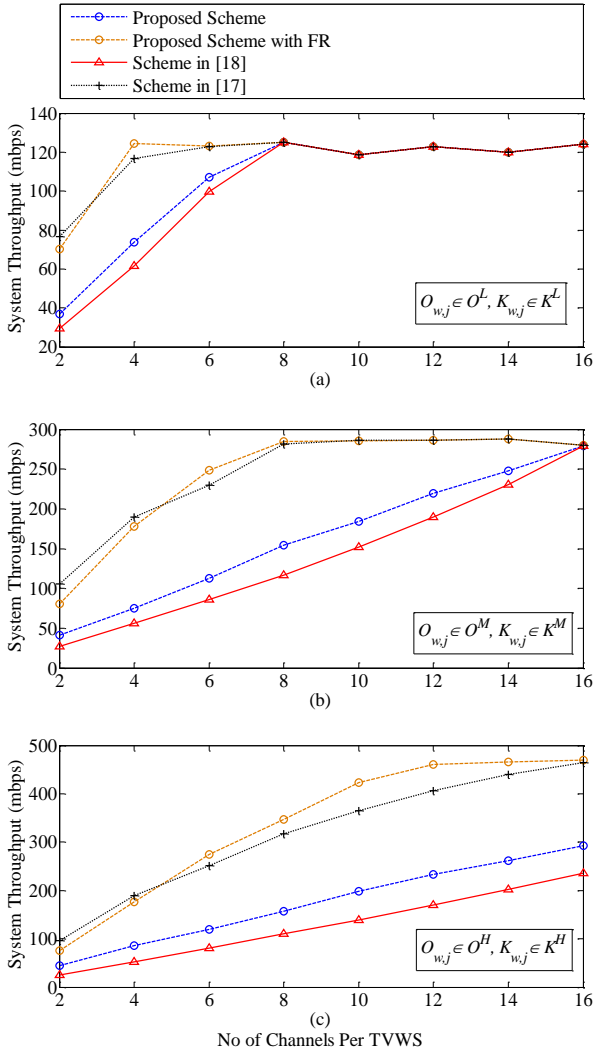


Fig. 4. System throughput for 32 WSOs registered in all CMs for a varying number of TV channels in the system.

are discussed as follows.

1) System Throughput

Fig. 4 shows the system throughput (ST) achieved by the three allocation schemes. Given the allocation matrix \mathbf{X} , and the SINR values, the ST is defined using Shannon capacity formula [45] as,

$$ST = \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}} \sum_{w \in \mathcal{W}} x_{w,j}^c O_{w,j}^c b_j \log_2(1 + SINR_{w,j}^c). \quad (26)$$

It is shown in Fig. 4 that, for most of the channels in the system, the proposed scheme achieves higher ST gain than the comparative TVWS sharing schemes. However, the proposed scheme with FR implementation achieves slightly lower ST than the Scheme in [17] for the case when the number of channels in the system is two. This is because the Scheme in [17] focuses on maximizing the throughput in the TVWS allocation process while the proposed scheme focuses on making a balance among the contradicting QoS metrics; ST and fairness in allocation. Consequently, the WSOs with lower channel quality (here lower SINR value) also get a proportion of the available TVWS which reduces the total ST gain in the proposed scheme. However, as the number of channels in the system reaches to four and above,

the proposed scheme achieves higher ST gain and remains so until both the schemes converge to the maximum achievable ST. The reason for such improvement is that the proposed scheme applies a joint time-frequency FR concept which accommodates a higher number of WSOs in the available TV channels while the Scheme in [17] applies FR concept in frequency domain only. Note that the ST gain in this study is defined as maximum if all of the WSOs in all the CMs get their desired channel demands.

The effect of variability in the $(O_{w,j}^c, K_{w,j}^c)$ pair values on the ST gain of the three allocation schemes is shown in Fig. 4(a), 4(b) and 4(c), respectively. The three allocation schemes converge to the maximum ST, as the number of channels in the system reaches 8 and 16, as shown in Fig. 4(a) and 4(b), respectively. However, in high subdomain case (Fig. 4(c)), none of the allocation scheme achieves the maximum ST. The reason is that the high channel occupancy demand of WSOs results in a few WSOs to saturate the available TVWS while leaving no channel share for rest of the WSOs.

Another notable property of the ST study is that, as the $(O_{w,j}^c, K_{w,j}^c)$ pair values increases from low to high subdomains, the ST gain of the proposed scheme improves over ST gains in the comparative scheme, as shown in Fig 4(a) through Fig. 4(c), respectively. This improvement is attributed to the combined effect of the use of the proportional fairness in the allocation and implementing FR in a joint time-frequency domain in the proposed scheme, as defined in the Section IV and V respectively.

2) Fairness

The fairness in allocation among CMs in the system is measured using equation (21) where the variability in CMs'

normalized throughput vector, $\mathbf{T} = (T^1, T^2, \dots, T^C)'$ is used as a fairness metric to compute the fairness index (FI) value. The FI result, as shown in Fig. 5, confirms that the proposed scheme achieves the highest FI value due to the combined use of the proportional fairness method and the FR implementation in the joint time-frequency domain. On the other hand, though both, the Scheme in [17] and the Scheme in [18], optimize the fairness in allocation. However, both the schemes make an orthogonal TV channel allocation thus, resulting in lesser number of WSOs to get the channel which reduces FI value. Moreover, the constraint of maintaining the lexicographically ordered throughputs of the WSOs in the Scheme in [18] further reduces the degree of the fairness in allocation.

The effect of varying the values of the $(O_{w,j}^c, K_{w,j}^c)$ pair in low, medium and high subdomains is shown in Fig. 5(a), 5(b) and 5(c), respectively. It is shown in Fig. 5(a) and Fig. 5(b) that the FI values of all the comparative allocation schemes converge to the maximum FI value, i.e., 1, as the number of channels in the system reaches 8 and 16, respectively. However, in the high subdomain case (Fig. 5(c)), none of the comparative allocation schemes converge to the maximum FI value except for the proposed scheme with the FR implementation. It is because, in all other schemes, their orthogonal channel allocation policy result in a few WSOs to saturate the available TVWS while in the proposed scheme, the spatial reuse of the TVWS in a joint time-frequency domain

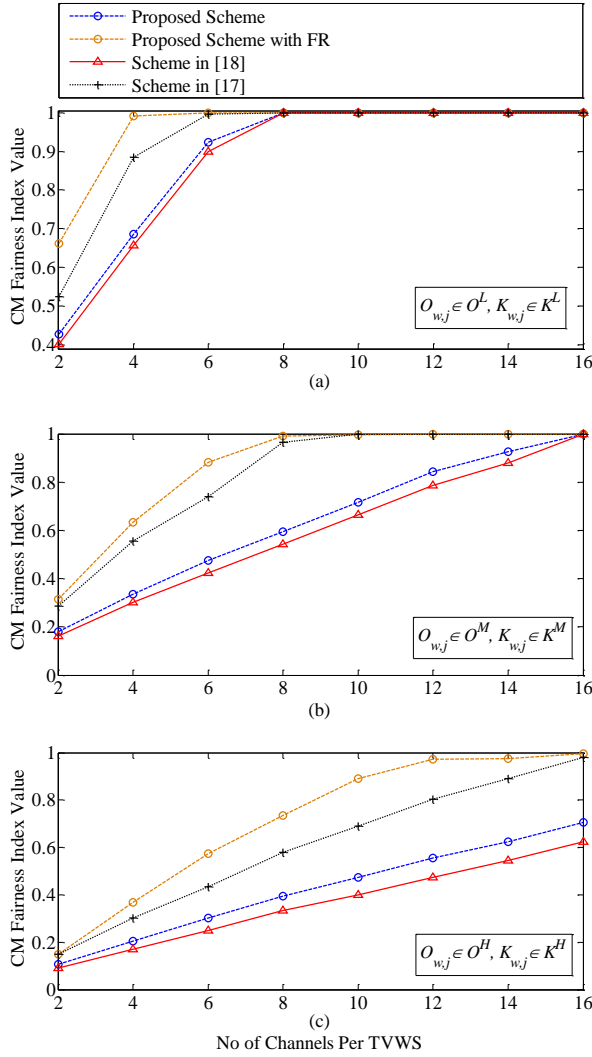


Fig. 5. Fairness index value calculated using normalized throughput vector of CMs for a varying number of TV channels in the system.

accommodates as many as WSOs, registered in the CMs which improves fairness in allocation.

3) WSO Satisfaction

In this study, we analyze the performance of the three allocation schemes the third objective of the TVWS sharing problem defined in Section III-B. in this study, a WSO is considered satisfied from allocation if it gets its desired channel demand for the duration of desired channel occupancy. The system-wide WSO satisfaction percentage (S) is then calculated using percentage of the mean satisfaction as,

$$S = 100 \sum_{c \in \mathcal{C}} \frac{\sum_{w \in \mathcal{W}^c} \frac{\sum_{j \in \mathcal{J}} x_{w,j}^c}{n_w}}{\mathcal{W}^c} \quad (27)$$

Fig. 6 shows the simulation result of the satisfaction study of the three allocation schemes. This figure shows that the proposed scheme and the Scheme in [18] achieves similar satisfaction result as their lines overlap each other. However, the proposed scheme with FR implementation achieves better satisfaction result than that of the Scheme in [17]. It is because, the TVWS allocation in a joint time-frequency domain enables the proposed

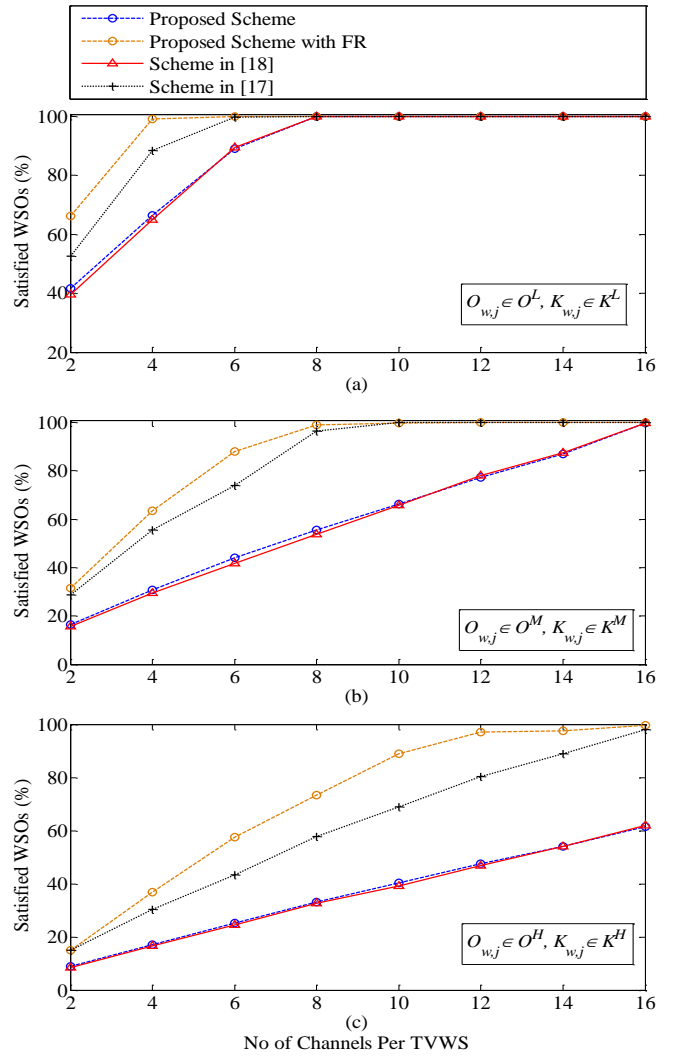


Fig. 6. Percentage of total 32 WSOs satisfied from the allocation.

scheme to accommodate as many as WSOs in the available TVWS while the third objective in the TVWS sharing problem, in Section III-B, requires the proposed scheme to satisfy the channel demand of each allotted WSO. Such an allocation strategy improves the satisfaction result of the proposed scheme.

From the results in Fig. 4 to Fig. 6, it is clear that none of the comparative schemes results in better performance than the proposed scheme in any of the performance metric. The proposed scheme, however, gives fairer channel allocation among all comparative allocation schemes. The proposed scheme with the FR implementation, however, outperforms the comparative schemes, in most of the TV channels in the system, in all the three performance metrics as shown in Fig. 4 to Fig. 6.

D. Increasing WSO Density

In this section, the effect of increasing the number of coexisting WSOs in the performance of the proposed allocation scheme is evaluated. The performance is measured using the metric like system throughput and WSO satisfaction, for the three subdomain cases, i.e., low, medium and high. The number of WSOs registered in each CM in the system varies in a set,

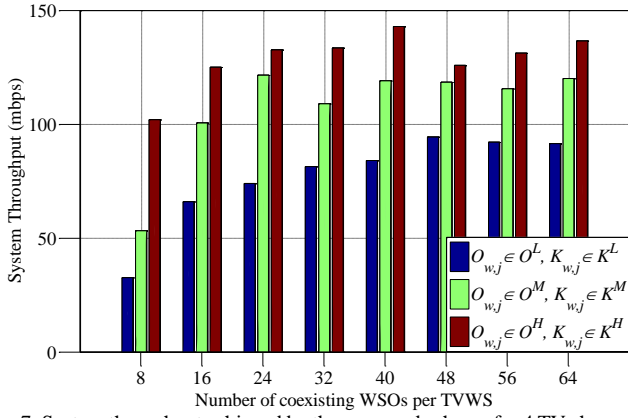


Fig. 7. System throughput achieved by the proposed scheme for 4 TV channels and a varying number of WSOs.

$W \in \{8, 16, 24, \dots, 64\}$. The number of available TV channels remains constant at 4, and the other simulation parameters are same as defined in Section VI-B. The results of the performance study are shown in Fig. 7 and Fig. 8.

Fig. 7 shows that the highest throughput gain is achieved in the high subdomain case, i.e., when $(O_{w,j} \in O^M, K_{w,j} \in K^M)$. The reason is that the proportional fairness method in the proposed scheme selects the WSOs with high throughput gain to share the available TVWS. While spatially reusing the frequency further helps the proposed scheme to accommodate as many as WSOs in the available TVWS. Consequently, the ST increases in high subdomain case. On the other hand, the achieved throughput is the least in low subdomain case, i.e., when $(O_{w,j} \in O^L, K_{w,j} \in K^L)$. It is because; the low channel occupancy demand of the WSOs could not saturate the available whitespace.

Fig. 8 shows the percentage of the number of WSOs satisfied from the allocation, calculated using (27). This figure shows that the satisfaction is the highest in the low subdomain, followed by the medium subdomain, especially in the case when $W=8$, for each CM. The reason is that a relatively greater number of WSOs can be satisfied per TVWS when $W=8$. The WSP value then sharply declines as the number of WSOs in the system increases, especially for the medium and high subdomain cases.

The results in Fig. 7 and Fig. 8 shall facilitate the modeling of a channel sharing system such that given the statistics of channel quality, the WSOs channel demands and the WSO density in the system, one can estimate an optimal number of WSOs that can be accommodated on the available TVWS such that the resource utilization is maximized.

E. Algorithm Scalability Test

The scalability of the proposed algorithm in terms of time taken to complete the allocation process is evaluated. In this experiment, the total number of coexisting WSOs registered in all the CMs in the system varies geometrically as, 2^W where $W \in \{3, 4, 5, 6, 7\}$. The number of TV channels in the system increases at a constant interval of 4 as, $J \in \{4, 8, 12, \dots, 48\}$. The remaining simulation parameters are same as defined in Section VI-B. The specifications of the computer system used for the scalability test is listed in Table 4. Using the above parameters,

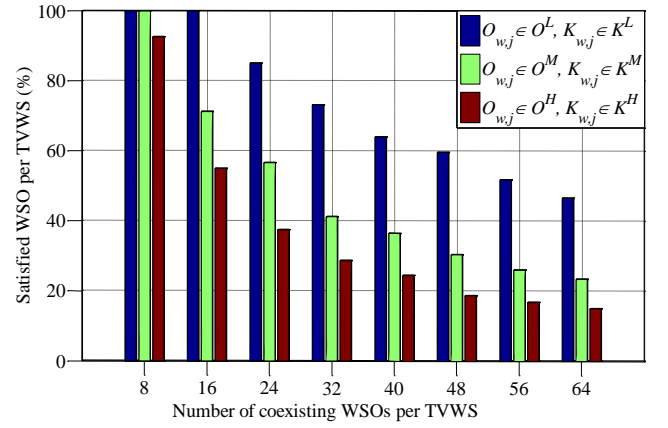


Fig. 8. WSOs satisfied from allocation with varying WSO density in the region. The number of TV channels in the system is 4.

the `intlinprog` routine solves the proposed TVWS sharing problem. The routine uses the branch and bound method to find an optimal solution point \mathbf{X} . The branch and bound split the problem into sub-problems, and each sub-problem is expanded until a solution is found as long as its cost does not exceed the set upper bound. The exact computational complexity of any branching algorithm is hard to find as time complexity of such a branching algorithm is usually analyzed by the method of branching vector. However, it has been mentioned in [46] that when the best-first search branch and bound technique is used, the upper bound to generate an expected solution is $\sum_{i=0}^n T(i) \leq \sum_{i=0}^n n-i+1 \leq (n+1)^2$ where n is the number of nodes visited. Thus, the complexity of such an algorithm is $\mathcal{O}(n^2)$.

In this experiment, we measure the simulation time taken using the MATLAB® tic-toc stopwatch timer. The time recorded for the high domain channel assignment is shown in Fig. 9. The result in this figure is generated using the average time required to complete allocation for the high subdomain case, i.e., $O_{w,j} \in O^H$ and $K_{w,j} \in K^H$. In this study, we perform the population engineering step, as defined in Section VI-B, using the `intlinprog` routine of the MATLAB. The figure indicates that for defined simulation parameters, the channel sharing process took a few hundreds of milliseconds to complete the allocation process which is quite acceptable for real-time implementation of the algorithm. The Fig. 9 shows that the algorithm execution time does not grow geometrically as the number of WSOs in the system increases. Rather, the algorithm has linear time allocation behavior as shown in Fig. 9.

VII. CONCLUSION

In this study, we investigated the channel sharing problem in a TVWS sharing domain with the objective of maximizing the resource utilization. The defined TVWS sharing problem optimizes the system throughput under a minimum fairness in allocation while constrained to satisfy the WSO channel occupancy demand on each allocated channel. To solve the defined problem, we proposed a channel allocation algorithm that shares the available TVWS among coexisting WSOs operating on incompatible network technologies. In order to

TABLE 4: COMPUTER SYSTEM CONFIGURATION

Symbol	Description	Quantity
Processor	Intel quad core i5-2500k	CPU = 3.30 GHz
Onboard memory	-	8555 MB
Memory used by MATLAB®	-	1289 MB

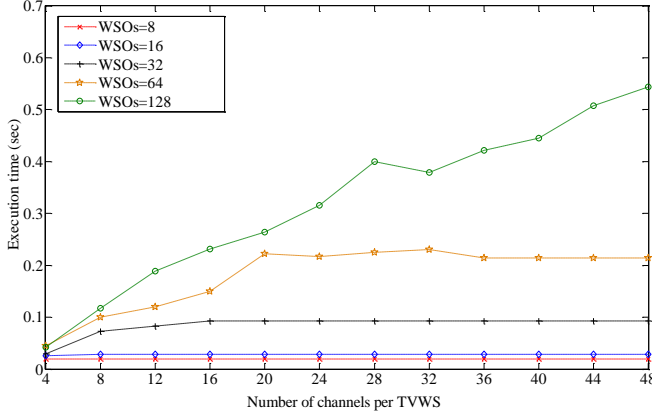


Fig. 9. Algorithm execution time for varying number of WSOs and varying number of TV channels in the system.

improve the TVWS utilization and to accommodate as many as WSOs in the available TVWS, the proposed algorithm spatially reuse the available TV spectrum. The simulation results show that the frequency reuse property of the proposed algorithm results in comparatively higher WSO satisfaction from the allocation, better fairness in allocation and higher system throughput gain. Moreover, the fast allocation process of the proposed algorithm makes it a promising candidate for implementation in 802.19.1 based coexistence system. The proposed algorithm can be implemented by a centralized decision-making entity, i.e., the master CM in the IEEE 802.19.1 system without requiring any major change in the baseline architecture of 802.19.1 TVWS sharing system.

Appendix A

In this section, we apply tangent plane approximation to linearize the objective function in (14a).

Let for some given points on the graph, $(q_1 = x_{1,j}^c, q_2 = x_{2,j}^c)$,

and $F = \log(q+1)$, where $q = \frac{q_1 r_{1,j}^c}{O_{1,j}^c + \delta_{O_{1,j}^c, 0}} + \frac{q_2 r_{2,j}^c}{O_{2,j}^c + \delta_{O_{2,j}^c, 0}}$. If

$\log(\mathcal{U}_{c,j} + 1)$ is differentiable at (q_1, q_2) , then the surface has tangent plane at (q_1, q_2, F) . The equation of the tangent plane at (q_1, q_2, F) is given by,

$$\frac{\partial y}{\partial x_{1,j}^c}(q_1, q_2)(x_{1,j}^c - q_1) + \frac{\partial y}{\partial x_{2,j}^c}(q_1, q_2)(x_{2,j}^c - q_2) - (F - F) = 0$$

where y denotes multivariate objective function $\log(\mathcal{U}_{c,j} + 1)$

and $F = \log(\mathcal{U}_{c,j} + 1)$.

The tangent plane equation is rearranged as,

$$F = F + \frac{\partial y}{\partial x_{1,j}^c}(q_1, q_2)(x_{1,j}^c - q_1) + \frac{\partial y}{\partial x_{2,j}^c}(q_1, q_2)(x_{2,j}^c - q_2)$$

where $\frac{\partial y}{\partial x_{1,j}^c} = \frac{r_{1,j}^c}{(\mathcal{U}_{c,j} + 1)(O_{1,j}^c + \delta_{O_{1,j}^c, 0})}$ denotes partial

derivative of log function at $x_{1,j}^c$. Thus, if F is differentiable at (q_1, q_2) , then the tangent plane to the surface at (q_1, q_2) provides a good approximation to F near (q_1, q_2) ,

$$F \approx F + \frac{\partial y}{\partial x_{1,j}^c}(q_1, q_2)(x_{1,j}^c - q_1) + \frac{\partial y}{\partial x_{2,j}^c}(q_1, q_2)(x_{2,j}^c - q_2)$$

which is called as linear approximation of y near (q_1, q_2) .

For a general case with $|\mathcal{W}^c| = n$, and near to some given point,

$\mathbf{q} = q_1 = x_{1,j}^c, \dots, q_n = x_{n,j}^c$, we define linear approximation of y as,

$$F \approx F + \frac{\partial y}{\partial x_{1,j}^c}(\mathbf{q})(x_{1,j}^c - q_1) + \dots + \frac{\partial y}{\partial x_{n,j}^c}(\mathbf{q})(x_{n,j}^c - q_n).$$

Appendix B

In this section, we aim to discuss the convergence property of the algorithm in Table 2. Note that our discussion here closely follows the discussion on the convergence of sub-gradient algorithm defined in [47]. Interested readers are referred to [47] for seeking knowledge beyond what is presented in this short discussion.

Given $\lambda^0 \in E^W$ and the sequence $\{t_k\}$ of positive scalars, called *step sizes*, in Table 2, define the sequence $\{\lambda^k\}$ as defined in Step 5-b) in Table 2,

$$\lambda^{k+1} = \max\{\lambda^k + t_k \nabla h(\lambda^k), 0\}.$$

For any λ , the maximum of (17) is assumed for at least one value of the index k . Since (17) is piecewise linear, there then exists at least one point λ^* such that $h(\lambda^*) = h^* = \max P(\mathbf{X}, \lambda^*)$. Then, $h(\lambda^k)$ will converge to its optimum h^* under the conditions,

$$\lim_{k \rightarrow \infty} t^k \rightarrow 0, \quad \sum_{k=0}^{\infty} t_k = \infty.$$

For the proof of the convergence of subgradient algorithm, the interested readers are encouraged to consult [47].

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIP) under Grant NRF-2015R1A2A1A05001826.

REFERENCES

- [1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021" San Jose, CA, USA, Tech. Rep., Feb. 2017.
- [2] Wikipedia, "White spaces (Radio)," [https://en.wikipedia.org/wiki/White_spaces_\(radio\)](https://en.wikipedia.org/wiki/White_spaces_(radio)), 2011.
- [3] FCC, "Third memorandum opinion and order in the matter of unlicensed operation in the TV broadcast bands, additional spectrum for unlicensed devices below 900 MHz and in the 3 GHz band," docket 02-380, Apr. 2012.
- [4] Ofcom, "Implementing TV White Spaces," Feb. 2015.
- [5] Industry Canada, "Consultation on a Policy and Technical Framework for the Use of Non-Broadcasting Applications in the Television Broadcasting Bands Below 698 MHz," Aug. 2011.
- [6] T. Baykas, M. Kasslin, M. Cummings, H. Kang, J. Kwak, R. Paine, A. Reznik, R. Saeed, S. J. Shellhammer, "Developing a Standard for TV White Space Coexistence: Technical Challenges and Solution Approaches," *IEEE Wireless Comm.*, vol. 9, no. 1, pp. 10-22, Feb. 2012.
- [7] C. Ghosh, S. Roy, and D. Cavalcanti, "Coexistence Challenges for Heterogeneous Cognitive Wireless Networks in TV Whitespaces," *IEEE Wireless Comm.*, vol. 18, no. 4, pp. 22-31, Aug. 2011.
- [8] L. Akter and B. Natarajan, "Modeling Fairness in Resource Allocation for Secondary Users in a Competitive Cognitive Radio Network," *Proc. IEEE WTS*, pp. 1-6, 2010.
- [9] G. Villardi, Y. D. Alemseged, C. Sun, C. S. Sum, T. H. Nguyen, T. Baykas, and H. Harada, "Enabling Coexistence of Multiple Cognitive Networks in TV White Space," *IEEE Wireless Comm.*, Aug. 2011.
- [10] IEEE Standard 802.19.1, "TV white space coexistence methods," May, 2014.
- [11] B. Bahrak and J. Park, "Coexistence Decision Making for Spectrum Sharing Among Heterogeneous Wireless Systems," *IEEE TWC*, vol. 31, no. 3, pp. 1298-1307, Mar. 2014.
- [12] IEEE Standard 802.15.2, "coexistence of wireless personal area networks with other wireless devices operating in unlicensed frequency bands," 2003.
- [13] S. Pollin, I. Tan, B. Hodge, C. Chun, and A. Bahai, "Harmful Coexistence between 802.15.4 and 802.11: a Measurement-Based Study," *Proc. 2008 Int. Conf. Cognitive Radio Oriented Wireless Netw.*
- [14] IEEE standard 802.11af, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications— Television white spaces (TVWS) operation," 2014.
- [15] IEEE Standard 802.22.1, "Standard to Enhance Harmful Interference Protection for Low-Power Licensed Devices Operating in TV Broadcast Bands," 2010.
- [16] *MAC and PHY for Operation in TV White Space*, std. ECMA-392, Jun. 2012.
- [17] F. Hessar and S. Roy, "Resource allocation techniques for cellular networks in TV white space spectrum," in *Dynamic Spectrum Access Networks (DYSPAN)*, 2014 IEEE International Symposium on, April 2014, pp. 72-81.
- [18] T. Bansal, D. Li, and P. Sinha, "Opportunistic Channel Sharing in Cognitive Radio Networks," *IEEE Trans. on Mobile Computing*, vol. 13, no. 4, pp. 852-865, Apr. 2014, doi: 10.1109/TMC.2013.59.
- [19] T. Marler and J. Arora, "Survey of Multi-objective Optimization Methods for Engineering," *Structural Multidisciplinary Optimization*, vol. 26, no. 6, pp. 369-395, 2004.
- [20] K. Khalil, G. Farhadi, and A. Ito, "Coexistence Management for Heterogeneous Networks in White Spaces," *Proc. International Conference on Computing, Networking, and Communications, Wireless Network Symposium*, pp. 691-697, Feb. 2014, doi:10.1109/ICCNC.2014.6785420.
- [21] D. Zhang, Q. Liu, L. Chen, W. Xu, "Ecology-based Coexistence Mechanism in Heterogeneous Cognitive Radio Networks," IEEE GLOBECOM, 2015.
- [22] K. Bian, J. M. Park and B. Gao, "Frequency reuse over a single TV white space channel," USA, Elsevier, pp 77-100, 2014.
- [23] B. Gao, J-M J. Park, and Y. Tang, "Uplink soft frequency reuse for self-coexistence of cognitive radio networks," *IEEE Trans. on Mob. Comp.*, Vol.13, Issue 6, June 2014. Pp. 1536-1233.
- [24] L. M. O. Khanbary, and D. P. Vidyarthi, "Reliability-based channel allocation using genetic algorithm in mobile computing," *IEEE Trans. Vehicular Tech.*, vol. 58, no. 8, pp. 4248-4256, Oct. 2009.
- [25] A. P. Shrestha, J. Won, S.-J. Yoo, M. Seo, and H.-W. Cho, "Genetic algorithm based sensing and channel allocation in cognitive ad-hoc networks," *Int. Conf. on ICT Convergence (ICTC)*, pp. 109-111, 2016.
- [26] Y. Jao, and I. Joe, "Energy-efficient resource allocation for heterogeneous cognitive radio network based on two-tier crossover genetic algorithm," *Jr. Commun. Networks*, vol. 18, no. 1, pp. 112-122, Feb. 2016.
- [27] D. T. Ngo, C. Tellambura, and H. H. Nguyen, "Efficient resource allocation for OFDMA multicast systems with spectrum-sharing control," *IEEE Trans. Vehicular Tech.*, vol. 58, no. 9, pp. 4878-4889, Nov. 2009.
- [28] M. R. Sherif, I. W. Habib, M. Nagshineh, and P. Kermani, "Adaptive allocation of resources and call admission control for wireless ATM using genetic algorithms," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 2, pp. 268-282, Feb. 2000.
- [29] M. Alias, S. Chen, and L. Hanzo, "Multiple-antenna-aided OFDM employing genetic-algorithm-assisted minimum bit error rate multiuser detection," *IEEE Trans. Veh. Technol.*, vol. 54, no. 5, pp. 1713-1721, Sep. 2005.
- [30] F. Hessar and S. Roy, "Capacity considerations for secondary networks in tv white space," *IEEE Trans. on Mob. Comp.*, Vol.14, Issue 9, Sep. 2015. Pp. 1780-1793.
- [31] L. E. Li, M. Pal, Y. R. Yang, "Proportional Fairness in Multi-rate Wireless LANs", *IEEE INFOCOMM.*, 2008, pp. 1004-1012.
- [32] T. Bonald, L. Massouli, A. Prouti, and J. Virtamo, "A Queuing Analysis of Max-min Fairness, Proportional Fairness, and Balanced Fairness," *Queueing Systems: Theory and Applications*, vol. 53, no. 1-2, 2006.
- [33] L. B. Jiang and S. C. Liew, "Proportional Fairness in Wireless LANs and Ad-hoc Networks," *IEEE WCNC*, 2005.
- [34] F. Kelly, "Charging and Rate Control for Elastic Traffic," *European Trans. on Telecomm.*, vol. 8, no. 1, pp. 33-37, 1997.
- [35] T. Bu ; L. Li ; R. Ramjee, "Generalized Proportional Fair Scheduling in Third Generation Wireless Data Networks," in *Proceedings IEEE INFOCOM 2006*, April 2006.
- [36] C. Cano, D. J. Leith, "Coexistence of WiFi and LTE in Unlicensed Bands: A Proportional Fair Allocation Scheme," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 2288 - 2293, June 2015.
- [37] FP Kelly, AK Maulloo and DKH Tan, "Rate control for communication networks: shadow prices, proportional fairness, and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237-252, Mar. 1998.
- [38] S. S. Byun and C. Yoo, "Minimum DVS Gateway Deployment in DVS-based Overlay Streaming," *Computer Comm.*, vol. 62, no. 3, pp. 537-550, 2008.
- [39] L. Daniel and K. Narayanan, "Congestion Control 2: Utility, Fairness, and Optimization in Resource Allocation," *Lecture Notes*, March 2013.
- [40] W. Murray and K. M. Ng, "An Algorithm for Nonlinear Optimization Problems with Binary Variables," *Computational Optimization and Applications*, vol. 47, no. 2, pp. 257-288, Oct. 2010.
- [41] M. Pioro and D. Medhi, "General Optimization Methods for Network Design," *Routing, Flow, and Capacity Design in Communication and Computer Networks*, USA, Elsevier, pp. 178-184, 2004.
- [42] R. Jain, "Selection of Techniques and Metrics," *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design Measurements Simulation and Modeling*, Wiley Computer Publishing, pp. 30-40, 1991.
- [43] R. Jain, D.M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Technical Report, Eastern Research Lab., Dig. Equip. Corp., Hudson, MA, available at: <http://www1.cse.wustl.edu/jain/papers/ftp/fairness.pdf>, 1984.
- [44] IEEE Standard for information technology, Part 22: Cognitive wireless RAN medium access control (MAC) and physical layer (PHY) specifications: Policies and procedures for operation in the TV Bands, IEEE Std 802.22-2011, Jul. 2011.
- [45] T. Robert, "Data Communications Fundamentals," *Data Communications: An Introduction to Concepts and Design*, USA, Elsevier, pp. 68-69, 2013.
- [46] A. Kugel, "Proofs for the paper: Average Case Complexity of Branch-and-Bound Algorithms on Random b-ary Trees," Proofs Report, Faculty of Engineering and Computer Sciences, Ulm University, 89069 Ulm, available at https://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.190/Mitarbeiter/kuegel/proofs.pdf.
- [47] M. Held, P. Wolfe, and H. P. Crowder, "Validation of Subgradient Optimization," *Math Programming*. Vol. 6, 1974, pp. 62 - 88, IBM Systems Research Institute, New York, U.S.A.,



Computer Aided Diagnostic System for Ultrasound Liver Images: A Systematic Review

Mohamed Yaseen, Heung-No Lee*

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea

Abstract

In this article an in-depth overview is presented on Computer aided diagnostic (CAD) system's usage for liver cancer. Besides, in a broader sense highlighting the technical aspects developed for medical ultrasound images is also discussed. CAD system is a process that provides adequate information that helps to analyze the Ultrasound images and helps to accurately detecting different types of liver cancer. However, the system performance is still not significantly improved. In this paper, firstly, we categorize the CAD system according to the four primary stages including data preprocessing, lesion segmentation, feature extraction, selection, and Classifier. In each stage, we review specific methods that are commonly used in most of the algorithms proposed for computerized tissue characterization and discuss their advantages and drawbacks. Then, recent proposed algorithms are presented in summarize form that have shown clinical value or specific possibility to the computerized analysis of setback for ultrasound liver images. These techniques or their combinations are the ones that are mostly used in the past few decades by the majority of work published in the Computer aided diagnosis domain.

© 2016 Elsevier Ltd. All rights reserved

Keywords: Ultrasound Image; Liver Cancer Diagnosis; Computer Aided Diagnostic System; Segmentation; Artificial Intelligence

1. Introduction

In therapeutic imaging and analytical radiology as a key subject matter of research Computer aided diagnostic (CAD) has emerged. CAD has applicability in numerous medical imaging modalities. Some of which are computerized tomography (CT), magnetic resonance imaging (MRI), ultrasound (US) imaging, and nuclear medicine [1], [2], [3], [4]. For US liver image diagnosis CAD strongly depend by the quality of data.

* Corresponding author.

E-mail address: Heungno@gist.ac.kr.

The comparatively substandard of clinical US images reduces the success of early liver ailment finding and analysis. There are distinctive objectives [5] which make the system assignment complex, such as speckle, attenuation, signal dropout and shadows. It is due to the orientation dependence of acquisition that can outcome in losing boundaries. Additional complexities appear due to the highly variable shape of the liver, reduced contrast and intensity inhomogeneity within liver, weak boundaries to its nearest organs say heart and stomach, and intensity homogeneity to nearer organs. So, liver diagnosing from US device seems an exigent task that has called the attention of many researchers in recent times.

The data generated by the automated computer processes while diagnosis is helpful to the radiologist to realize the US liver images. So, the precision of image diagnosis is better, and the time required by regular methods in peruse an image is reduced [6]. Henceforth, utilizing CAD the analysis of diseases has become a vibrant area of research [7]. There is greater requirement of precisely analyzing the therapeutic images and lessening the time requisite for proper analysis of liver cancer.

The key objective of this review is emphasizing on the potentiality of intelligent computer systems to be utilized in clinical application to support pathologists to analyze and classify US liver cancerous tissue images. On the basis of methodical analysis of various liver conditions, CAD methods and organized summary of algorithm, we categorize the computerized system according to the four primary stages of analyzing liver US image. Here the using of general procedures including data preprocessing, lesion segmentation, feature extraction and selection, and depicting of cancer by means of a classifier [8] better summarizing the performance of each category leads to find the ideal solution for automatic computerized system performance and the four stages are given in Figure 1.

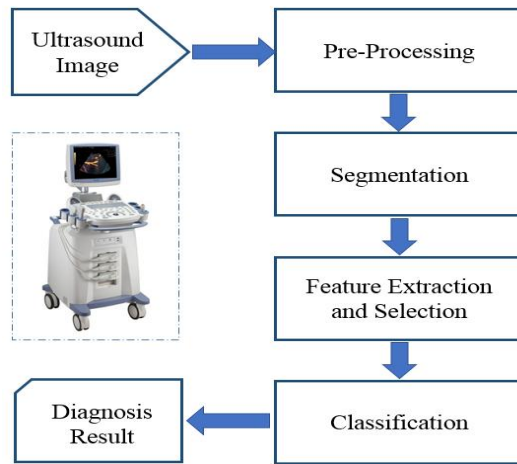


Fig. 1. Flow diagram of CAD for liver cancer.

1. Data preprocessing: The preprocessing task is to restrain the noise and to increase the image without eliminating the important features of Liver US images.
2. Image segmentation: Here image is divided into a number of small portions, and it forms the background the lesions detached. The edges of the lesions are outlined for feature withdrawal.
3. Feature extraction and selection: The stage identifies a feature set of liver cancer lesions that can precisely differentiate normal tissue or abnormal cancer tissue. The feature space could be vast and intricate, so withdrawal and choosing the finest features is decisive.

4. Classification: After the selected features, the apprehensive regions will be characterized into distinctive classifications, say normal tissue or cancerous tissue.

In following manner, this paper is arranged. Classification of liver cancer is presented in section 2. The literature review about the four stages of the CAD system in US liver diagnosis is discussed in section 3. Relevant research works are detailed in section 4. In section 5 concluding part is presented.

2. Classification of liver cancer

Globally liver cancer is much popular malignant disease, mainly in Southeast Asia and sub Saharan Africa. Worldwide, liver cancer has sixth position as the most familiar cancer with a half a million people affected each year. The number of people who develop liver cancer is increasing around the globe [9].

In human body liver is one of the indispensable organs. It's extremely hard to live without a sound liver because of its impacts on every other body parts. Mainly focal liver diseases and diffused liver diseases influence the liver. Diffused liver diseases, for example cirrhotic and fatty, harm the total surface of liver. Focal liver diseases which affect the small area of the liver surface, such as hepatocellular carcinoma (HCC), hemangioma (Hem), and cyst. Figure 2 presents three focal liver diseases in the US images. The hepatocellular carcinoma (HCC) and echo type in the liver based on US image representation, five types of primary carcinoma of liver tumour are there, they are correlative to low echo type, equal echo type, high echo type, mixed echo type and diffuse type correspondingly.

Complexity of the liver tumour in patients having chronic viral illnesses, can be classified from asymptomatic strong carriers to patients with liver cirrhosis [10], [11], [12]. In general, the US appearance of Cyst, Hem, and HCC visible similar. Brightness mode (B-mode) ultrasound [13] diagnosis is the foremost choice in well-liked analyses because of its effectiveness, non-invasiveness, and economy. All types of liver cancers are not correctly diagnosable in the US images, in case there is benign appearance of deadly tumours and observer's diagnostic level is not good, careless mistakes and visual fatigue. Hence, utilizing traditional method it is extremely difficult in decision making between them. A fully efficient automatic computer system is required to be developed for disease detection and diagnose with high performance.

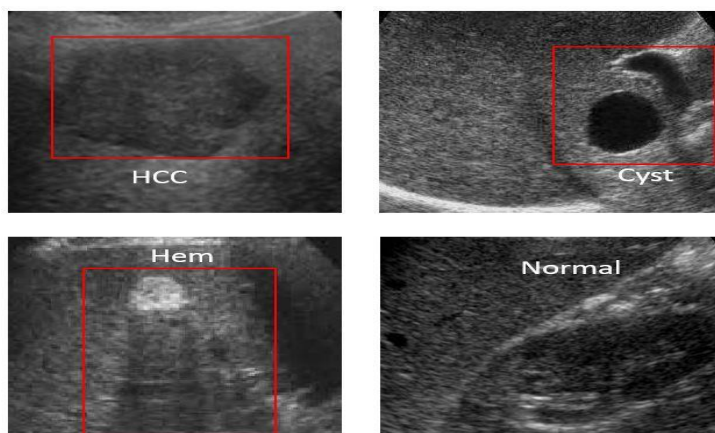


Fig. 2. Normal liver and focal liver US images.

3. Overview of the CAD system schemes

The most recent success in automatic diagnosing of liver US images is reviewed in this section. The four main steps in the CAD including data preprocessing, lesion segmentation, feature extraction and selection, and classifier of lesions are discussed in detail.

3.1. Data Preprocessing

Data preprocessing is aimed at filtering speckle noise, which impinge on the diagnostic value of the US image [14]. It makes image detail unclear and hazy drastically, demeans the image feature. Likewise, it decreased the pace and correctness of US image processing tasks say- division and classification. Hence, in US image processing tasks, speckle noise reduction is always an important prior requirement. Figure 3 depicts an example for speckle noise image and enhanced image.

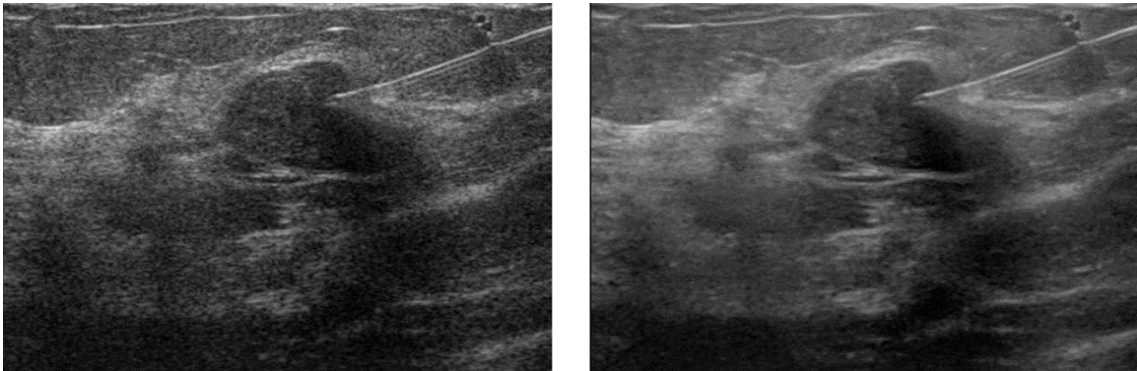


Fig. 3. (Left) Speckle noise image and (Right) Despeckled US image.

In this paper, we categorized the speckle reduction techniques into two major classifications, namely a) spatial filtering methods and b) multiscale methods. Those methods are effective in eliminating the speckle noise and conserving the analytical information in US images.

a. Spatial filtering process

The basis of spatial filter is proportion of local statistics. It is helpful to improve smoothing in uniform regions of the US images where speckle is completely visible. This method lessens smoothing substantially in surrounding areas of the image to conserve the helpful particulars of the image [15]. Lee and Kuan Spatial filters use local statistics to perform straight on the intensity of the image [16], [17], [18].

Various sorts of filters are utilized as a part of uses of speckle lessening in US imaging. The most usually utilized sorts of filters are:

- Mean filter [19] is easy to apply, also its a plain filter . However, speckles are not eliminated by it, but in the data averaged by it. It is an attractive technique for speckle noise diminishing as it can make loss of resolution and accuracy Image can be obscured by it. It has amazing quality for added substance Gaussian noise, though the speckled image comply a multiplicative form with non-Gaussian noise since the mean filter is not an ideal selection.
- Median filter [20] is very efficient against impulsive type noise and edge conserving characteristics. It generates least obscure images in comparison to mean filter. It requires listing of all near values into

numerical order to figure out the median and here it is the drawback. Moreover, it takes additional calculation time to list the intensity value of all set.

- Wiener filter [21], [22] replace images in the existence of noise and blur. Decreasing the quantity of noise presence in a signal by comparing with an assessment of the preferred noiseless signal is the aim of algorithm here. The approach of filter towards image smoothing is on the basis of calculation of local image variation. The smoothing becomes less when the local variation of the image is immense. The filter does more smoothing when the difference is small. This calculation over linear filtering over linear filtering. It conserves edges and other high recurrence information of the images, however takes more time for calculation than linear filtering.
- Enhanced Frost and Lee filter [23] is utilized to vary the capability in light of the limit value. The filter works out like a stern all pass filter when the local coefficient of variation is over a greater limit. On the other part as the local coefficient of variation goes under a poor limit then pure averaging is actuated. The stability amongst averaging and identity operation is processed when the coefficient of variation stands at middle of lower and higher thresholds.
- Gamma Map filter [24] is like preceding filter aside from that the local coefficient of variation takes place amid the two limits; the filtered pixel value depends on the Gamma estimation of the contrast proportions inside the proper filter window. It is utilized to reduce the loss of texture information. The filter needs suppositions about the dispersion of the genuine procedure and the degradation model.
- Frost filter [25] is an adaptive and exponentially weighted average making filter in light of the proportion of the local standard deviation to the local mean of the debase image. Within the $n \times n$ moving core it interchanges the region of interest with a weighted sum of the values. The weighting aspects lessen with difference from the region of interest. The weighting aspects increment in the mid region as difference inside the core grows.
- Lee filter [16], [17] relies upon the multiplicative speckle model. It can utilize local statistics to conserve borders and attributes adequately. It also uses the process like when the variance over an area is poor, then the smoothing will be done. When the difference is much similar to near borders, no smoothing will be done.
- Kaun filter [26] Irelies upon an image's Equivalent Number of Looks (ENL) to decide an unlike weighting function to do the noise reduction. The filter model is a local linear least square inaccurate filter relies on multiplicative model regarded as to be finer to Lee filter. It makes no estimation of the noise variance inside the filter window.
- Diffusion filter is for smoothing images on a nonstop area, nonlinear partial differential equation was implemented by Perona and Malik [27], which has since been extended and enhanced [28], [29]. Through many years, other denoising processes with extremely fascinating ability are developed for example: Bilateral filter along with derivatives [30]. In Speckle reducing anisotropic diffusion (SRAD) [31] dispersion of speckled images is edge-sensitive. Its preference is a rapid and a decent speckle lessening impact. In SRAD, the instantaneous coefficient of variation goes about as the border identifier. Here algorithm displays maximum gains at the borders and creates least gains in consistent areas. This way, it guarantees the mean-preserving conduct in the uniform areas, and conserves and improve the borders.

The noted diffusion methods can save or even improve important edges when taking out speckles. Even so, the techniques have one basic constraint in holding unobtrusive features of minute cyst and lesion in US images.

b. Multiscale process

For US imaging numerous speckle reduction algorithms are proposed in light of contourlet, curvelet and wavelet.

- **Wavelet Transform:** The key target of speckle diminishment is eliminating the speckle noise by not missing much data included in an image. To be successful this target Wavelet transform have set up since it gives an ideal representation for 1D (single dimensional) piecewise smooth signals, for example, an image's scan lines [32]. The complex wavelet transform (CWT) just requires $O(N)$ computational to enhance directional selectivity. Yet, in the past intricate wavelet change not broadly utilized, as it is hard to devise intricate wavelets with impeccable recreation properties and good filter attributes [33], [34]. Kingsbury proposed the technique known as dual-tree CWT in articles [35], [36], which can add faultless reform to the other appealing properties of difficult wavelets, incorporating limited redundancy, estimated shift invariance, six directional selectivity's, and proficient $O(N)$ calculation. To build 2D complicated wavelets here Tensor-product 1D wavelets are utilized. The directional discerning dispensed by complicated wavelets (six directions) is vastly improved from what is acquired by the discrete wavelet transform (three directions), however is by now fewer. Such undesirable practices demonstrate that further potent representations are asked in upper dimensions.
- **Contourlet Transform:** a contourlet transform utilizing 2D transform process for image delineation and study executed by Do and Vetterli [37]. It was implemented in the detached space. likewise, the researchers justify its union in the uninterrupted space. It was implemented in a detached space multiple direction and a multiple resolution extension utilizing non-distinguishable filter banks. This brought about an adaptable multi-resolution, directional and local image extension using contour segmented region, and hence it is known as contourlet transform. As specified before by utilizing a filter bank that decouples the multiscale decomposition contourlet was completed and finished by Laplacian pyramid and then directional decompositions, which are completed utilizing a directional filter bank.
- The advantages of contourlet alter are as follows. 1) The rectangular grids are utilized to portray contourlet expansions, and thus offer an impeccable interpretation to the distinct world, where based on a rectangular grid the image pixel's sample is taken. The main disparity between the contourlet [37] and the curvelet framework [38] are to attain the rectangular grid attribute, the contourlet kernel roles have to be diverse courses and by just turning a lone function cannot be acquired. 2) As a consequence of rectangular grids, contourlet have 2D division on centric squares, besides centric circles for curvelets and polar coordinates to depict other systems. 3) Since usage of iterated filter banks for wavelets, contourlet transform utilizes quick bank calculations and adaptable tree structures. 4) This calculation gives a space multi-resolution action plan which gives lithe improvements of the spatial and angular resolution. The contourlet change characterizes a multiple scale and multiple directional delineation of an image. Likewise it is simply adaptable for identifying superior attributes in any placement at a variety of scale levels [39] ensuing in fine probable for efficient image examination.
- **Curvelet Transform:** A new algorithm is presented by Candes and Donoho in article [38] on the continuous 2D (two-dimensional) space R^2 utilizing curvelets. This calculation showed a fundamentally ideal estimation manner for 2D per piece plane functions that are curves. First generation curvelet transforms are delineated in the uninterrupted domain [38] through multiscale filtering and after that on each bandpass image applied a block ridgelet transform [40]. The second generation curvelet transform [41] was produced utilizing no ridgelet transform but using frequency partitioning. However, for implementing both curvelet generation want a turning maneuver and ought to match with a 2D frequency division on the basis of on the polar coordinate. It gives the curvelet creation easily in the uninterrupted sphere, yet makes it critical sampling appears to be troublesome in discretized structures. The curvelet change is extremely proficient in representing curve-like edges. In any case, this transforms have two key disadvantages: 1) the discrete curvelet transform is superfluous. 2) They are not ideal beyond c^2 singularities for sparse approximation of curve features.

On US images most standard speckle filters perform fine, yet a few limits exist with them, which lead to image resolution degradation. In this way, while developing an efficient and strong denoising algorithm for data preprocessing stage in CAD one needs to consider various factors. In the design of despeckling methods, choices of despeckling filter and speckle model have important part. In above most usually favored models and filters were reviewed with its pros and cons.

3.2. Segmentation

Segmentation process is a mandatory step in CAD systems that frequently refers to the delineation of specific structures [42]. Segmentation's key objective is to convert the image to provide more significant data that can be effortlessly examined. It is used to distinguish the various boundaries and objects in images. Due to poor contrasts, different types of noise and missing the boundaries in medical images make segmentation Harder. Depending on anyone between the two vital traits of intensity values that are similarity and discontinuity Medical image Segmentation approaches are mostly based [43]. In subsection, the different segmentation procedures of the medical images are reviewed and it is composed into four basic classifications as appeared in figure 4.

- **Region based method:** On the basis of pixel likeness in a region, these process is developed. It is used to approximate the region straight [44], [45]. This method classify the pixels with comparable attributes (like intensity) into regions. Classification of Region based methods have two methodologies such as- a) region expanding approach and b) region combining/dividing approach. In this approach, the procedure initiated by choosing a seed region (pixel). Adams and Bischof [46] proposed the first seeded region growing. The region develops by including the neighbor pixels having comparative established in advance standard with the seed, for example- texture, potency, difference, texture or gray level, etc. When no pixel is present for inclusion then the procedure stops. The issues with this approach are- the user has to choose the seed point and it will miss the efficacy when the region is inhomogeneous. Within the region, combining/splitting mode, the technique starts with the entire image as a seed. At that point, the seed is partitioned into various sub regions, most often into four sub regions. Thus, continuity of the process goes on till there are no regions of the partition by using each sub region as a seed. Lastly, based on same properties, for example intensity, variance or gray-level combine any adjacent regions.

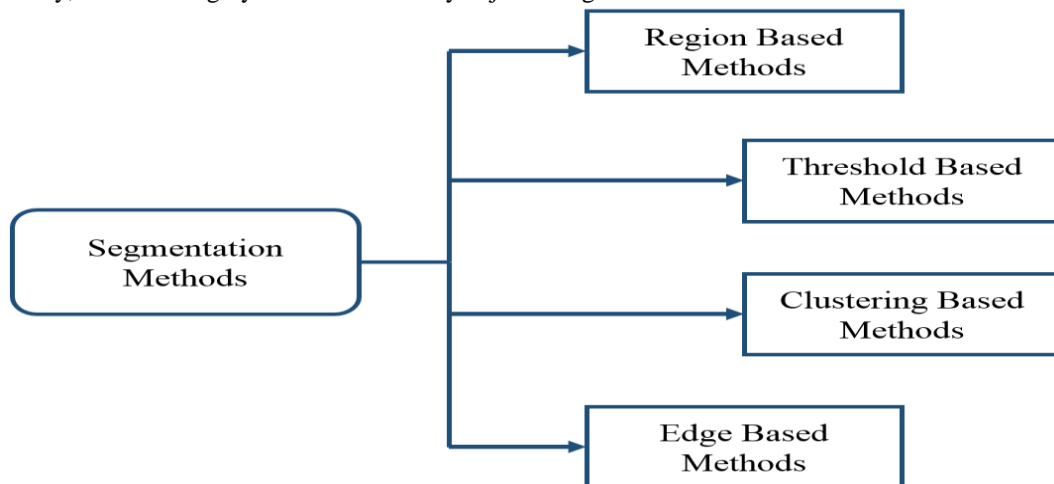


Fig. 4. Image segmentation methods.

- **Edge based methods:** Edge based methodology [44], [47] utilized for distinguishing the discontinuities in an intensity value for image segmentation. It is the sudden changes in potency level at the region borders of US images. The image border can be described as the perimeter isolated by different areas that vary in the level of potency [48]. Utilization of the borders are to identify the items' measurement and differentiate items from the background. Usage of edge detectors are needed to trace the distinctive points in the image where the potency actually changes. Border detection is a vital tool for the success of segmentation and interpreting the US image content, mainly when dealing with feature extraction and feature detection. There are two main techniques in order to detect an edge in US images such as searching and zero crossing techniques. Figuring the gradient magnitude by making use of first order derivative expressions takes place at first place in search-based technique. Subsequently, with the utilization of gradient direction local directional maxima of the gradient enormity is searched. In zero crossing technique, looking for a zero crossings in the second imitative of the image takes place. Finding zeros in the second imitative of image orders are detected, at this time the value of first imitative is high and zero is the value of second imitative. It is named Laplacian approach on the basis of edge detection. The edge based division method's disadvantage is, when there is presence of lots of edges in the image it does not work well.
- **Threshold based methods:** For image segmentation, Thresholding [49] is one of the imperative techniques used. It is helpful in a separate frontal region from the background region of the image. The gray level image can be changed to binary image by choosing a sufficient limit value T , All necessary data regarding the shape and position of the objects of interest (foreground) and the image's (background) other areas ought to be contained in the binary image. To acquire data easily gray image conversation to binary image is done, that result in the generalization of the categorization phase. Pixels having unique concentration less than the threshold rate is named "black pixels" (0) and belongs to the background. On the other hand, pixels over the threshold rate is called "white pixels" (1) and becomes object's part. There are two sorts of thresholding systems: a) global thresholding and b) local thresholding. There is fixed value of threshold T in global thresholding. Such threshold's difficulty is, if the background of the image holds unlike enlightenment, failure of segmentation procedure may occur. In local limiting, the threshold value T is not fixed, as such the problem of particular enlightenment can be sorted out by using numerous thresholds. Automatic threshold scheme utilizing different routines for example Mean, Edge maximization technique (EMT), Histogram dependent technique (HDT) is a system where [44] threshold value T for every image is automatically selected by the system exclusive of human intrusion.



Fig. 5. Examples of two types of thresholding system from left to right: Original cyst US image, local and global threshold segmented image.

The histogram-based techniques are dependent on achieving the estimated threshold value T . This threshold value T divides the two uniform backgrounds and area of the item in the image. The image having a uniform area of the item and background and separated by heterogeneous region between them, the HDT is appropriate for it.

In mean based framework, the threshold value T utilizes pixel's mean value and work fine in stringent cases of the images that have generally partly of the pixels connected with the objects and the rest half associated with the background.

The EMT segmentation system relies on finding the most border limit in the image to begin segmenting with the guide of border recognition process works. It is applied when the image holds excess of one uniform area or where there is an alteration in lighting between its background and item. As it occurs, the object sections may be united with the backdrop or a portion of the backdrop may be united with the item.

The inconvenience of thresholding method is barely two categories are created, and it cannot be tried on multichannel pictures. Thresholding does not judge the spatial distinctiveness of an image so it is irritable to noise. This distorts the histogram of the image, making the separation more troublesome. In general, thresholding procedures are reasonable for images that hold more and clear separation between the homogenous regions. It has resulted to enhance the effectiveness of the threshold technique.

- Clustering based method: It is an unattended learning undertaking, where it perceives a limited set of classes known as clusters to categorize pixels [50]. Clustering performs by either grouping pixels or partitioning pixels. In the grouping type, it starts with every component as a distinct bunch and combines all the distinct clusters in forming bigger clusters. While, in partitioning type, it begins to split into successively smaller clusters from the entire image.

The clustering techniques are divided to attended clustering and unattended clustering. In attended clustering to decide the clustering properties, Human interactions required. Whereas in unsupervised clustering technique, by own help the clustering properties are defined. There are two popular algorithms for unsupervised clustering that are, K-mean clustering and fuzzy clustering.

1) K-mean clustering: It is an unattended clustering calculation. It categorizes the input data points into numerous categories based on natural space from one another. Here data vectors are organized into predefined number of clusters. At first, the centroids of the predefined clusters are initialized arbitrarily. The centroid and data vector dimensions are same. Every pixel is consigned to the cluster on the basis of nearness, estimated by the Euclidian distance measure. Following every one of the pixels are clustered, the average of every cluster is recalculated. Repetition of this procedure goes on until no noteworthy vary outcome for some fixed number of iterations or for each cluster mean [51].

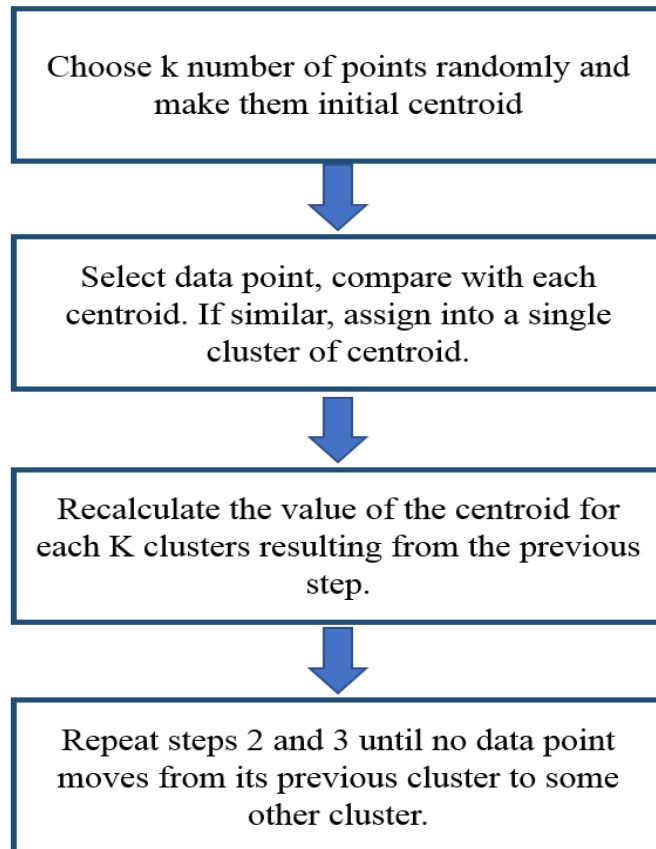


Fig. 6. K-mean clustering algorithm

2) Fuzzy clustering/Fuzzy c-means (FCM): it is an unsupervised clustering algorithm. Here a dataset is grouped into n clusters with all data point in the dataset staying with all bunches to an exact degree [52]. The Fuzzy clustering technique can be thought to be better than those of their harder counterparts, since they can represent the affiliation between the input pattern data and clusters more actually. Fuzzy c-means is a most prominent soft clustering technique; its viability to a great extent is limited spherical clusters. It has extra advantage as it is extra lithe than the corresponding stiff clustering algorithm.

3.3. Feature drawing out and choice

In the detection and classification of liver cancer feature drawing out and choice [53] are significant course of action. However, in computer-aided system only texture features are used as inputs. Texture feature extraction is the basic and traditional techniques. Different examination techniques are used to extract helpful attributes for US liver cancer image classifications. A few general utilization systems are:

- Laws Texture Energy Measure (TEM): In order to find various texture types of TEM [54] using convolution masks of 5×5 . It works to produce 25 2D masks by convolving based on 5 basic 1D masks. Afterwards texture picture is filtered with produced masks by extracting helpful attributes.

- Gray Level Difference Statistics (GLDS): It is the Probability Density Function of pair pixels lying at a particular difference and holding a discrete potency value variation. Least variation of coarse texture and large variation of fine texture in Inter pixel gray level values.
- Spatial Gray level Dependence Matrices (SGLDM) [55]: It counts how often pixels with a potency i and j happen at a particular offset to calculate matrix. It makes use of spatial relationships amid gray levels of a picture furnishes to total texture properties of the picture
- Gray level Run length Statistics (GLRLM) [56]: Its rough texture consists of comparatively long runs than short runs. It utilizes fact that containing similar gray level along a particular course of the successive points in the image.
- Gray Level Histogram: Texture parameters are obtained by using the intensity distribution of the image.
- Fourier Power Spectrum (FPS): FPS used for normal wave like forms with a fixed interval. Fourier conversion gives the frequency of form and direction.
- Edge Frequency on the basis of Texture Features: It is opposite to the autocorrelation role and depending on difference concerned gradient little and big distance operator is detected using Micro-edges and macro-edges.
- Wavelet Features: It is derived from Region of Interest (ROI) or wavelet transform of the image. Foremost types are quincunx, Gabor and dyadic.
- First Order Parameters (FOP): It defines only diffuse variation and echogenicity characteristics and these are sovereign of spatial concern amid pixels.

Successive texture analysis methods depend on selecting appropriate features. Some important textural features include- contrast (CO), Short Run Emphasis (SRE), Local Homogeneity (LH), Energy (E), Gray Level Distribution (GLD), Variance (VAR), Homogeneity (H), Uniformity (U), Sum Entropy (SENT), Dissimilarity (D), Angular Second Moment (ASM), Run Length Distribution (RLD), Mean (M), Inverse Difference Moment (IDM) and Standard Deviation (SD).

3.4. Classification

After extraction of feature and selection process, we have to classify the images into lesion/non-lesion or normal/abnormal classes. Classifiers are divided into two types - statistical and neural network, which can be classified using unattended as well as attended procedure. An example for numerical unattended classifiers such as K means clustering [57] and for statistical attained classifier e.g. Support Vector Machine (SVM) [58], [59]. In the meantime for unattended neural grid like Self Organizing Map (SOM) [60] and for attended neural grid such as Multi-Layer Perceptron (MLP) [61], [62] are utilized to classify liver images. We summarize the different US liver detection and classification techniques are listed below

- Fuzzy neural network (FNN): Diverse stochastic associations are find out by it, which represent the attributes of a picture. The diverse sorts of stochastic are grouped (set of attributes) in which the elements of this set of attributes are blurry. It gives the scope to define various classes of stochastic attributes in the comparable type [63]. Accomplishment and correctness depend on the limit choice and unclear integral. The drawback of fuzzy neural network is exclusive of previous information output is not fine and accurate result depends on the route of decision.
- Support vector machine (SVM): SVM targets to reduce the superior bound of generalization fault by increasing the periphery amid the parting hyperplane and the data [64]. Fine division is gained by the hyper plane that has the biggest difference to the closest training data point of any class (operating margin), usually bigger the margin lowers the generalization mistake of the categorizer. SVM utilized Nonparametric with binary classifier approach. It can manage additional input data extremely proficiently. Accomplishment and accurateness depend on the hyperplane choice and core limit.

SVM reduces the calculating difficulty, easy to administer decision rule intricacy and fault occurrence. The drawbacks of SVM are low result transparency, training is time consuming and finding out of finest limits is not trouble-free in presence of nonlinearly separable training data.

- Artificial neural network (ANN): It is the combination of arithmetic techniques inspired by the characteristics of biological nervous system and the tasks of versatile biological learning models. Its plain formations are neurons that can be interrelated in distinctive arrangements. It utilizes a nonparametric method. Accomplishment and accuracy depends on the grid formation and the quantity of inputs. There are many types of ANN classifier, but only few algorithm proven efficiency in neural network learning like multi-layer back propagation [65], [66]. The advantages of ANN is a data driven self-adaptive process, competently controls noisy inputs, calculation rate is good and its major problem is taking more time for training data and complexity in selecting the type grid architecture.
- Probabilistic neural network (PNN): There are input, output and hidden (summation) layer in Feed Forward Neural Network. Pattern layer formed by the input data set with the product of the weights tracked by the summing up layer that gets results related to the given class. The output layer contains the classification results. The main advantage of PNN classifier is its maximum training speed [67]. The scale factor of the exponential activation functions used to control the smoothing parameter (σ) of this classifier.
- Decision tree: In medical image study it is used as a attended categorizer. This process is comprised of 3 parts- Dividing the nodes, locate the terminal nodes and sharing of class labels to terminal nodes. A node in a tree represents a test for a exact attribute, and each part of that node represents the likely result of the test [68]. A pathway in the tree, from the root of the tree to an end leaf, details the categorization, with the ending leaf representing an object class. The Decision tree is based on the hierarchical statute based method and utilizes the nonparametric process. It is simple and computational efficiency is good, but becomes a difficult computation when diverse values are undecided or when variety of results are correlated.
- K-nearest neighbor (K-NN): It is a process to analyze image feature on basis of closest training illustrations in the feature space. It utilizes a separate calculate to make guess the class of the new test sample. This technique is one of the least complex of all machine learning calculation: a feature is arranged by a maximum vote of its neighbors, with the item being allocated to the class most ordinary amongst its k adjacent neighbors when k is small and the item is merely assigned to the class of its adjacent neighbor when K=1 [69].
- Bayesian neural network (BNN): It used in many areas of medicine. In US features prophetic of malignancy have been widely analyzed and the reactivity and specificity of these attributes for malignancy are easily obtainable [70]. The scheme of BNN is to cast the work of training a grid as a difficult of inference, which is sorted out utilizing Bayes theory [71]. A Bayesian neural network is extra optimized and strong in comparison to traditional neural grids, particularly when the training data set is not big.

3.5. Performance estimation

Quantitative measurement of system correctness is measured in term of true positive (TP), true negative (TN), false positive (FP), false negative (FN) with relation to positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity and accurateness [72]. It is given by:

$$PPV = \frac{(TP)}{(TP + FP)} \quad (1)$$

$$NPV = \frac{(TN)}{(TN + FN)} \quad (2)$$

$$\text{Sensitivity} = \frac{(TP)}{(TP + FN)} \quad (3)$$

$$\text{Specificity} = \frac{(TN)}{(TN + FP)} \quad (4)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN) + FP + FN} \quad (5)$$

- *TP* represents number of diseased lesions that is rightly classify as diseased.
- *TN* represents number of non-diseased lesions that is properly classify as non-diseased.
- *FP* represents number of non-diseased lesions that is incorrectly categorized as diseased.
- *FN* represents number of diseased lesions that is incorrectly categorized as non-diseased.
- The accuracy used to diagnose diseased and benign cases, the sensitivity calculated for the classification model to categorize diseased cases and the specificity used to evaluate benign cases.
- *PPV* represents percentage of predictive positives that is always positive.
- *NPV* represents percentage of predictive negatives that is always negative.

Some authors have evaluated their proposed system just by manual inspection performed by radiologist Specialists while others have exploited area under receiver operating characteristic (ROC) curve analysis. ROC applied to demonstrate the competence of the trade-off between the *TP* and the *FP* [73].

4. Related work

In recent years, the liver cancer analysis using CAD system has turned into a dynamic area of research. There are different approaches that are proposed for liver ailment analysis on the basis of medical picture analysis. In this section, we elaborate various techniques.

In article [62], a CAD system is proposed by Mittal et al. by which doctors can be guided in diagnosis of focal liver ailment from 2D mode US images. The suggested technique has been utilized for detecting and analysis of four types of focal liver ailment and compared them with normal liver. At the first image noise are reduced, then they divided the areas of interest to 800 segmented areas. Next, on the basis of the texture 208 features are extracted from each segmented region. Finally, they proposed use of Artificial Neural Networks (ANN) in reducing the training errors with two phases to diagnose the ailment. The general precision achieved by the CAD system was 86.4%.

Authors on paper [63] proposed an algorithm using Fuzzy Neural Network (FNN) to automatically characterize diffused liver diseases. For classification utilizing RUNL, GLDS, FOP, SGLDM and FDTA 12 texture features were taken out. Then again, the features were reduced to six utilizing multiple feature combinations. After that to produce blurry sets and create class edges in a statistical way voronoi diagram of training patterns was built which was utilized by FNN. The Authors showed 82.67% classification accuracy for verification using 150 liver images.

Design made from M-mode motion curve of liver and B-mode US liver picture on the basis of feature extraction by Guohui et.al. [74]. They took out 25 features utilizing M-mode movement curve through GLDS, FOP, RUNL, and a couple of additional extraordinary attributes. After taking out attribute, they used Fisher linear decision rule for choosing 20 helpful features depending on the minimum classification error. Experiment's outcome divulged that features gained utilizing movement curve were further reasonable for discerning ordinary or cirrhosis, liver in expressions of reactivity and specificity.

For US liver images categorization, Cao et al. suggested different process [75]. For taking out feature SGLDM and FDTA on 64x64 pixels sub-image were utilized by them. In this way the joint feature vector was gained, which was utilized to differentiate 273 sound and 99 fibrosis liver pictures. Fisher linear classifier and

SVM (leave-one-out calculation) were utilized. It was found effective in expression of categorization rate. Yet it is proved that the joint feature vector is a bit better.

The author's of paper [76], used an algorithm to recognize diffused liver ailment utilizing Gabor wavelet and categorized ultrasonic liver picture into usual, hepatitis and cirrhosis categories. Familiar three advantages of Gabor wavelets were used by them, which is invariance to swing of picture contents, maximum joint space frequency resolution, and littler feature vector. Attributes were extricated and pictures were categorized into various classes utilizing Gabor wavelet change, dyadic wavelet change, statistical moments and attributes.

Researchers under lead of Balasubramanian suggested a method [77] for automatically categorized benign, malignant, cyst and regular liver pictures utilizing texture attributes via TEM, SGLDM, RUNL, and Gabor wavelets. By manual selection and on the basis of Principle Component Analysis (PCA) on the basis of idea attributes, eight attributes were selected. K-means clustering calculation used by PCA based features whereas physically chosen attributes were categorized by BPN. Finally, it is proved that categorization outcome of BPN were improved than K-means. Poonguzhali et.al., [78] authors classified same liver diseases. The attribute taking out from the ROI of US pictures through Autocorrelation, TEM, Edge Frequency method and SGLDM. Optimal attribute sets selected from extracted features using PCA. For K-means categorization afterwards optimal features were utilized.

Jeon et al. presented a technique to classify focal liver lesion based on multiple ROI, to obtain more reliable and better classification performance [79]. This technique can be utilized to classify focal liver ailment, for example Hem, Cyst, and Malignancies. From the complete US image the ROI features are extricated at first. Lastly, the categorization of cysts and hemangiomas, categorization of cysts and malignancies, and categorization of hemangiomas and malignancies are classified using the SVM classifier. The preprocessing stage is complicated since it affects the subsequent stages and improves the quality of the images. Their method has shown the overall accuracy of 80%.

Ribeiro et.al., [80] implemented an algorithm using three dissimilar classifiers to classify the different chronic liver ailment. The classifiers utilized are SVM, KNN, and decision tree. The outcome showed that the SVM gained superior performance than the KNN and decision tree classifier. The classification's precision was 73.20%utilizing SVM with a radial basis kernel. Yet, the general accurateness of this process is not high. In another paper [81], the authors given a partly automatic method to categorizing unceasing liver ailment from US liver pictures. For this approach the data, which is collected from laboratories and clinic, are generated by utilized SVM classifier with a polynomial core of the fourth degree. The data achieved 91.67% of sensitivity better than previous approaches. In the coming works, they will extend their method of merging more textural features.

A few more compact summaries of different author's algorithms and accuracy of their proposed technique up to recent years presented in table 1.

Table 1. Summary of various researches on CAD for liver diagnosis

Authors/Year	Number of samples	Features	Classifier	Performance
Fayazul et al., /2012 [82]	88	Wavelet packet transform	SVM	~95%
Acharya et al., /2012 [83]	100	Wavelet and Higher order spectra feature	DT	93.3%
Jitender et al., /2012 [84]	56	Wavelet packet transform	SVM	88.8%
Jitender et al., /2013 [85]	31	Wavelet packet transform and Gabor Wavelet transform	SVM	98.3%
Jitender et al., /2013 [86]	108	FOS, GLCM, GLRLM, FPS, GWT, TEM	BPNN	87.7%

Jitender et al., /2013 [87]	108	FOS, GLCM, GLRLM, FPS, TEM, Gabor	SVM	87.2%
Nivedita et al., /2014 [88]	42	GLCM	SOM and MLP	81.5%
Jitender et al., /2014 [89]	108	FOS, GLCM, GLRLM, FPS, TEM, Gabor	Neural network ensemble	88.7-95%
Rivas et al., /2015 [90]	7	GLCM	Binary logistic regression	95.45%
Wu et al., /2015 [91]	288	Mean, SD, Kurtosis, Skew	SVM and Random forest	72.81%
Hwang et al., /2015 [92]	115	FOS, GLCM, TEM, Ecogenecity	Baysian regularity learning	96%
Acharya et al., /2016 [93]	100	GIST descriptor	PNN	98%

5. Conclusion

This study proposed in a new way of categorizing and summarizing the different stages of the computerized system scheme applied to ultrasound (US) with focus on liver cancer diagnosis. The up to date review of existing approaches in the literature has been reviewed. To the best of our knowledge, there has been no unique consensus on computer aided diagnosis (CAD) system. Many different algorithms mentioned in the state of art used to find and design optimal solution for automatic liver cancer diagnosis scheme. In our opinion, there should be a trade-off, strengths and weaknesses associated with the choice of the algorithm used for image analysis. In the future, researchers must pay attention to data pre-processing stage meanwhile minimizing motion artifacts, image noise and tolerable classification time using optimized neural network. It might be possible that integrating multiple effective techniques, potentially improve the general correctness, exactness and techniques concerned to speed of segmentation, also lessening the quantity of manual interactions of user. Moreover, greater part of the work in the literature concentrated on detection and classification of B-Mode US imaging. Our research in the future will be directed by introducing a novel CAD algorithm for 3D high resolution ultrasound imaging device that accurately characterize and detect liver lesions by including more features and classification techniques such as duct extension, Microlobulation and compressive sensing (CS) framework. In the upcoming years, CAD system will be a useful device for discovery sooner and treatment of liver lesions by radiologist and diagnostic examinations in everyday clinical work.

Acknowledgment

This work done by the support of the National Research Foundation of Korea (NRF) grant provided by the Korean government (NRF-2015R1A2A1A05001826).

References

- [1] K. Doi, "Overview on research and development of computer-aided diagnostic schemes," *Semin. Ultrasound, CT MRI*, vol. 25, no. 5, pp. 404–410, 2004.
- [2] M. L. Giger, "Computerized analysis of images in the detection and diagnosis of breast cancer," *Semin. Ultrasound, CT MRI*, vol. 25, no. 5, pp. 411–418, 2004.
- [3] H. Yoshida and A. H. Dachman, "Computer-aided diagnosis for CT colonography.," *Semin. Ultrasound. CT. MR*, vol. 25, no. 5, pp. 419–31, 2004.
- [4] Q. Li, F. Li, K. Suzuki, J. Shiraiishi, H. Abe, R. Engelmann, Y. Nie, H. MacMahon, and K. Doi, "Computer-aided diagnosis in thoracic CT.," *Semin. Ultrasound. CT. MR*, vol. 26, no. 5, pp. 357–363, 2005.

- [5] F. W. Kremkau and K. J. Taylor, “Artifacts in ultrasound imaging,” *J. Ultrasound Med.*, vol. 5, no. 4, pp. 227–237, 1986.
- [6] S. Marwaha, H. Monga, S. M. Student, and H. Cse, “Automatic Diagnosis Systems Using Image Processing- A systematic Study,” *IRACST -International J. Comput. Sci. Inf. Technol. Secur.*, vol. 2, no. 2, pp. 2249–9555, 2012.
- [7] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, “Computer-aided detection of prostate cancer in MRI,” *IEEE Trans. Med. Imaging*, vol. 33, no. 5, pp. 1083–1092, 2014.
- [8] Gunasundari S, “A Study of Textural Analysis Methods for the Diagnosis of Liver Diseases from Abdominal Computed Tomography,” *Int. J. Comput. Appl.*, vol. 74, no. 11, pp. 0975–8887, 2013.
- [9] A. Jemal, F. Bray, and J. Ferlay, “Global Cancer Statistics,” vol. 61, no. 2, 2011.
- [10] H. Yoshida, D. D. Casalino, B. Keserci, A. Coskun, O. Ozturk, and A. Savranlar, “Wavelet-packet-based texture analysis for differentiation between benign and malignant liver tumours in ultrasound images,” *Phys. Med. Biol.*, vol. 48, no. 22, pp. 3735–53, Nov. 2003.
- [11] M. E. Mayerhoefer, M. Breitenhofer, G. Amann, and M. Dominkus, “Are signal intensity and homogeneity useful parameters for distinguishing between benign and malignant soft tissue masses on MR images? Objective evaluation by means of texture analysis,” *Magn. Reson. Imaging*, vol. 26, no. 9, pp. 1316–22, 2008.
- [12] G. Xian, “An identification method of malignant and benign liver tumors from ultrasonography based on GLCM texture features and fuzzy SVM,” *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6737–6741, 2010.
- [13] R. S. C. Cobbold, *Foundations of Biomedical Ultrasound*. Oxford University Press, 2007.
- [14] S. Tang and S.-P. Chen, “An Effective Data Preprocessing Mechanism of Ultrasound Image Recognition,” *Bioinforma. Biomed. Eng. 2008. ICBBE 2008. 2nd Int. Conf.*, pp. 2708–2711, 2008.
- [15] Sk. Narayanan, “A View on Despeckling in Ultrasound Imaging,” *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 2, no. 3, 2009.
- [16] J.-S. Lee, “Digital Image Enhancement and Noise Filtering by Use of Local Statistics,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 2, pp. 165–168, 1980.
- [17] J.-S. Lee, “Refined filtering of image noise using local statistics,” *Comput. Graph. Image Process.*, vol. 15, no. 4, pp. 380–389, 1981.
- [18] D. Kuan, A. Sawchuk, T. Strand, and P. Chavel, “Adaptive restoration of images with speckle,” *IEEE Trans. Acoust.*, vol. 35, no. 3, pp. 373–383, Mar. 1987.
- [19] R. C. Gonzalez and R. E. Woods, “Digital Image Processing,” *Addison-Wesley Publ. Co.*, 2002.
- [20] W. N. M. and P. L. A. T. Loupas, “An Adaptive Weighted Median Filter for Speckle Suppression in Medical Ultrasound Images,” *IEEE Trans. Circuits Syst.*, vol. 36, no. January, pp. 129–135, 1989.
- [21] R. C. G. and R.E.Woods, “Digital Image Processing,” *3rd ed. Pearson Educ.*, 2008.
- [22] A. K. Jain, *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.
- [23] A. Lopes, R. Touzi, and E. Nezry, “Adaptive Speckle Filters and Scene Heterogeneity,” *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 6, pp. 992–1000, 1990.
- [24] a. Lopes, E. Nezry, R. Touzi, and H. Laur, “Maximum A Posteriori Speckle Filtering And First Order Texture Models In Sar Images,” *10th Annu. Int. Symp. Geosci. Remote Sens.*, pp. 8–11, 1990.
- [25] V. S. Frost, J. A. Stiles, K. S. Shanmugan, and J. C. Holtzman, “A model for radar images and its application to adaptive digital filtering of multiplicative noise,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 4, no. 2, pp. 157–66, 1982.
- [26] D. T. Kuan, a a Sawchuk, T. C. Strand, and P. Chavel, “Adaptive noise smoothing filter for images with signal-dependent noise,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 2, pp. 165–177, 1985.
- [27] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” *IEEE Trans.*

- Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, 1990.
- [28] R. de Luis-García, R. Deriche, M. Rousson, and C. Alberola-López, “Tensor Processing for Texture and Colour Segmentation,” in *Image Analysis*, 2005, pp. 1117–1127.
- [29] T. D., “Anisotropic diffusion PDE’s for image regularization and visualization,” in *Handbook of Mathematical Methods in Imaging, 1st Edition*, Springer, 2010.
- [30] C. Tomasi and R. Manduchi, “Bilateral Filtering for Gray and Color Images,” *Int. Conf. Comput. Vis.*, pp. 839–846, 1998.
- [31] S. T. Acton, “Deconvolutional speckle reducing anisotropic diffusion,” *Proc. - Int. Conf. Image Process. ICIP*, vol. 1, no. 11, pp. 5–8, 2005.
- [32] S. Mallat, “A Wavelet Tour of Signal Processing,” *2nd ed. Acad. Press*, 1999.
- [33] X. Q. Gao, T. Q. Nguyen, and G. Strang, “A study of two-channel complex-valued filterbanks and wavelets with orthogonality and symmetry properties,” *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 824–833, 2002.
- [34] J. Neumann and G. Steidl, “Dual-Tree complex wavelet transform in the frequency domain and an application to signal classification,” *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 03, no. 01, pp. 43–65, 2005.
- [35] N. Kingsbury, “Complex Wavelets for Shift Invariant Analysis and Filtering of Signals,” *Appl. Comput. Harmon. Anal.*, vol. 10, pp. 234–253, 2001.
- [36] N. Kingsbury, “Image processing with complex wavelets,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 357, no. 1760, pp. 2543–2560, 1999.
- [37] M. N. Do and M. Vetterli, “Contourlets: a directional multiresolution image representation,” in *Proceedings. International Conference on Image Processing*, 2002, vol. 1, pp. I–357–I–360.
- [38] Emmanuel J. Candès and David L. Donoho, *Curvelets – A Surprisingly Effective Nonadaptive Representation For Objects with Edges*, Saint-Malo. Vanderbilt University Press, 2000.
- [39] A. M. Eskicioglu and P. S. Fisher, “Image quality measures and their performance,” *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, 1995.
- [40] E. J. Can and D. L. Donoho, “Ridgelets: a key to higher-dimensional intermittency?,” *Phil. Trans. R. Soc. Lond.A*, vol. 14, pp. 423–438, 1999.
- [41] E. J. Candès and D. L. Donoho, “New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities,” *Commun. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, 2004.
- [42] K. D. Toennies, “Guide to Medical Image Analysis Methods and Algorithms,” *Adv. Comput. Vis. Pattern Recognit.*, no. 2191–6586, pp. 1–19, 2012.
- [43] T. Sheela and M. S. Abirami, “Analysis of Image Segmentation Techniques for Medical Images,” in *Proceedings of International Conference on Emerging Research in Computing, Information, Communication and Applications*, 2014.
- [44] E. P. Thakur and E. N. Madaan, “A survey of image segmentation techniques,” *Int. J. Res. Comput. Appl. Robot. www.ijrcar.com*, vol. 24, pp. 158–165, 2014.
- [45] M. Suganthi and P. Elayaraja, “A survey of image segmentation techniques,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 11, 2014.
- [46] R. Adams and L. Bischof, “Seeded region growing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.
- [47] R. M. S. and K. S. Selvanayaki, “A Survey on Image Segmentation Techniques for Edge Detection,” in *International Conference on Innovation in Communication, Information and Computing (ICICIC)*, 2013.
- [48] A. A. M. Rajiv Kumar, “A Comparative Study of Image Segmentation Using Edge-Based Approach,” *Int. J. Math. Comput.*, vol. 7, no. 3, 2013.

- [49] S. Saleh Al-Amri, N. V Kalyankar, and K. S. D, “Image Segmentation by Using Thershod Techniques,” *J. Comput.*, vol. 2, no. 5, pp. 2151–9617, 2010.
- [50] V. D. Santanu Bhowmik, “A Survey on Clustering Based Image Segmentation,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 5, pp. 2278–1323, 2012.
- [51] O. Oke, T. Adedeji, O. Alade, and E. Adewusi, “Fuzzy kc-means Clustering Algorithm for Medical Image Segmentation,” *J. Inf. Eng. Appl.*, vol. 2, no. 6, pp. 21–33, 2012.
- [52] R. R. and K. Tejaswini, “A Survey of Image Segmentation Algorithms Based On Fuzzy Clustering,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 2, pp. 200–206, 2013.
- [53] I. Levner, V. Bulitko, and G. Lin, “Feature Extraction for Classification of Proteomic Mass Spectra: A Comparative Study,” in *Feature Extraction*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 607–624.
- [54] K. I. Laws, “Texture energy measures,” in *Image understanding workshop*, 1979, pp. 47–51.
- [55] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Trans. Syst. Man. Cybern.*, vol. 3, no. 6, pp. 610–621, 1973.
- [56] M.M.Galloway, “Texture classification using gray level run lengths,” *Comput. Graph. Image Process.*, vol. 4, pp. 172–179, 1975.
- [57] D. Gaitini, Y. Baruch, E. Ghersin, E. Veitsman, H. Kerner, B. Shalem, G. Yaniv, C. Sarfaty, and H. Azhari, “Feasibility study of ultrasonic fatty liver biopsy: Texture vs. attenuation and backscatter,” *Ultrasound Med. Biol.*, vol. 30, no. 10, pp. 1321–1327, Oct. 2004.
- [58] W.-C. Yeh, S.-W. Huang, and P.-C. Li, “Liver fibrosis grade classification with B-mode ultrasound,” *Ultrasound Med. Biol.*, vol. 29, no. 9, pp. 1229–35, Sep. 2003.
- [59] Z. Jiang, K. Yamauchi, K. Yoshioka, K. Aoki, S. Kuroyanagi, A. Iwata, J. Yang, K. Wang, Z. Jiang, K. Yamauchi, J. Yang, . K. Wang, K. Yoshioka, K. Aoki, S. Kuroyanagi, and A. Iwata, “Support Vector Machine-Based Feature Selection for Classification of Liver Fibrosis Grade in Chronic Hepatitis C,” *J Med Syst*, vol. 30, pp. 389–394, 2006.
- [60] M. S. Gebbinck, J. T. Verhoeven, J. M. Thijssen, and T. E. Schouten, “Application of neural networks for the classification of diffuse liver disease by quantitative echography,” *Ultrason. Imaging*, vol. 15, no. 3, pp. 205–17, 1993.
- [61] N. H. K.Ogawa,k. Kubota, “Computer-aided Diagnostic System for Diffuse Liver Diseases with Ultrasonography by Neural Networks,” *IEEE Trans. Nucl. Sci.*, vol. 45, no. 6, pp. 3069–3074, 1998.
- [62] D. Mittal, V. Kumar, S. C. Saxena, N. Khandelwal, and N. Kalra, “Neural network based focal liver lesion diagnosis using ultrasound images,” *Comput. Med. Imaging Graph.*, vol. 35, no. 4, pp. 315–323, 2011.
- [63] S. Pavlopoulos, E. Kyriacou, D. Koutsouris, K. Blekas, A. Stafylopatis, and P. Zoumpoulis, “Fuzzy neural network-based texture analysis of ultrasonic images,” *IEEE Eng. Med. Biol. Mag.*, vol. 19, no. 1, pp. 39–47, 2000.
- [64] S.-T. J. Christianini N, “An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods,” *United Kingdom: Cambridge University Press*, 2000.
- [65] Y. Hirose, K. Yamashita, and S. Hijiya, “Back-propagation algorithm which varies the number of hidden units,” *Neural Networks*, vol. 4, no. 1, pp. 61–66, 1991.
- [66] G. E. H. and R. J. W. David E.Eumelhart, “Learning representation by back-propagation errors,” *Lett. to Nat.*, vol. 323, pp. 533–536, 1986.
- [67] P.D.Wasserman, *Advanced Methods in Neural Computing*. Van Nostrand Reinhold, 1993.
- [68] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Chichester, UK: John Wiley & Sons, Ltd, 2004.

- [69] Y. M. Kadah, A. A. Farag, J. M. Zurada, A. M. Badawi, and A. B. M. Youssef, "Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images," *IEEE Trans. Med. Imaging*, vol. 15, no. 4, pp. 466–478, 1996.
- [70] Y. I. Liu, A. Kamaya, T. S. Desser, and D. L. Rubin, "A Bayesian classifier for differentiating benign versus malignant thyroid nodules using sonographic features.," in *AMIA symposium proceeding*, 2008, pp. 419–23.
- [71] H. B. P. P.C.Bhat, "Bayesian Neural Networks," in *Proceedings of PHYSTAT 05*, 2006, pp. 151–155.
- [72] W. Zhu, N. Zeng, and N. Wang, "Sensitivity , Specificity , Accuracy , Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations K & L consulting services , Inc , Fort Washington , PA Octagon Research Solutions , Wayne," *NESUG Heal. Care Life Sci.*, pp. 1–9, 2010.
- [73] J. R. Beck and E. K. Shultz, "The use of relative operating characteristic (ROC) curves in test performance evaluation.," *Arch. Pathol. Lab. Med.*, vol. 110, no. 1, pp. 13–20, 1986.
- [74] G. Zhou, Y. Wang, W. Wang, Y. Sun, and Y. Chen, "Decision of Cirrhosis Using Liver's Ultrasonic Images.," in *Proceedings of the 2005 IEEE Engineering in Medicine and Biology*, 2005, vol. 4, pp. 3351–4.
- [75] G. Cao, P. Shi, and B. Hu, "Liver fibrosis identification based on ultrasound images.," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 2005, vol. 6, pp. 6317–20.
- [76] A. Ahmadian, A. Mostafa, M. Abolhassani, and Y. Salimpour, "A texture classification method for diffused liver diseases using Gabor wavelets.," in *IEEE Engineering in Medicine and Biology Society.*, 2005, vol. 2, pp. 1567–70.
- [77] D. Balasubramanian, P. Srinivasan, and R. Gurupatham, "Automatic classification of focal lesions in ultrasound liver images using principal component analysis and neural networks.," in *IEEE Engineering in Medicine and Biology Society*, 2007, pp. 2134–7.
- [78] S. Poonguzhali, B. Deepalakshmi, and G. Ravindran, "Optimal Feature Selection and Automatic Classification of Abnormal Masses in Ultrasound Liver Images," in *2007 International Conference on Signal Processing, Communications and Networking*, 2007, pp. 503–506.
- [79] J. H. Jeon, J. Y. Choi, S. Lee, and Y. M. Ro, "Multiple ROI selection based focal liver lesion classification in ultrasound images," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 450–457, 2013.
- [80] R. Ribeiro, R. Marinho, J. Velosa, F. Ramalho, and J. M. Sanches, "Chronic liver disease staging classification based on ultrasound, clinical and laboratorial data," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 707–710.
- [81] R. Ribeiro, R. T. Marinho, J. Velosa, F. Ramalho, J. M. Sanches, and J. S. Suri, "The usefulness of ultrasound in the classification of chronic liver disease.," in *IEEE Engineering in Medicine and Biology Society*, 2011, pp. 5132–5.
- [82] F. U. A. A. Minhas, D. Sabih, and M. Hussain, "Automated classification of liver disorders using ultrasound images," *J. Med. Syst.*, vol. 36, no. 5, pp. 3163–3172, 2012.
- [83] J. S. S. U. Rajendra Acharya, S. Vinitha Sree, Ricardo Ribeiro, Ganapathy Krishnamurthi, Rui Tato Marinho, João Sanches, "Data mining framework for fatty liver disease classification in ultrasound: A hybrid feature extraction paradigm," *Med. Phys.*, vol. 39, 2012.
- [84] J. Virmani, V. Kumar, N. Kalra, and N. Khandelwal, "SVM-based characterization of liver ultrasound images using wavelet packet texture descriptors," *J. Digit. Imaging*, vol. 26, no. 3, pp. 530–543, 2012.
- [85] N. K. Jitendra Virmani, Vinod Kumar Naveen Kalra, "Prediction of liver cirrhosis based on multiresolution texture descriptors from B-mode ultrasound," *Int. J. Conver. Comput.*, vol. 1, pp. 1–19, 2013.
- [86] J. Virmani, V. Kumar, N. Kalra, and N. Khandelwal, "A comparative study of computer-aided

- classification systems for focal hepatic lesions from B-mode ultrasound,,” *J. Med. Eng. Technol.*, vol. 37, no. 4, pp. 292–306, 2013.
- [87] N. K. Jitendra Virmani, Vinod Kumar, Naveen Kalra, “PCA -SVM based CA D System for Focal Liver Lesions using B-Mode Ultrasound Images,” *Def. Sci. J.*, vol. 63, pp. 478–486, 2013.
- [88] M. Roy, “Classification of Ultrasonography Images of Human Fatty and Normal Livers using GLCM Textural Features,” vol. 4, 2014.
- [89] V. J., K. V., K. N., and K. N., “Neural network ensemble based CAD system for focal liver lesions from B-mode ultrasound,” *J. Digit. Imaging*, vol. 27, no. 4, pp. 520–537, 2014.
- [90] V. Morocho, P. Vanegas, and R. Medina, “Hepatic Steatosis Detection Using the Co-occurrence Matrix in Tomography and Ultrasound Images,” in *2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*, 2015, pp. 1–7.
- [91] J. Y. Wu, M. Beland, J. Konrad, A. Tuomi, D. Glidden, D. Grand, and D. Merck, “Quantitative ultrasound texture analysis for clinical decision making support,” in *Medical Imaging 2015: Ultrasonic Imaging and Tomography*, 2015, vol. 9419, p. 94190W.
- [92] Y. N. Hwang, J. H. Lee, G. Y. Kim, Y. Y. Jiang, and S. M. Kim, “Classification of focal liver lesions on ultrasound images by extracting hybrid textural features and using an artificial neural network,” in *Bio-Medical Materials and Engineering*, 2015, vol. 26, pp. S1599–S1611.
- [93] U. R. Acharya, H. Fujita, S. Bhat, U. Raghavendra, A. Gudigar, F. Molinari, A. Vijayanathan, and K. Hoong Ng, “Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images,” *Inf. Fusion*, vol. 29, pp. 32–39, 2016.

A Cooperative Wireless Sensor Network for Indoor Industrial Monitoring

Zafar Iqbal, Kiseon Kim, *Senior Member, IEEE*, and Heung-No Lee, *Senior Member, IEEE*

Abstract—Industrial wireless sensor networks (IWSNs) are getting popular for indoor monitoring of heavy machinery and large factories to make a reliable decision on the state of machines in a certain area of interest. However, the indoor wireless communication channel is not always reliable, and observations of some sensors cannot be reported successfully to the base station. In order to deal with this problem, we propose a cooperative WSN scheme by introducing a novel cooperation mechanism and a medium access control (MAC) protocol. The proposed scheme effectively increases the probability of correct decision about the state of the machine, reduces the probability of false alarms at a given signal level, and reduces the overall energy consumption as compared to non-cooperative schemes. We also present a closed-form expression for the symbol-error rate analysis of the proposed scheme, which shows that our proposed scheme achieves full diversity order offered by the cooperation scheme.

Index Terms—machine condition monitoring, industrial wireless sensor networks, cooperative communication, medium access control, indoor industrial monitoring.

I. INTRODUCTION

WIRELESS sensors are widely used for machine condition monitoring (MCM) and maintenance, especially the machines which are located in inaccessible areas or are hard to be monitored by human, such as nuclear plants, unmanned underwater vehicles (UUVs), or in large factories. In addition, wireless sensors are also used for environmental monitoring, surveillance, healthcare, and security services [1]. But these sensors are prone to failure and the wireless communication channel may also fail sometimes due to severe conditions. Therefore, it is a good idea to use cooperation among the sensor nodes that communicate with a central base station to ensure the accuracy and timeliness of information gathered from the nodes [2].

The wireless communication channel in indoor industrial environment suffers from severe conditions such as propagation loss, time variation, and multipath fading etc. The quality of a wireless communication link is very important for

transmitting the information collected by the sensors to a central signal processing unit without significant amount of error. There has been a lot of research on multipath fading in wireless networks and channel characterization for industrial environments, such as [3], and a work that deals with the underground link quality characteristics [4]. However, a recent work and the references therein, provides a suitable channel model for indoor industrial environments [5].

A bad communication link results in higher energy consumption because of repeated transmissions or use of higher transmit power by the nodes, and reduces the overall throughput of the network. Similarly, the amount of data transmitted by the network nodes and the amount of processing at the receiver also contributes towards the energy consumption per bit of the network. Various techniques have been proposed to deal with these issues in wireless networks, such as user cooperation in communication [6]-[8] for improved spatial diversity, time-slot reassignment [9], and sleep scheduling [10] strategies used to improve the energy efficiency of the WSN. In the case of cooperation among sensor nodes, data aggregation at the intermediate nodes is an important factor of multi-hop communication. Since the size of data packets is usually small and are addressed to a single destination, therefore, reducing the number of transmissions and the size of control packet overhead, improves the energy efficiency and throughput of the system [11].

In this paper, we propose a cooperation scheme for IWSNs, in which the network consists of small cooperation groups of sensors. Each node in the cooperation group shares its information with all others in the first phase. In the second phase each node forms a cooperative data packet and sends it to the base station (BS). In this way, the nodes help relay information for its neighbor nodes with a significant reduction in energy consumption at the cost of an acceptable reduced throughput.

A. Related Works and Contributions of this Paper

Recently, network coding has become one of the most widely used techniques for cooperation among nodes in a wireless communication network. Some works that deal with improving the energy efficiency and packet delivery ratio include, a Reliable Reactive Routing Enhancement (R3E) algorithm for IWSN, which finds a guide path towards the sink and provides a reliable and energy-efficient packet delivery against the unreliable wireless links [2]. A physical-layer cooperative transceiver, which can use either amplify-and-

Manuscript received Feb. 12, 2016; revised Jul. 06, 2016 and Aug. 19, 2016; accepted Sep. 14, 2016.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2015R1A2A1A05001826).

The authors are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, 61005, South Korea (Phone: 0082-62-715-2237, emails: {zafar, kskim, heungno}@gist.ac.kr).

forward (AF) or decode-and-forward (DF) relaying to improve the packet error rate, was proposed in [6]. The work in [7] presents an adaptive-gain M -relay AF cooperative system with conventional relay (CR) and best relay (BR) selection schemes and shows that the BR scheme provides higher asymptotic error limits than that of the CR scheme. A generalized dynamic-network code (GDNC) for a network of M users sending independent information to a common base station using independent block fading channels was proposed in [8]. The proposed scheme offers a much better tradeoff between rate and diversity as compared to the DNC. Similarly, [12] presents a selective cooperative relaying protocol with periodic, adaptive, and reactive relay selection mechanism. The scheme improves packet delivery ratio and reduces the number of retransmissions for successful delivery.

An adaptive and energy-efficient TDMA-based MAC protocol called receiver-driven MAC (RMAC), which uses a timeslot stealing and timeslot reassignment mechanism, was proposed in [9]. RMAC performs better in terms of average packet delay and average power consumption per packet as compared to S-MAC. An energy-aware sleep scheduling mechanism for wireless sensor networks was presented in [10], which significantly reduces the variation in energy level among sensors and extends the lifetime of the network by around 18%. A practical wireless model-based predictive networked control system (W-MBPNCs) was proposed in [13], in order to achieve a decent control under severe impairments, such as unbounded delay, burst of packet loss, and ambient wireless traffic.

Another solution, used for MCM in large factories, distributes the signal processing operations among the central unit and the sensor nodes to reduce the energy consumption in data transmission and improve the network throughput, was proposed in [14]. Similarly, [15] presents an IWSN-based MCM system which overcomes false alarms caused by loss of data, interference, or invalid data. An improvement in the SNR and false alarm detection rate, after Dempster-Shafer Theory (DST)-based fusion method, was observed.

Most of the above-mentioned works use cooperative and selective relaying to improve packet delivery and energy consumption of the network. However, relay selection comes with an extra overhead of reduced network throughput and the problem becomes more evident in the case of multi-hop and multiple cluster sensor networks. We propose a method in which the cooperation groups are fixed in the organization stage of the network. A source node acts as a relay node in the cooperative phase of transmission. A relay node uses data aggregation and AF relaying to send the cooperative packet to the BS. The contributions of this paper are as follows,

- We propose a novel two-phase cooperation scheme that works in a dual-hop manner.
- Our proposed scheme does not involve the extra overhead of relay selection and retransmission to ensure successful packet delivery, unlike [2], [7], and [12].
- The relay does not need to check whether the data was

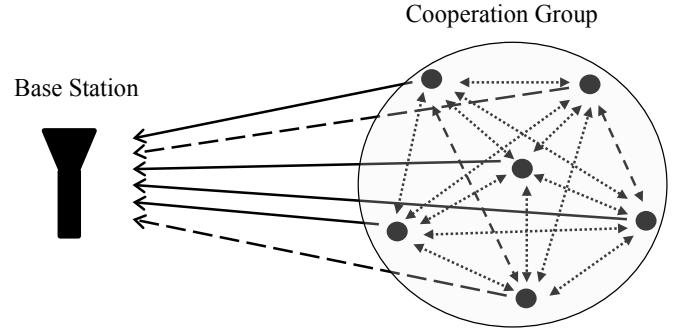


Fig. 1. Cooperative wireless sensor network

- correctly received. It forwards the detected binary symbols without regard to the error induced in it in the first hop.
- We also propose a TDMA-based MAC protocol for the organization and operation of the sensor network.
- A closed-form expression has been derived for the symbol error rate analysis and it is shown that the proposed method achieves full diversity order.
- We have carried out the throughput and energy consumption analysis to show the effectiveness of the proposed scheme.

The rest of the paper is organized as follows. Section II describes the network design. Section III explains the MAC and cross-layer design. Section IV explains the fusion mechanism. Section V presents performance analysis. Section VI presents the simulation results, and Section VII concludes the paper.

II. NETWORK DESIGN

Fig. 1 shows a cooperative WSN where the sensor nodes share their information with each other in the first phase and send the cooperative information to the base station in the second phase. Inter-sensor channels are shown by dotted lines while the channels from sensor to base station are shown by solid lines. We assume that some of the communication links between the sensors and from a sensor to the base station might be broken at a particular time instance, shown by long-dashed lines.

A. Sensor Deployment

Sensor deployment deals with the problem of coverage and connectivity of the sensor network while minimizing the power consumption for prolonged network lifetime and to transmit the sensed data timely and efficiently to the BS. In our case of indoor industrial area monitoring, the sensors could be deployed according to a pre-planned location map around huge machines in the factories. Considering these scenarios, our coverage problem becomes a static coverage problem, where the nodes do not change their positions. Assume that the sensing range of a sensor is r , the minimum number of sensors required to cover the area of interest [16], is given as,

$$\frac{N \times r^2 \pi}{P_{AREA}} = \frac{2\pi}{\sqrt{27}} \quad (1)$$

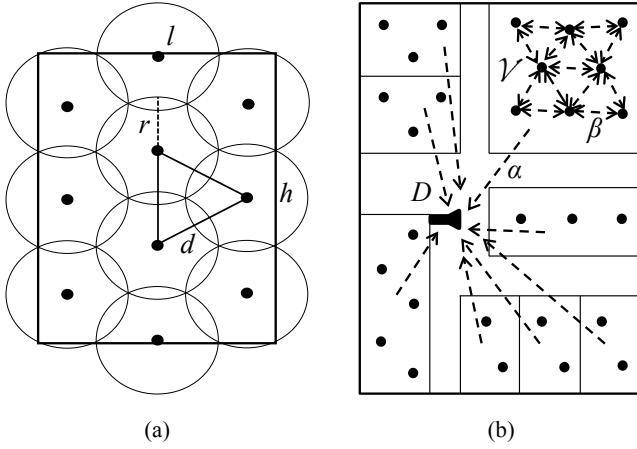


Fig. 2. Sensor nodes deployed in a rectangular area. (a) Triangular-grid, ensuring the coverage of the whole area with minimal overlap. (b) Indoor communication scenario showing a floor layout.

where N is the minimal number of nodes needed to cover the area of interest, P_{AREA} . This kind of optimal regular deployment is shown in Fig. 2(a). Every three nodes, whose sensing ranges intersect, form an equilateral triangle with each side $d = r\sqrt{3}$.

In order to ensure connectivity, we use the argument in [17] for minimum number of neighbor nodes, which says that for a network to be connected, $\Theta(\log N)$ ($0.074\log N$ to $5.1774\log N$) neighbors are necessary and sufficient. Therefore, we choose the minimum number of neighbors for a sensor to be equal to 6, with which it can communicate in a single-hop manner. This is used to enable cooperative transmission to the base station for a combined decision on the sensed data.

B. Sensor Localization

Sensor localization is used to locate the sensor positions and time of the observed information in the network. We consider a medium-sized fixed sensor network. Each sensor contains its local coordinate information, which is sent to the base station along with its observation. The local coordinate system (LCS) field in the transmitted packet contains geographic coordinates and floor number in case of multi-story buildings. The received signal strength (RSS) and angle-of-arrival (AOA) information could be used to find the sensors' distance and angular position, respectively, but the LCS field in the transmitted packet already contains the position information. Therefore, the only information that needs to be determined is the time of the event. Thus, along with the LCS, we use the time-of-arrival (TOA) information in order to locate the time of the event. This will help us localize the received information in both time and geographical location of the observation.

The transmitted packet structure and alarm information by each sensor is shown in Fig. 3. This information will be decoded at the base station by the fusion center to find out the nature of the observations at a particular location in the network and activate response mechanisms on time.

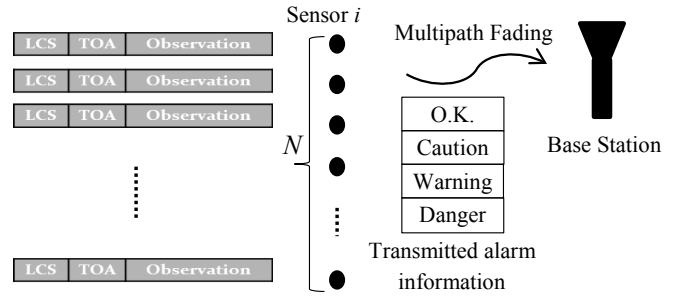


Fig. 3. Information transmission to the base station.

C. Time Synchronization

In our problem of a medium-sized network, most of the computations are done by the base station and the sensors are supposed to be in harsh environmental conditions which make it difficult for fine-grained synchronization algorithms to be used. Therefore, we adapt the Wisden system [18] of coarse-grained synchronization. In our synchronization technique, each sensor records the delay from the time of generation/reception of a sample to the time it is transmitted to the next hop or BS. Also, the cooperating node will record its own time delay for the packet that it processes before sending it to the base station. The TOA field in the transmitted packet contains this time delay information of all the nodes that the packet has traversed before reaching the base station.

Assume that the time spent by each packet k at the sensor node i is λ_i^k . Let the number of hops the packet traverses be n , and the time of arrival of the packet k , at the base station D , be T_D^k . Then, the start time of the packet at the origin node s , can be calculated as,

$$T_s^k = T_D^k - \sum_{i=1}^n \lambda_i^k. \quad (2)$$

The second term in (2) represents the time spent by the packet in the network. Thus, we can get the time of origination of the observation at the base station by subtracting the total time spent in the network from the current time at the BS. The BS is assumed to have an accurate reference clock periodically synchronized with the GPS time reference while each sensor node has its own local clock. This method of achieving time synchronization is simple, cost-effective, and robust to many sources of latency that contribute to error but is vulnerable to varying clock drifts in the intermediate nodes. But we assume a well-maintained medium-sized network of nodes and a moderate accuracy requirement; therefore, clock drift is not a very critical issue and can be traded off with the simplicity of the approach.

D. Wireless Link Characteristics

We consider a medium-sized indoor industrial WSN with mixed line-of-sight (LOS) and non-line-of-sight (NLOS) configurations, therefore we will use the statistical one-slope radio propagation model for path-loss [5], given as

$$P_{r,dB}(d) = P_{t,dB} - PL_{dB}(d) \quad (3)$$

$$PL_{dB}(d) = PL_{dB}(d_0) + 10\eta \log_{10}\left(\frac{d}{d_0}\right) + X_{\sigma,dB}$$

where, $P_{r,dB}$ and $P_{t,dB}$ are the received and transmitted powers in dB, PL_{dB} is the Friis free-space path-loss in dB with distance d from the transmitter, η is the path-loss exponent indicating the rate of decay of the mean signal with respect to distance, d_0 is a reference distance, and $X_{\sigma,dB}$ is a zero-mean Gaussian random variable with standard deviation σ . The model in (3) provides a very good approximation for the indoor industrial wireless channel by considering the multipath and shadowing effects present in the environment. However, the values of η and σ need to be carefully chosen according to the environment, as described in [5] and the references therein.

Fig. 2(b) shows a floor map of a building with sensors scattered all over the floor that communicate to a common BS. Each link in the network is modeled by using (3) and incorporating η and σ with respect to indoor communication scenario. The inter-node channels, β_i , and the node-destination channels, α_i , are modeled as lognormal distributed Rayleigh fading channels.

The wireless nodes are clustered into different cooperation groups by their geographic locations. The cooperative transmission is done within each cooperation group, $\mathcal{V} = \{V_i\}_{i=1}^{\mathcal{N}}$, where \mathcal{N} is the maximum number of nodes in a cooperation group. The wireless nodes V_i 's in a cooperation group are physically close to each other and the destination node is relatively far away from the group. Further assumptions are that the channels from each node V_i to the destination D is modeled as a lognormal fading channel with fading coefficient α_i , which is assumed to be fixed for a sufficiently longer period of time. As the group of wireless nodes is collocated and the destination is relatively far away, the fading coefficients α_i 's are assumed to have the same average magnitude determined by the path loss from V_i to D . Also, the fading channels from V_i to D , are independent, thus, the fading coefficients α_i 's, from V_i to D , are i.i.d. lognormal random variables. The channel from a transmitting node V_i to a node V_j within a group are also modeled as lognormal fading channels with fading coefficients $\beta_{i,j}$. To further simplify the analysis, it is assumed that the relative distance among these nodes is almost the same. Under this assumption, $\beta_{i,j}$'s are also i.i.d. lognormal random variables with the same average magnitude determined by the path loss among them.

III. MAC AND CROSS-LAYER DESIGN

Some of the major sources of energy wastage in WSNs are packet collisions, overhearing, packet overhead, and idle listening [19]. In order to reduce the energy loss to collisions and overhearing, we will use a TDMA-based MAC scheme in a two-phase communication model. Scheduling reduces packet collisions over the air, while the overheard information by

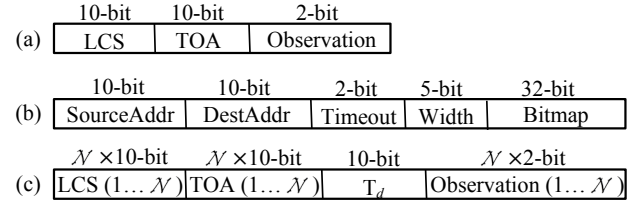


Fig. 4. (a) Data packet structure of each node in *Phase 1*. (b) Schedule packet structure used for organization of the network. (c) Data packet structure of each node in *Phase 2* for cooperative data transmission. (d) Cooperation group of 12 nodes in the sensor network.

each sensor is used to reduce the error rate and improve data transfer to the base station, which helps reduce the energy wastage in the network. For a medium-sized network of fixed sensors, we propose this protocol to meet our needs of scheduling, to reduce the header length and computation time.

A. Design of the Data Packet

As mentioned earlier, the data packet contains LCS, TOA, and Observation fields. The LCS contains the location information of the sensor which is embedded in it during the network deployment stage. It contains the following information fields:

- *Floor Number*: Although this parameter may vary according to the design under consideration, we choose this value to be 3 bit in order to cover up to 8-story buildings with our design.
- *Sensor ID*: Each sensor on a floor is assigned a unique identification number. We assign 7 bits to this field to allow up to 128 sensors on each floor of the building.
- *TOA*: This field is of 10 bits and contains the time duration between the time-of-arrival/observation of information on the current sensor and the time it was transmitted.
- *Observation*: The 2-bit observation field contains the alarm information, i.e., OK, Caution, Warning, and Danger.

The resulting data packet is a 22-bit packet as shown in Fig. 4(a). If Δ is the total time taken by the network to transmit one sensing event to the base station in a TDMA manner, then in the worst-case scenario, each sensor may have to wait for Δ seconds before it can send its data to the base station.

B. Design of the Schedule Packet

In our proposed scheme, each sensor transmits a schedule packet to select the winner of a given time slot, which is called Schedule, as shown in Fig. 4(b).

The source and destination node addresses of the current schedule packet are called SourceAddr and DestAddr, respectively. These fields contain the LCS information of the corresponding nodes and are therefore 10 bits each.

Timeout is used to resend the scheduling information to the next node in case it did not respond at the first time. The number of retries is limited to 4, after which the node is considered dead, its bitmap is set to 1, and the DestAddr is changed to next node in the schedule.

Width defines the number of nodes in a cooperation group and in turn the size of the data packet each node has to send to

the BS. This value is set to 5 bit in our design for a maximum of 32 nodes in the cooperation group.

Bitmap contains a bitmap of all the sensors in the network. A '0' in the bitmap means the node is not organized and a '1' means the node is organized in the network. A node ID corresponding to the bitmap is saved at each node in the network, so that it knows which bit represents which node in the network.

C. Network Organization

Our proposed network consists of fixed nodes and therefore mobility issues are not considered. Further, the network is assumed to have local groups of nodes that communicate cooperatively with the destination in a dual-hop manner. There are no cluster-heads formed because all the nodes in a group will schedule their communication links independently and in collaboration with other nodes in their vicinity. As mentioned earlier, a node is able to communicate with a minimum number of neighboring nodes in the network. Each node keeps a list of 6 to 20 neighboring nodes by saving the source address of these nodes which will be broadcasted using a low-frequency control channel. A node will decode the received information only from the nodes within its neighbor list. The rest of the received information will be discarded. The cooperation group will be updated periodically depending upon the application and conditions of the sensor nodes.

Based on the above described scheme, we propose a TDMA-based MAC protocol for the operation of the WSN. It consists of two main steps, organization of the nodes and operation of the network, and therefore referred to as Organize and Operate Protocol (OOP). The OOP is described as follows,

1. Organize

- (i) *BS sends Organize message to all the nodes in the network, using the Schedule packet described earlier.*
 - *The Bitmap is set to all 0's, i.e., none of the nodes is organized as yet.*
 - *It also contains the address of the first node to start the Organize process from. This address will be generated randomly on each Organize message.*
- (ii) *Upon receiving the Organize message, each node turns to Organize mode, i.e.,*
 - *Stop all the current transmit/receive operations.*
 - *Update its current list of neighbor nodes.*
 - *Listen to the received Schedule packet from neighbor node.*
 - *After receiving a Schedule packet, each node updates its information in the Schedule packet, sets its corresponding bit in the Bitmap field to '1' and passes the Schedule packet to the next node.*
- (iii) *When the bitmap becomes all 1's,*
 - *The current node transmits this information to the BS, by setting the DestAddr to that of the BS.*
 - *The BS, upon receiving this packet, sends a global message to all the nodes indicating to start normal sense and transmit operations mode, called Operate.*

2. Operate

Upon receiving the Operate message from BS, each sensor then,

- (i) *Senses the surrounding environment and wait for its turn to transmit.*
- (ii) *Shares the data with the nodes in its neighbor list.*
 - *Each node in the neighbor list receives this data and stores it in its local memory.*
- (iii) *Upon receiving the data from all the nodes in its neighbor list,*
 - *IF this was the first node in Organize stage*,
Transmit the cooperative data packet to the BS,
ELSE
Wait for its turn to transmit.*
- (iv) *Go to (i)*

**Each node stores the bitmap and next node DestAddr information during the Organize stage, which is also used in the operation scheduling.*

D. Cooperative Communication

Assume that the network is organized in sub groups of nodes that cooperate with each other, called cooperation group. We use the AF relaying protocol at the relays. The communication is done in two phases, as follows,

1) Phase 1

After sensing the information from its surrounding area, each sensor in the cooperation group shares this information with the nodes in its neighbor list in a TDMA manner. Every node in the neighbor list that receives this data, stores it in its local memory. The received signal $r_{i,j}$ at the relay node V_i , from the source node V_j , in phase 1 is,

$$r_{i,j} = \sqrt{E_{s1}} v_j \beta_{i,j} + n_{i,j} \quad (4)$$

where E_{s1} is the transmitted symbol power in phase 1, v_j is the BSPK-modulated symbol sent from node V_j , and $n_{i,j}$ is the additive white Gaussian noise at node i from node j , with variance, N_0 . The data packet sent by each node in this phase is shown in Fig. 4(a).

2) Phase 2

The size of the data packets sent by a sensor is usually small and sending each packet separately to the BS requires a large number of transmissions, which increases the energy consumption. Therefore, aggregation of data is used at the intermediate nodes to reduce the control packet overhead and the number of transmissions required to send the same amount of data to the BS. The aggregated data is forwarded to the BS by using the AF protocol, in which, the relay equalizes the channel fades between the source and the relay by amplifying the received signal by a factor that is inversely proportional to the received power.

Each node V_i , combines the received information from the nodes within its cooperation group, \mathcal{V} , to form a cooperative data packet. The cooperative data packet represented by x_i at

a node i , consists of a concatenation of the received and amplified packets from all the nodes in the cooperation group ($\zeta_{i,j}\hat{u}_{i,j}$, $j=1,2,\dots,N$ and $j \neq i$) and its own information v_i . $\hat{u}_{i,j}$ is the detected signal and $\zeta_{i,j}$ is the amplification factor at the relay node V_i with a corresponding source node V_j . The cooperative data packet is formed as,

$$x_i = \begin{cases} \zeta_{i,j}\hat{u}_{i,j} & j \neq i \\ v_i & j = i \end{cases}, \quad (5)$$

where

$$\zeta_{i,j} = \frac{1}{\sqrt{E_{s1}|\beta_{i,j}|^2 + N_{0,i,j}}}, \quad (6)$$

and $N_{0,i,j}$ is the input noise variance at the relay i from node j . The cooperative data packet sent by each node in this phase is shown in Fig. 4(c). In the cooperative packet, all the LCS and TOA information of the cooperation group including self-information is concatenated in sequential order. T_d contains the time spent at the relay node and Observation contains the observed alarm information by each node in the cooperation group, \mathcal{V} .

Upon its turn, every node transmits the cooperative data packet to the BS. The received signal at the BS, $y_{i,D}$, can be written as,

$$y_{i,D} = \sqrt{E_{s2}}x_i\alpha_{i,D} + n_{i,D} \quad (7)$$

where $\alpha_{i,D}$ is the lognormal fading channel coefficient from node V_i to the destination D and E_{s2} is the transmitted symbol power in phase 2. $n_{i,D}$ is the additive white Gaussian noise at destination D from node i , with power spectral density, N_0 . More specifically, the received signal at the destination D can be written as,

$$y_{i,D} = \frac{\sqrt{E_{s1}E_{s2}}}{\sqrt{E_{s1}|\beta_{i,j}|^2 + N_0}}\alpha_{i,D}\beta_{i,j}v_i + n'_{i,D}, \quad (8)$$

where

$$n'_{i,D} = \frac{\sqrt{E_{s2}}}{\sqrt{E_{s1}|\beta_{i,j}|^2 + N_0}}\alpha_{i,D}n_{i,j} + n_{i,D}. \quad (9)$$

Since the noise terms $n_{i,j}$ and $n_{i,D}$ can be assumed independent, then the equivalent noise $n'_{i,D}$ is a zero-mean complex Gaussian random variable with variance given as

$$N'_0 = \left(\frac{E_{s2}|\alpha_{i,D}|^2}{E_{s1}|\beta_{i,j}|^2 + N_0} + 1 \right) N_0. \quad (10)$$

IV. FUSION AT THE BASE STATION

The information from each cooperation group is received at the base station, decoded, and combined at the fusion center. Each packet contains its sensor ID and cooperation group ID as well as the observed information. Each node sends its own as well as the observation from all other sensors in its cooperation group to the BS in a combined packet. A *majority*

TABLE I. DATA FUSION AT THE BASE STATION

$s_i \backslash j$	s_1	s_2	s_3	s_4	s_5	s_6	s_7	$R(j)$
1	D	O	O	O	C	C	W	O
2	C	C	W	W	W	W	D	W
3	O	O	C	C	C	C	C	C
4	W	W	D	D	D	D	D	D
5	D	D	W	C	W	C	C	C
6	O	C	W	W	W	W	W	W
7	W	C	C	O	O	O	O	O

rule decision is made on the observations after collecting the received information from each sensor in the cooperation group. This helps increase the probability of correct decision at the BS even in bad channel conditions. This is illustrated in Table I, where j is the index of the cooperating node whose information is received from the sensor s_i . Here, O, C, W, and D represent OK, Caution, Warning, and Danger, respectively. A final result $R(j)$ is obtained based on majority rule as shown in Table I. A majority vote decision, which consists of votes from sensors in the cooperation group \mathcal{V} , can be mathematically represented as follows,

$$R(j) = \arg \max_X \sum_{i=1}^{\mathcal{N}} w_i I(s_i(j) = X) \quad (11)$$

where $s_i(j)$ is the j th cooperative symbol received from a sensor s_i with the information X . $I(\cdot)$ is an indicator function given as, $I(x) = \begin{cases} 1 & x \text{ is true} \\ 0 & x \text{ is false} \end{cases}$. For example, in the case of alarm information, $X = \{O, C, W, D\}$. Therefore, $I(x)$ will be true if the received information $s_i(j)$ is equal to one of O, C, W, or D, otherwise it will be false. w_i is the weight associated with each sensor's information. In this work, the channels are assumed to have equivalent average magnitude, therefore the weights are set to $1/\mathcal{N}$. Note that, if the weights w_i are set to $1/\mathcal{N}$, (11) results in the mode of $s_1, s_2, s_3, \dots, s_{\mathcal{N}}$.

V. PERFORMANCE ANALYSIS

Fig. 5 shows the proposed dual-hop multiple-branch communication system where each relay has multiple branch inputs and a single branch output, each working in an orthogonal manner based on TDMA. AF scheme is used at the relays in order to repeat the symbols for the neighbor nodes. The resulting symbol-error rate (SER) can be approximated as stated in the following theorem.

Theorem 1: If all of the channel links of the proposed multi-hop multi-branch cooperative system are known, the SER of a sensor node i at the destination D in the proposed system, can be tightly approximated as,

$$P_s(\gamma_{eq,i,D}) = F \left(1 + \frac{g_{PSK}}{N_0 \sin^2 \theta} \left(\frac{\sigma_{i,D}^2 \prod_{j=1}^{N-1} \sigma_{i,j}^2}{\prod_{j=1}^{N-1} \sigma_{i,j}^2 + \sigma_{i,D}^2 + 1} \right) \right) \quad (12)$$

where $F(x(\theta)) = \frac{1}{\pi} \int_0^{(M-1)\pi/M} \frac{1}{x(\theta)} d\theta$, M is the modulation symbol size, $g_{PSK} = \sin^2(\pi/M)$, $\gamma_{eq,i,D}$ represents the instantaneous SNR per relay node at the destination, and $\sigma_{i,j}^2$, $\sigma_{i,D}^2$ are the variances of the Rayleigh fading channel coefficients $\beta_{i,j}$ and $\alpha_{i,D}$, respectively.

Proof: See Appendix A.

Since, we use the majority voting rule in the fusion process, to decide a final outcome. Therefore, the probability of error, in the result after fusion, can be computed by using the Binomial theorem. Let, P_s , be the probability that the information sent by a sensor has error, and $l = \lceil \frac{N+1}{2} \rceil$ be the minimum number of votes needed for majority, then the probability of error in the consensus is given as,

$$P_e(\mathcal{N}) = \sum_{m=l}^{\mathcal{N}} \binom{\mathcal{N}}{m} P_s^m (1-P_s)^{\mathcal{N}-m}. \quad (13)$$

With (13), we expect to obtain a diversity order of l in the final SER of the proposed system.

A. Throughput and Energy Consumption of the Network

In this subsection, we aim to compare the non-cooperative and cooperative schemes in terms of throughput and energy consumption of the network. For the sake of a fair comparison, we assume a traditional dual-hop communication scheme for the non-cooperative mechanism, in which each node's data is forwarded by a relay node in the second hop towards the BS without any cooperative mechanism. Let B represent the number of bits per symbol, and the symbol duration is given by $T_s = 1/f_s$, where f_s is the symbol rate. Then, the throughput in case of non-cooperative (T_{nc}) and cooperative (T_c) dual-hop communication is given as,

$$\begin{aligned} T_{nc} &= \frac{\mathcal{N}B}{\mathcal{N}T_s + \mathcal{N}T_s} \text{ bps} \\ T_c &= \frac{\mathcal{N}B}{\mathcal{N}T_s + \mathcal{N}\mathcal{N}T_s} \text{ bps} \end{aligned}, \quad (14)$$

where the addition in denominator represents the time taken by two hops to transmit the symbol to BS. The additional \mathcal{N} in the denominator for T_c comes from the fact that each node relays the data of \mathcal{N} nodes in the second phase. The time taken by \mathcal{N} nodes to transmit \mathcal{N} packets to the BS in the case of non-cooperative (\mathcal{D}_{nc}) and cooperative (\mathcal{D}_c) scheme is then computed as,

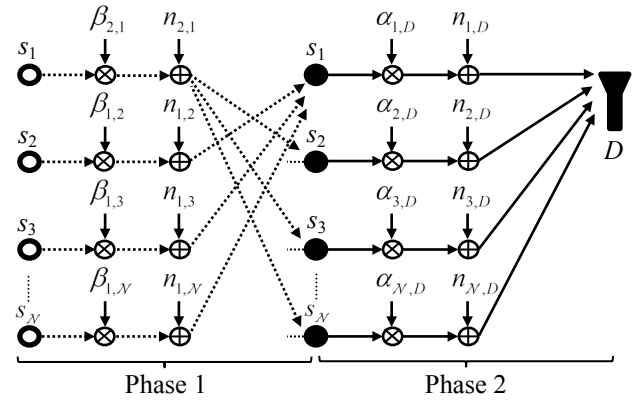


Fig. 5. The proposed two-phase communication system. In Phase 1, sensor s_1 in the cooperation group sends its information to all other sensors during its time slot. Similarly, all the other sensors send their information to s_1 during their allocated time slots. In Phase 2, the sensors then make a cooperative packet and send it to the destination, D .

$$\begin{aligned} \mathcal{D}_{nc} &= \frac{\mathcal{N} \times \text{size of data packet (bits)}}{T_{nc} \text{ (bps)}} \\ \mathcal{D}_c &= \frac{\mathcal{N} \times \text{size of data packet (bits)}}{T_c \text{ (bps)}} \end{aligned}. \quad (15)$$

Using $T_s = 15 \mu\text{s}$, the time delay given by (15) is plotted in Fig. 6(a).

In order to compute the energy consumption, let E_t , E_i , E_r , and E_f represent the energy consumed by the transmit operation by a sensor, idle listening, reception at a sensor node/BS, and fusion operation, respectively. In the case of non-cooperative dual-hop communication, each node transmits with energy E_t in phase 1 and the other $\mathcal{N}-1$ nodes receive this information with energy E_r . This process is repeated \mathcal{N} times. In phase 2, each node transmits with energy E_t to the BS while the other $\mathcal{N}-1$ nodes remain idle, and the BS receives each node's data with energy E_r . Thus the total energy consumed (E_{nc}) is given as,

$$\begin{aligned} E_{nc} &= \mathcal{N} (E_t + (\mathcal{N}-1)E_r) \\ &+ \mathcal{N} (E_t + (\mathcal{N}-1)E_i + E_r) \end{aligned}. \quad (16)$$

In the case of the proposed cooperative dual-hop communication, the total energy consumed (E_c) is given as,

$$\begin{aligned} E_c &= \mathcal{N} (E_t + (\mathcal{N}-1)E_r) \\ &+ \mathcal{N} (E_t + (\mathcal{N}-1)E_i + E_r) + \mathcal{N}^2 E_f \end{aligned}, \quad (17)$$

where E_f is the additional energy spent in fusion at the BS and \mathcal{N}^2 represent the number of multiply-and-accumulate operations performed to compute the fusion result for \mathcal{N} cooperative packets each containing \mathcal{N} number of observations given in (11). Using $E_t = 31.6 \text{ mW}$, $E_i = 2.8 \mu\text{W}$, $E_r = 17.4 \text{ mW}$ [20], and $E_f = 13.3 \text{ mW}$ [21], the results of (16) and (17) are plotted in Fig. 6(b).

Fig. 6(a) shows that the time required transmitting a certain amount of data to the BS increases in the case of our proposed cooperation scheme. But the given delay is still acceptable as it is 336 ms for $\mathcal{N} = 18$ and can go up to 957 ms for $\mathcal{N} = 30$. This amount of delay is not very critical and can be accepted

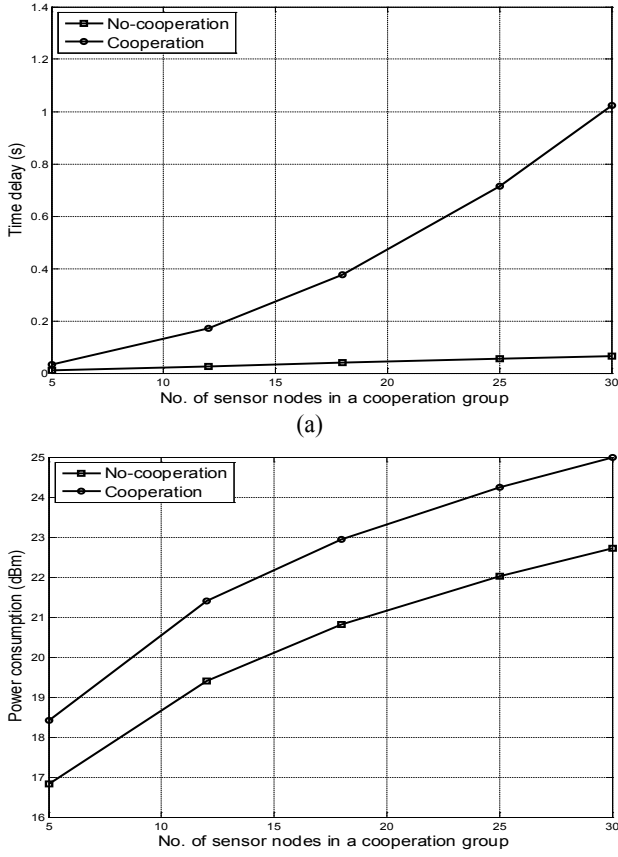


Fig. 6. Results of (15), (16), and (17). (a) Time delay for packet delivery. (b) Energy consumption of the network.

in return for improved robustness and reliability. Fig. 6(b) shows that the cooperation mechanism increases the amount of energy consumption by about 2 dB for $\mathcal{N}=12$ and remains below 3 dB for $\mathcal{N}=30$. This increase in the energy consumption is easily offset by the gain in SNR which is achieved by our proposed scheme, given in Section VI.

B. Comparative Analysis

The false alarm rate (FAR) and packet delivery rate (PDR) metrics are used to compare our results with some of the previous works mentioned in Section 1-A. The FAR and PDR for our work was calculated and averaged over a range of SNR (0 to 30 dB) and a total of 12,000 packets. In order to make a fair comparison, we use the PDR reported by [2] for IWSN and the PDR reported by [12], when no relay selection mechanism is used. As shown in Table II, our work shows a significant improvement in the FAR as compared to [14] and [15]. The PDR of our scheme is higher than that of [13] and is significantly higher than [2] and [12]. For improving the PDR, these works involve a significant overhead of retransmission, guide-path discovery, and relay selection mechanism, respectively. In contrast, our work does not involve guide-path discovery, relay selection, and retransmission overhead but still gives a higher PDR and very low FAR.

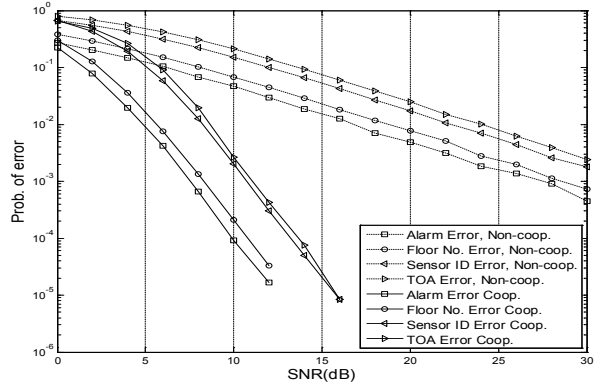


Fig. 7. Simulation results for 12-node cooperation group. (a) Simulation field of information. (b) Prob. of error for the received information at the BS.

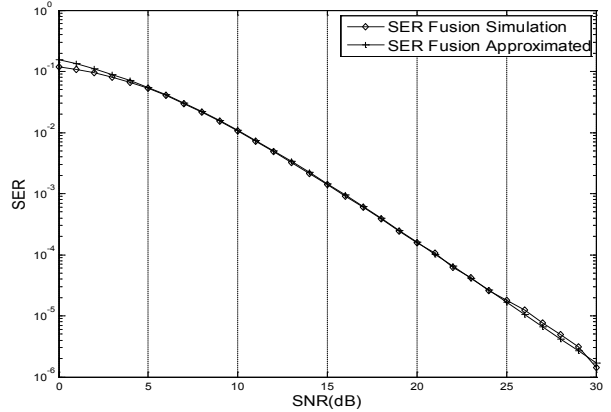


Fig. 8. Comparison of simulation result after fusion at the BS and the approximated SER given in (13).

TABLE II. COMPARISON WITH RELATED WORKS

Performance Metrics	[2]	[12]	[13]	[14]	[15]	Our Work
FAR	—	—	—	3.8%	10.5%	1.8%
PDR	~70%	~73%	~84%	—	—	~86%

VI. SIMULATION RESULTS

Assume an indoor communication environment of $100 \times 100 \text{ m}^2$ with hard-partitioned rooms. Some machines are scattered inside this area that generate some kind of radiation information i.e., temperature. Suppose that a higher temperature at a certain location represents a fault in the operation or state of the machine at that location. We model this information over the entire area as a Gaussian random field. The field varies from high temperature to low, which generates four different kinds of alarms i.e., Danger, Warning, Caution, and OK, respectively. The inter-sensor channels and the channels from sensor to BS are modeled as Rayleigh faded with lognormal shadowing for indoor environments with $\sigma = 7$ and $\eta = 3$. Each sensor has a sensing range of 18 m and 2.4 GHz ISM band carrier frequency is used. We assume the destination location at the edge of the area under consideration, and the nodes deployed according to the scheme discussed in Section II-A. The results are averaged over 20,000 sensing operations and compare our proposed scheme with that of a non-cooperative dual-hop communication.

Fig. 7 shows the probability of error for the alarms generated at the BS, floor number, sensor ID, and TOA for 12-node cooperation. We can see a clear advantage by using the proposed cooperation schemes, which achieves, on average, 10^{-3} probability of error at almost 20 dB lower SNR compared with the non-cooperative scheme. By taking into account the 2 dB increase in the cooperative transmission to the BS for $\mathcal{N}=12$, we can still get ~ 18 dB savings in the SNR as compared to the traditional dual-hop transmission without cooperation.

Fig. 8 shows the numerical result obtained in (13) for the SER of a cooperation group of 3 nodes, using majority vote fusion scheme at the BS compared with the simulated result. We can see that the approximated result matches that of simulation, especially at high SNR. The result also verifies that our proposed cooperation and fusion scheme is able to achieve the full diversity order of $l=2$ here.

VII. CONCLUSION

In this paper, we have proposed a relay based dual-hop cooperative WSN to monitor the state of an indoor industrial environment. By applying the proposed cooperation scheme, we obtain a much better performance in terms of SER and achieve a highly accurate decision at the base station. The packet overhead and energy consumption is reduced by combining a limited number of sensors' data into one packet for transmission. The energy saving provided by the proposed scheme is almost 18 dB, which is very significant for the harsh indoor industrial environment. The proposed cooperation protocol is robust to communication link failures and adapts to changing link conditions in the wireless channel. We also derived a closed-form solution for the SER of the proposed scheme, which verifies the diversity benefit of the scheme.

As a future work, this scheme can be extended to multi-hop and mobile sensor networks. Furthermore, the MAC design proposed in this paper can be further developed in future.

APPENDIX A PROOF OF THEOREM 1

In order to find the SNR at the destination D , we need to calculate the signal power and noise power components at the destination. The signal power for a single link is $(\beta_{1,j}^2 \zeta_{1,j}^2)(\alpha_{1,D}^2)$. Since, each node sends independent information, we take average to approximate the received signal power for each node at the destination. The signal power received from i th relay node is given as,

$$\begin{aligned} \text{SP}_i &= \left[\begin{aligned} &(\beta_{i,1}^2 \zeta_{i,1}^2) \times (\beta_{i,2}^2 \zeta_{i,2}^2) \times (\beta_{i,3}^2 \zeta_{i,3}^2) \times \dots \\ &\times (\beta_{i,\mathcal{N}-1}^2 \zeta_{i,\mathcal{N}-1}^2) \end{aligned} \right] (\alpha_{i,D}^2) \\ &= \alpha_{i,D}^2 \prod_{j=1}^{\mathcal{N}-1} \beta_{i,j}^2 \zeta_{i,j}^2 \end{aligned} \quad (18)$$

Similarly, the noise power for a single link is

$(N_{0,i,j} \zeta_{i,j}^2)(\alpha_{i,D}^2) + N_{0,i,D}$. The total noise power at the destination can be calculated as follows,

$$\begin{aligned} \text{NP}_i &= \left[\begin{aligned} &(N_{0,i,1} \zeta_{i,1}^2) \times (N_{0,i,2} \zeta_{i,2}^2) \times \\ &(N_{0,i,3} \zeta_{i,3}^2) \times \dots \times (N_{0,i,\mathcal{N}-1} \zeta_{i,\mathcal{N}-1}^2) \end{aligned} \right] (\alpha_{i,D}^2) + N_{0,i,D} \\ &= N_{0,i,D} + \alpha_{i,D}^2 \prod_{j=1}^{\mathcal{N}-1} N_{0,i,j} \zeta_{i,j}^2 \end{aligned} \quad (19)$$

The equivalent SNR at the destination, $\gamma_{eq,i,D}$ with respect to the relay node i can then be calculated by dividing the signal power with noise power as follows,

$$\gamma_{eq,i,D} = \frac{\alpha_{i,D}^2 \prod_{j=1}^{\mathcal{N}-1} \beta_{i,j}^2 \zeta_{i,j}^2}{N_{0,i,D} + \alpha_{i,D}^2 \prod_{j=1}^{\mathcal{N}-1} N_{0,i,j} \zeta_{i,j}^2} \quad (20)$$

Dividing the numerator and the denominator by $N_{0,i,D} \prod_{j=1}^{\mathcal{N}-1} N_{0,i,j} \zeta_{i,j}^2$, (20) is simplified as follows,

$$\text{Numerator} = \gamma_{i,D} \prod_{j=1}^{\mathcal{N}-1} \gamma_{i,j}, \quad (21)$$

$$\text{Denominator} = \frac{1}{\prod_{j=1}^{\mathcal{N}-1} N_{0,i,j} \zeta_{i,j}^2} + \frac{\alpha_{i,D}^2}{N_{0,i,D}} \quad (22)$$

Putting $\zeta_{i,j} = \frac{1}{\sqrt{E_{s1} |\beta_{i,j}|^2 + N_{0,i,j}}}$ in (22), we get the following,

$$\text{Denominator} = \prod_{j=1}^{\mathcal{N}-1} \gamma_{i,j} + \gamma_{i,D} + 1. \quad (23)$$

Therefore, the equivalent SNR at the destination D with respect to a sensor node i , is given as,

$$\gamma_{eq,i,D} = \frac{\gamma_{i,D} \prod_{j=1}^{\mathcal{N}-1} \gamma_{i,j}}{\prod_{j=1}^{\mathcal{N}-1} \gamma_{i,j} + \gamma_{i,D} + 1} \quad (24)$$

The SER formulation for the proposed system with M-PSK modulation, and conditioned upon known channel coefficients is given as,

$$\begin{aligned} P_s(\gamma_{eq,i,D}) &= \frac{1}{\pi} \int_0^{(M-1)\pi/M} e^{-\left(\frac{g_{PSK} \gamma_{eq,i,D}}{\sin^2 \theta}\right)} d\theta \\ &= \frac{1}{\pi} \int_0^{(M-1)\pi/M} e^{-\left(\frac{g_{PSK}}{\sin^2 \theta} \left(\frac{\gamma_{i,D} \prod_{j=1}^{\mathcal{N}-1} \gamma_{i,j}}{\prod_{j=1}^{\mathcal{N}-1} \gamma_{i,j} + \gamma_{i,D} + 1} \right)\right)} d\theta \end{aligned} \quad (25)$$

Since, each hop in the multi-hop multi-branch communication experiences independent fading and,

$$\int_0^\infty e^{\left(\frac{g_{PSK} E_s z}{N_0 \sin^2 \theta}\right)} p_{|h|^2}(z) dz = \frac{1}{1 + \frac{g_{PSK} E_s \sigma_h^2}{N_0 \sin^2 \theta}} \quad (26)$$

where h is the corresponding fading channel coefficient.

Therefore, we can write $P_s(\gamma_{eq,i,D})$ as

$$P_s(\gamma_{eq,i,D}) = F \left(1 + \frac{g_{PSK}}{N_0 \sin^2 \theta} \left(\frac{\sigma_{i,D}^2 \prod_{j=1}^{N-1} \sigma_{i,j}^2}{\prod_{j=1}^{N-1} \sigma_{i,j}^2 + \sigma_{i,D}^2 + 1} \right) \right) \quad (27)$$

where $F(x(\theta)) = \frac{1}{\pi} \int_0^{(M-1)\pi/M} \frac{1}{x(\theta)} d\theta$.

REFERENCES

- [1] L. D. Xu, W. He, and S. Li, "Internet of things in industries: a survey," *IEEE Trans. Ind. Inf.*, vol. 10, no. 4, pp. 2233-2243, Nov. 2014.
- [2] J. Niu, L. Cheng, Y. Gu, L. Shu, and S. K. Das, "R3E: Reliable reactive routing enhancing for wireless sensor networks," *IEEE Trans. Ind. Inf.*, vol. 10, no. 1, pp. 784-794, Feb. 2014.
- [3] L. Tang, K. C. Wang, Y. Huang, and F. Gu, "Channel characterization and link quality assessment of IEEE 802.15.4-compliant radio for factory environments," *IEEE Trans. Ind. Inf.*, vol. 3, no. 2, pp. 99-110, May 2007.
- [4] B. Silva, R. M. Fisher, A. Kumar, and G. P. Hancke, "Experimental link quality characterization of wireless sensor networks for underground monitoring," *IEEE Trans. Ind. Inf.*, vol. 11, no. 5, pp. 1099-1110, Oct. 2015.
- [5] Y. Ai, M. Cheffena, and Q. Li, "Radio frequency measurements and capacity analysis for industrial indoor environments," *Proc. 9th Eur. Conf. Ant. and Propag. (EuCAP)*, pp. 1-5, May 2015.
- [6] P. Murphy and A. Sabharwal, "Design, implemenation, and characterization of a cooperative communications system," *IEEE Trans. Veh. Tech.*, vol. 60, no. 6, pp. 2534-2544, Jul. 2011.
- [7] Y. M. Khattabi and M. M. Matalgah, "Performance analysis of multiple-relay AF cooperative systems over Rayleigh time-selective fading channels with imperfect channel estimation," *IEEE Trans. Veh. Tech.*, vol. 65, no. 1, pp. 427-434, Jan. 2016.
- [8] J. Rebelatto, B. Uchoa-Filho, Y. Li, and B. Vucetic, "Multiuser cooperative diversity through network coding based on classical coding theory," *IEEE Trans. Sig. Proc.*, vol. 60, no. 2, pp. 916-926, Feb. 2012.
- [9] W. L. Tan, W. C. Lau, and O. Yue, "Performance analysis of an adaptive, energy-efficient MAC protocol for wireless sensor networks," *Jr. Parallel. Distrib. Comput.*, vol. 72, no. 4, pp. 504-514, Apr. 2012.
- [10] H. S. AbdelSalam, and S. Olariu, "Toward adaptive sleep schedules for balancing energy consumption in wireless sensor networks," *IEEE Trans. Computers*, vol. 61, no. 10, pp. 1443-1458, Oct. 2012.
- [11] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-network aggregation techniques for wireless sensor networks: a survey," *IEEE Wireless Comm.*, vol. 14, no. 2, pp. 70-87, Apr. 2007.
- [12] N. Marchenko, T. Andre, G. Brandner, W. Masood, and C. Bettstetter, "An experimental study of selective cooperative relaying in industrial wireless sensor networks," *IEEE Trans. Ind. Inf.*, vol. 10, no. 3, pp. 1806-1816, Aug. 2014.
- [13] A. Ulusoy, O. Gurbuz, and A. Onat, "Wireless model-based predictive networked control system over cooperative wireless network," *IEEE Trans. Ind. Inf.*, vol. 7, no. 1, pp. 41-51, Feb. 2011.
- [14] J. Neuzil, O. Kreibich, and R. Smid, "A distributed fault detection system based on IWSN for machine condition monitoring," *IEEE Trans. Ind. Inf.*, vol. 10, no. 2, pp. 1118-1123, May 2014.
- [15] O. Kreibich, J. Neuzil, and R. Smid, "Quality-based multiple-sensor fusion in an industrial wireless sensor network for MCM," *IEEE Trans. Ind. Elect.*, vol. 61, no. 9, pp. 4903-4911, Sep. 2014.
- [16] R. Williams, *The Geometrical Foundation of Natural Structure: A Source Book of Design*. Dover Pub. Inc., New York, pp. 51-52, 1979.
- [17] F. Xue and P. R. Kumar, "The number of neighbors needed for connectivity of wireless networks," *Wireless Networks*, vol. 10, no. 2, pp. 169-181, Mar. 2004.
- [18] N. Xu, S. Rangwala, K. K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A wireless sensor network for structural monitoring," in *Proc. 2nd ACM Int. Conf. Emb. Net. Sen. Sys. (SenSys'04)*, Nov. 2004.
- [19] P. Huang, L. Xiao, S. Soltani, M. W. Mutka, and N. Xi, "The evolution of MAC protocols in wireless sensor networks: a survey," *IEEE Comm. Surv. Tuto.*, vol. 15, no. 1, pp. 101-120, 2013.
- [20] Silicon Labs, *EFR32 Flex Gecko Proprietary Wireless SoC*, www.silabs.com. Accessed, Jun. 2016.
- [21] K.-Y. Wu, C.-Y. Liang, K.-K. Yu, and S.-R. Kuang, "Multi-mode floating-point multiply-add fused unit for trading accuracy with power consumption," *IEEE/ACS Int. Conf. Comp. Inf. Sci.*, pp.429-435, Jun. 2013.



Zafar Iqbal received his undergraduate degree in computer engineering from COMSATS Institute of Information Technology, Islamabad, Pakistan in 2005 and M.S. in information and communications from the Gwangju Institute of Science and Technology (GIST), South Korea in 2010. He was with ZTE Corporation, Shanghai R&D Center, China from 2005 to 2008 and worked at Vieworks Co. Ltd. Korea, during 2011. Currently, he is pursuing a Ph.D. degree at the School of Electrical Engineering and Computer

Science in GIST. His research interests include wireless communication systems, digital signal processing, and design of VLSI circuits and systems. He was awarded the Korea IT Industry Promotion Agency scholarship for his M.S., and the Korean Government Scholarship for his Ph.D. study and research.



Kiseon Kim received the B.Eng and M.Eng degrees, all in electronics engineering, from the Seoul National University, Seoul, Korea, in 1978 and 1980, and the Ph.D degree from the University of Southern California, Los Angeles, in 1987, in Electrical Engineering – Systems.

Since joining Gwangju Institute of Science and Technology (GIST), Korea, in 1994, he is currently a Professor. His current interests include wideband digital communications system design, sensor network design for healthcare and smart grid, analysis and implementation both at the physical layer and at the resource management layer.



Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He then worked at HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member from 1999 to 2002. From 2002 to 2008, he worked as an Assistant Professor at the University of Pittsburgh, PA, USA. In 2009, he then moved to the School of Electrical Engineering and Computer Science, GIST, Korea, where he is currently affiliated. His areas of

research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He has received several prestigious national awards, including the Top 100 National Research and Development Award in 2012, the Top 50 Achievements of Fundamental Researches Award in 2013, and the Science/Engineer of the Month (January 2014).

Circular Sphere Decoding for Low Complexity Detection of MIMO Systems with General Two-dimensional Signal Constellations

Hwanchol Jang, Saeid Nooshabadi, *Senior Member, IEEE*, Kiseon Kim, *Senior Member, IEEE*, and
Heung-No Lee*, *Senior Member, IEEE*

Abstract—We propose a low complexity complex valued Sphere Decoding (CV-SD) algorithm, referred to as Circular Sphere Decoding (CSD) which is applicable to multiple-input multiple-output (MIMO) systems with arbitrary two dimensional (2D) constellations. CSD provides a new constraint test. This constraint test is carefully designed so that the element-wise dependency is removed in the metric computation for the test. As a result, the constraint test becomes simple to perform without restriction on its constellation structure. By additionally employing this simple test as a prescreening test, CSD reduces the complexity of the CV-SD search. We show that the complexity reduction is significant while its maximum-likelihood (ML) performance is not compromised. We also provide a powerful tool to estimate the pruning capacity of any particular search tree. Using this tool, we propose the Predict-And-Change strategy which leads to a further complexity reduction in CSD. Extension of the proposed methods to soft output SD is also presented.

Index Terms—multiple input multiple output (MIMO), circular sphere decoding (CSD), predict and change (PAC), sphere decoding (SD), complex-valued, arbitrary constellation

I. INTRODUCTION

SPHERE decoding (SD) is a promising multiple-input multiple-output (MIMO) detection strategy because it can achieve the error rate of maximum-likelihood (ML) detection with significantly less complexity compared to the straight-forward ML detector [1], [2]. The standard and most widely used SDs are real valued SDs (RV-SDs) which are only

directly applicable to real valued systems. For the usages of RV-SDs, a complex valued system is decomposed into its real and imaginary parts, and it becomes a real valued equivalent system with twice the dimension of its complex valued counterpart. It is worthwhile to note that these standard SDs assume that each data substream is modulated independently. This independency is maintained in the equivalent real valued system when each substream is modulated by a rectangular quadrature amplitude modulation (QAM). Pham *et al.* [3] and Mozos and Garcia [4] state that the application of RV-SD is permissible only for rectangular QAMs because otherwise invalid candidates may arise during the search; it is because the independency between the substream is broken.

It is desirable for an SD to be applicable directly to complex valued systems in the senses of *i*) its flexibility on the choice of signal constellations and *ii*) its efficiency in very large scale integration (VLSI) implementations. *i*): Complex valued SDs (CV-SDs), which are applicable directly to complex valued systems, do not require the decomposition of the systems, and therefore eliminate the limitation of its applicability to independent real and imaginary modulations like a rectangular QAM. Each data substream can be modulated using any two dimensional (2D) constellations. There are many constellations which are desirable to be employed in terms of many aspects of communications performance over rectangular QAMs. The benefits include the reduction in peak-to-average power ratio (PAPR) at each transmit antenna, signal-to-noise ratio (SNR) efficiency, and the increased range of choice in data rate. To list a couple of examples, star QAM reduces the PAPR [5], and near-Gaussian constellations yield a superior shaping gain [6], [7]. Rectangular QAMs depart significantly from these constellations. *ii*): The throughput of VLSI implementation of SD is inversely proportional to the product of the number of visited nodes and the time complexity for each node visit; we assume that one node is visited in each cycle. Burg *et al.* found that the expected number of nodes visited in the SD search is nearly doubled when a complex valued system is decomposed into its equivalent real valued system [8]. For the compensation to this increase, it requires the time complexity for each visit in RV-SD to decrease to half of its CV-SD equivalent. However, the time complexity of RV-SD is almost identical to that of CV-SD [8]. They conclude that CV-SD is the appropriate choice for high throughput VLSI implementations [8].

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2015R1A2A1A05001826). This paper was presented in part at the 45th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 6-9, Nov. 2011.

Hwanchol Jang was pursuing Ph.D. at Gwangju Institute of Science and Technology (GIST) while completing this work, and is currently with Advanced Photonics Research Institute, GIST, Gwangju, 500-712 Korea, e-mail: hcjang@gist.ac.kr. Saeid Nooshabadi is with the Department of Electrical and Computer Engineering and the Department of Computer Science, the Michigan Technological University, Houghton, MI, 49931, e-mail:saeid@mtu.edu. Kiseon Kim and Heung-No Lee are with the School of Information and Communications, GIST, e-mail: {kskim, heungno}@gist.ac.kr. The asterisk * indicates the corresponding author.

In CV-SDs, however, the main operation of SD, the pruning test is hard. The test relies on partial Euclidean distance (PED) computations. Here, PED computations are hard, actually the most expensive operations in SD, and thus, the CV-SD has high complexity. In previous CV-SDs, it is considered to restrict the applicable constellation to *i*) those whose elements are aligned in concentric rings of different radii [3], [4], [8], [9] and *ii*) those whose elements are aligned with several vertical or horizontal lines [10] for complexity reduction. Nevertheless, the general CV-SD without such restriction on its constellation is still remained to be computationally expensive.

It is our goal in this paper, therefore, to develop a low complexity CV-SD for general 2D constellations. We aim to do this while guaranteeing the ML performance. To this end, we take a new approach which takes advantage of a simple necessary condition, rather than an equivalent one, of the original pruning constraint, the sphere constraint (SC).

The main contributions of this paper are summarized as follows:

- We derive a necessary condition with which the metric for the constraint test becomes the magnitude of a scalar, not the Euclidean distance of a vector, thus the constraint test becomes very simple.
- We use the simpler metric and devise a novel complexity reduced CV-SD algorithm. The proposed constraint and the proposed CV-SD algorithm employing the constraint is referred to as the circular constraint (CC) and circular sphere decoding (CSD), respectively. CSD employs a two-step constraint test. In the prescreening step, those constellation points which are not promising are eliminated by the simple CC tests. The pruning by SC tests, in the second step, requiring expensive PED calculations is performed only for those candidates which have survived the CC tests. Thus, many expensive PED computations are avoided. Significant savings in the complexity of the CV-SD can be made with CSD without sacrificing the ML performance.
- We also propose a novel Predict-And-Change (PAC) strategy which further utilizes CC and reorganizes the tree so that tree pruning close to the root is increased. This leads to a substantial complexity reduction in CSD.
- We provide an extension of CSD for soft output SD for coded MIMO systems.

The rest of this paper is organized as follows. In Section II, the system model is presented. In Section III, the underlying principle in SD, and the difference between RV-SD and CV-SD are studied. In Section IV and V, the proposed CSD algorithm and the proposed PAC strategy are developed. In Section VI, an extension of CSD to list SD is provided. In Section VII, the complexity analysis for the proposed CSD is given. In Section VIII, we discuss the system simulation results. Section IX concludes the paper.

II. SYSTEM MODEL AND NOTATION

We consider a complex valued baseband MIMO channel

model with m receive and n transmit antennas ($m \geq n$). Consider the system model,

$$\mathbf{r} = \mathbf{H}\bar{\mathbf{s}} + \mathbf{v}, \quad (1)$$

where $\mathbf{r} \in \mathbb{C}^m$ denotes the received signal, $\mathbf{H} \in \mathbb{C}^{m \times n}$ denotes the $m \times n$ block Rayleigh fading channel matrix whose entries are independent and identically distributed (i.i.d.) circularly-symmetric complex Gaussian (CSCG) random variables $\mathcal{CN}(0,1)$, $\bar{\mathbf{s}} \in \mathcal{O}^n \subset \mathbb{C}^n$ is the transmitted symbol vector where $\mathcal{O} := \{o_1, o_2, \dots, o_L\}$ can be any discrete 2D constellation set with size L ; $o_i \in \mathbb{C}$ for $1 \leq i \leq L$. The components of $\bar{\mathbf{s}}$ are i.i.d. and take the values uniformly from \mathcal{O} , $\bar{s}_k \in \mathcal{O}$, and they are scaled to have $E|\bar{s}_k|^2 = 1$ for $\forall 1 \leq k \leq n$. The $m \times 1$ vector $\mathbf{v} \in \mathbb{C}^m$ is the additive noise whose entries are i.i.d. CSCG random variables $\mathcal{CN}(0, \sigma^2)$. The channel \mathbf{H} is assumed to be known at the receiver.

Variables that denote vectors and matrices are set, respectively, lowercase and uppercase boldface. $\mathbf{H}_{k,-}$ and $\mathbf{H}_{-,j}$ denotes the k^{th} row and the j^{th} column of matrix \mathbf{H} , respectively. \mathbf{H}^* and \mathbf{H}^\dagger denote the conjugate transpose and the pseudo-inverse of \mathbf{H} , respectively. An individual component of a matrix or a vector is identified by the subscript. For example, $H_{i,j}$ and s_k are the (i,j) component of \mathbf{H} and, the k^{th} component of \mathbf{s} , respectively. $\mathbf{s}_{k:n}$ is the vector taking the last $n-k+1$ components of \mathbf{s} . $|\mathcal{O}|$ denotes the cardinality of the set \mathcal{O} . $|a|$ denotes the magnitude of $a \in \mathbb{C}$. $\|\mathbf{s}\|$ denotes the 2nd norm of vector \mathbf{s} . \mathbb{R} and \mathbb{C} denote the real and the complex domains respectively. $\Re(s_k)$ and $\Im(s_k)$ denote the real and the imaginary parts of s_k respectively. $(a)_+$ is a if $a > 0$, and 0 otherwise.

III. COMPARISON BETWEEN RV-SD AND CV-SD

In this section, we describe the SD principle and the complexity problem in CV-SD.

A. SD principle

The standard procedure of SD is *i*) to identify all the candidates \mathbf{s} that satisfy the sphere constraint (SC), and *ii*) to choose the candidate with the minimum distance to the received signal \mathbf{r} as the solution. The SC is expressed by

$$d(\mathbf{s}) := \|\mathbf{r} - \mathbf{H}\mathbf{s}\|^2 \leq C, \quad (2)$$

where $\mathbf{s} \in \mathcal{O}^n$. But, this is not efficient for implementation. For a better implementation, the SC can be expressed as

$$\begin{aligned} d_k(\mathbf{s}_{k:n}) &:= \sum_{i=k}^n \left| y_i - \mathbf{R}_{i,i:n} \mathbf{s}_{i:n} \right|^2 \\ &= d_{k+1}(\mathbf{s}_{k+1:n}) + \left| b_{k+1}(\mathbf{s}_{k+1:n}) - R_{k,k} s_k \right|^2 \leq C, \end{aligned} \quad (3)$$

for $1 \leq k \leq n$ where \mathbf{R} is the upper triangular matrix from the

QR decomposition of \mathbf{H} , $\mathbf{H}=\mathbf{QR}$, $\mathbf{y}:=\mathbf{Q}^*\mathbf{r}$, $d_{n+1}=0$, and $b_{k+1}(\mathbf{s}_{k+1:n}):=y_k - \mathbf{R}_{k,k+1:n}\mathbf{s}_{k+1:n}$. Here, $d_k(\mathbf{s}_{k:n})$ is referred to as partial Euclidean distance (PED) and it depends on the partial vector $\mathbf{s}_{k:n}$, the last $n-k+1$ components of \mathbf{s} , which is associated with the nodes at the k^{th} level of the tree. As PED monotonically increases as k decreases, PEDs for the remaining levels of k do not need to be computed once PED of a level is found to violate the SC in (3). This gives computational savings to SD. As a result, SDs provide complexity reductions for ML solution search.

Definition 1: (element-wise independent metric) We call a metric is *element-wise independent* if the metric at a level of tree, say the k^{th} level, depends only on the corresponding element s_k in \mathbf{s} but not on any other s_i where $i \in \{1, 2, \dots, n\}$ and $i \neq k$. That is, among all the elements of \mathbf{s} , s_k is the only argument of the metric.

Here, we should note that the SC tests in (3) are still hard as the PEDs are element-wise dependent. Assume that the value of s_i for any $1 \leq i \leq n$ changes, the PED of the k^{th} level of the tree for level $k \leq i$ need to be calculated again. This incurs the number of required PED computations and the cost for each PED computation increased exponentially as k decreases. In the past, a simplified SC where explicit PED computations are not required was employed for a lower complexity algorithm. Further simplification on SC was obtained, as noted in the introduction, by employing rectangular QAMs and exploiting the characteristics that exists in the constellations, e. g., the independence between the real and the imaginary components.

B. The difference between RV-SD and CV-SD

In real valued systems where the components of \mathbf{r} , \mathbf{H} , $\bar{\mathbf{s}}$, \mathbf{s} , and \mathbf{v} are real valued, the SC in (3) can be simplified by the so called admissible interval (AI) which is expressed by the lower limit s_k^l and the upper limit s_k^u as follows

$$s_k \in [s_k^l, s_k^u], \quad (4)$$

where $s_k^l = \frac{b_{k+1}(\mathbf{s}_{k+1:n}) - \sqrt{C - d_{k+1}(\mathbf{s}_{k+1:n})}}{R_{k,k}}$ and $s_k^u = \frac{b_{k+1}(\mathbf{s}_{k+1:n}) + \sqrt{C - d_{k+1}(\mathbf{s}_{k+1:n})}}{R_{k,k}}$.

The AI is still element-wise dependent since $b_{k+1}(\mathbf{s}_{k+1:n})$ and $d_{k+1}(\mathbf{s}_{k+1:n})$, for the AI at the k^{th} level of the tree, are functions of $\mathbf{s}_{k+1:n}$, other than s_k . But it can be identified by calculating only the two values, s_k^l and s_k^u . The SC of (4) is used in RV-SD instead of the SC of (3) since it is simpler. The constellation points s_k which satisfy the SC can be identified without explicit PED computations $d_k(\mathbf{s}_{k:n})$ of candidates $\mathbf{s}_{k:n}$ for the current level k of the tree. PED computations are needed only for the nodes in the AI in RV-SD. They are not for pruning itself but for setting the AI, s_{k-1}^l and s_{k-1}^u , for the next level of the tree. However, note that, the AI simplification of the SC applies only to rectangular QAMs. References [1], [2] should be consulted for a more conceptual description of RV-SD.

In this work, the aim is for a CV-SD for general 2D

constellations. This makes it difficult to replace the expensive operations for the SC in (4) with ones that are much cheaper. Hence, every single pruning in the CV-SD is done by the explicit expensive PED computations for the SC test in (3) [8]. This results in high complexity in the CV-SD.

Let us see the difference on the number of PED computations of the CV-SD and the RV-SD, and the numbers of the floating point operations (FLOPs) of them. Consider a tree which is used for the SD search. The tree expands with the factor L as the level of the tree, k , goes down, starting from $k = n$ to $k = 1$. The number of nodes at the k^{th} level of the tree is L^{n-k+1} and each node represents a candidate for the partial vector $\mathbf{s}_{k:n}$. Let N_{k+1}^{sc} be the number of nodes that satisfy the SC at $(k+1)^{\text{th}}$ level of the tree. In the CV-SD, PEDs are calculated for all the children nodes of the surviving nodes. That is $L \cdot N_{k+1}^{sc}$ PED computations for the k^{th} level of the tree. Thus, the number of PED computations is expanded by L in CV-SD. We name this *L-expansion property* of the CV-SD. This property is trivial but it will be easier for us to recall this property later on. In the RV-SD, only the N_k^{sc} nodes which are inside the AI are required for PED computations at the k^{th} level of the tree no matter how large N_{k+1}^{sc} is. Surely, $L \cdot N_{k+1}^{sc} \geq N_k^{sc}$. The respective FLOPs of the PED computations for CV-SD and RV-SD are $(6(n-k)+8L)N_{k+1}^{sc}$ and $(2(n-k)+4)N_k^{sc}$.¹ Actually, the direct comparisons of the numbers of the PED computations and those of the FLOPs are not fair for the system models for RV-SD and for CV-SD are different. But, at least, we can see the inefficiency of the CV-SD for it includes $L \cdot N_{k+1}^{sc}$ factor rather than N_k^{sc} .

For low complexity CV-SDs, Hochwald and Brink [9], Burg *et al.* [8], Pham *et al.* [3], and Mozos and Garcia [4] consider restricting the applicable signal constellations only to those whose elements are aligned in several concentric rings with different sizes, rather than general 2D constellations, so that they can exploit the constellation structure. An AI is obtained for these complex valued constellations. There, the interval is not for the value of s_k itself but for its phase $\angle s_k$. Obviously the applicability of the method is limited to only those constellations with the specific shape. In addition, it requires costly trigonometric function and other computations. In this paper, we do not consider this approach.

C. Schnorr-Euchner enumeration in RV-SD and CV-SD

In Schnorr-Euchner (SE) enumeration, the children nodes of

¹ Computation of $b_{k+1}(\mathbf{s}_{k+1:n})$ requires $(n-k)$ complex multiplications and $(n-k)$ complex additions, totaling $6(n-k)$ FLOPs (one complex multiplication and one complex addition are equivalent to four FLOPs and two FLOPs, respectively). The remaining computation for the PED computation in (3) requires 8 FLOPs. The computation of $b_{k+1}(\mathbf{s}_{k+1:n})$ is required only once for it is common for the candidates as it does not depends on s_k . For RV-SD, computation of $b_{k+1}(\mathbf{s}_{k+1:n})$ requires $2(n-k)$ FLOPs and the remaining computation requires 4 FLOPs. Note that one more FLOP is required in case of a SC test for a comparison with C; consider this in Sec. IV-A.

a parent node are visited in the ascending order of their PEDs during the search. Therefore, once a preceding child node is found to violate the SC, it is definite that the remaining siblings also violate the SC and they do not need to be searched. This provides a considerable complexity reduction to SD.

In RV-SD, the sorting can be performed without explicit PED computations of the siblings. The first sibling s_k is determined by slicing $\frac{b_{k+1}(s_{k+1:n})}{R_{k,k}}$ to the closest constellation point. The sequence of the remaining siblings is then determined by the zigzag ordering of the neighboring constellation points of the first sibling [1]. Thus, the SE enumeration in RV-SD can be done very efficiently.

To the best knowledge of the authors, there is no efficient SE enumeration scheme for the CV-SD which is applicable to arbitrary 2D constellations and achieves the exact ML performance; we consider this in this paper. Thus, the SE enumeration for the CV-SDs requires explicit PED computations for all the siblings. Note that there are several efficient SE or SE-like enumeration schemes [3], [4], [8], [10], [11]. But, they are for the CV-SDs which *i)* limit their constellations to certain kinds and exploit the special structures of them, and/or *ii)* compromise the exact ML performance. *i)*: Pham *et al.* [3], Mozos and Garcia [4], and Burg *et al.* [8] apply SE enumerations on the constellations whose elements are aligned in several concentric rings with different sizes. The SE enumeration by Hess *et al.* [10] partitions the constellation into subsets consisting of rows or columns and reduces the enumeration overhead. But, this is efficient only for the constellations most of whose elements are aligned with several vertical or horizontal lines. *ii)*: The SE-like method by Wenk *et al.* [11] uses approximations, such as the l -1 norm or the so called l -infinity norm, of the PED for efficient enumeration. In this method, the remaining siblings are pruned once a preceded sibling node is found to violate the SC as they are in other SE enumerations. This may prune the node from which the ML solution originates; the method uses an approximation, not the PED itself.

IV. PROPOSED CIRCULAR SPHERE DECODING

The applicability to general 2D constellations is an important benefit of the general CV-SD. But, this generality gives an inherent problem in the CV-SD that no structure of a specific constellation can be exploited for a simplification of the SC. This makes it difficult to have a simpler but equivalent constraint to the SC in (3).

In this section, we simplify the SC of (3) by resorting to one of its necessary conditions, rather than to any specific structure in constellations. We refer to this necessary condition as the circular constraint (CC). The CC may not prune as many nodes as the SC does since the CC is a necessary condition. In order not to lose any pruning efficiency, we propose that SC tests are executed for those nodes which are not eliminated by the preceded CC tests. We call this CV-SD that employs CC tests circular sphere decoding (CSD). CSD prunes the nodes the same amount as the baseline CV-SD does but with a smaller

number of hard SC tests.

A. Circular constraint (CC)

SC tests in (3) are hard due to the element-wise dependence of the PEDs. Now, we aim to find a new constraint where the element-wise dependence of the metric is removed. We start from the SC in (3). The element-wise dependency is removed by eliminating the matrix \mathbf{R} inside the norm operator. The derivation is given by

$$\begin{aligned} C &\geq d_k(\mathbf{s}_{k:n}) \\ &= \|\mathbf{R}_{k:n,k:n}(\mathbf{x}_{k:n} - \mathbf{s}_{k:n})\|^2 \\ &= \frac{\|\mathbf{H}_{k,-}^\dagger\|^2}{\|\mathbf{H}_{k,-}^\dagger\|^2} \|\mathbf{R}_{k:n,k:n}(\mathbf{x}_{k:n} - \mathbf{s}_{k:n})\|^2 \\ &\stackrel{(a)}{\geq} \frac{|\mathbf{R}_{k,k:n}^\dagger \mathbf{R}_{k:n,k:n}(\mathbf{x}_{k:n} - \mathbf{s}_{k:n})|^2}{\|\mathbf{H}_{k,-}^\dagger\|^2} \\ &\stackrel{(b)}{=} \frac{|x_k - s_k|^2}{\|\mathbf{H}_{k,-}^\dagger\|^2}, \end{aligned} \quad (5)$$

where $\mathbf{x} := \mathbf{H}^\dagger \mathbf{r}$ with $\mathbf{H}^\dagger := (\mathbf{H}^* \mathbf{H})^{-1} \mathbf{H}^*$, (a) is from $\|\mathbf{H}_{k,-}^\dagger\| = \|\mathbf{R}_{k,k:n}^\dagger\|$ and the Cauchy-Schwarz inequality, and (b) is from the fact that $\mathbf{R}^\dagger \mathbf{R}(\mathbf{x} - \mathbf{s}) = \mathbf{x} - \mathbf{s}$ with the assumption that $\mathbf{H} \in \mathbb{C}^{m \times n}$ has the full rank with $m \geq n$. For a \mathbf{H} with rank deficiency, there are contributions in the metric derived in (5) from the elements of $\mathbf{x} - \mathbf{s}$ other than the k^{th} element of it. But, they are insignificant unless \mathbf{H} has serious rank deficiency.

Now, we have a new constraint, CC,

$$\Delta_k(s_k) \leq C \cdot \delta_k^2, \quad k = 1, 2, \dots, n, \quad (6)$$

where $\Delta_k(s_k) := |x_k - s_k|^2$ and $\delta_k^2 := \|\mathbf{H}_{k,-}^\dagger\|^2$. We name the metric $\Delta_k(s_k)$ as circular metric (C-metric). Here, the value of δ_k^2 can be computed before the SD search begins, and remains unchanged as long as \mathbf{H} does not change., thus it can be used while \mathbf{H} stays the same. Note that CC is a necessary condition for a candidate \mathbf{s} to satisfy the SC. It is because the metric derived in (5) is smaller than or equal to the PED for SC in (3).

The C-metric $\Delta_k(s_k)$ is element-wise independent since it depends only on s_k (Def. 1). This element-wise independence gives CC two beneficial features in terms of complexity. First, the required number of CC tests is fixed only to L for each level of the tree. It is because the C-metric does not depend on the elements of the parent node, and hence the C-metrics for the children nodes originated from one parent node are the same with those from any other parent nodes. Note that the number of required SC tests at the k^{th} level of the tree is $L \cdot N_{k+1}^{\text{sc}}$ (Sec. III-B); this ranges from L to L^{n-k+1} . Second, each CC test is simple. A CC test requires only six FLOPs while a SC test requires $6(n-k)+9$ FLOPs for the first sibling at the k^{th} level of the tree and 9 FLOPs for the remaining siblings.¹ Thanks to

these two features, the CC tests at the k^{th} level of the tree require only $6L$ FLOPs while the SC tests require $(6(n-k)+9L)N_{k+1}^{\text{sc}}$ FLOPs.

Note that there are some lower bounds used in other SDs [12], [13]. Stojnic *et al.* provide several lower bounds on the remaining path metric using ideas from H^∞ estimation theory and some of its special cases [12]. Barik and Vikalo obtain a lower bound on the metric by relaxing the metric minimization problem [13]. But, it is difficult to utilize them for prescreening on the SC tests because *i)* they are not lower bounds on the current PED but on the remaining path metric and *ii)* they are computationally expensive due to their element-wise dependent metric computations; they require $b_{k+1}(\mathbf{s}_{k+1:n})$.

B. Circular sphere decoding

In CSD, we utilize the simple CC for prescreening on the SC tests, thus reduce the computational complexity of the search. The strategy in CSD is *i)* to eliminate as many nodes as possible for a given level of the tree using the CC, and then *ii)* to perform SC tests only for the surviving nodes.

Consider Fig. 1 where we illustrate the CSD operations by providing a geometrical presentation of CC and SC. The CC is represented in the \mathbf{s} space by n separate circles, one for each element of \mathbf{s} . The SC is represented by the sphere in $\mathbf{H}\mathbf{s}$ space. In the prescreening step of CSD, the constellation points which are not inside each separate circle are excluded from the search. In the pruning step, SC tests are performed only for the prescreened candidates and vector points which are inside the sphere are identified. In CSD, many non-promising candidates are eliminated in the prescreening test even before their SC tests are performed. As it is shown in Fig. 1, only a portion of the points are pruned by the SC tests in CSD (Fig. 1 (b)), while whole pruning operations in the CV-SD are solely through the SC tests. Thus, the employment of the CC test in CSD reduces the complexity of the CV-SD. As a result, CSD outperforms the

CV-SD in terms of complexity (Section VIII). Note that the ML performance in CSD is not compromised since the CC test does not eliminate the ML solution; the CC is necessary for a candidate for the satisfaction of the SC.

The proposed CSD is a quite general MIMO detection framework that can be used together with other advantageous complexity reduction techniques or can be employed in other kinds of MIMO detection problems for further benefits. CSD can be employed with statistical pruning techniques such as that in [14] for additional complexity reduction. This can be done just by replacing C by one in [14]. Note also that CSD can be used to benefit those MIMO schemes employing space-time block codes in [15] and references therein. They have additional zero elements in the triangular matrix \mathbf{R} of their equivalent system models. The zero elements reduce the element-wise dependency in their PED computations. However, the PEDs are still not fully element-wise independent. CSD, with element-wise independent metric, can thus be applied to those additionally structured MIMO systems and provide further complexity reduction for ML detection.

The entire algorithm of the proposed CSD is given in Table I.

C. Circular enumeration

SE enumeration has a distinct benefit in reducing the complexity of the search. But, the SE enumeration for general CV-SDs is not suitable for CSD due to its heavy overhead. Fortunately, there are C-metrics available in CSD. C-metric of a node is a kind of a lower bound on the PED of the node (Eq. (5)). Thus, it can be used as a surrogate PED for the purpose of efficient enumeration.

We sort the constellation points in non-decreasing order with respect to their C-metrics as follows,

$$s_k^1, s_k^2, \dots, s_k^L \text{ s. t. } \Delta_k(s_k^1) \leq \Delta_k(s_k^2) \leq \dots \leq \Delta_k(s_k^L). \quad (7)$$

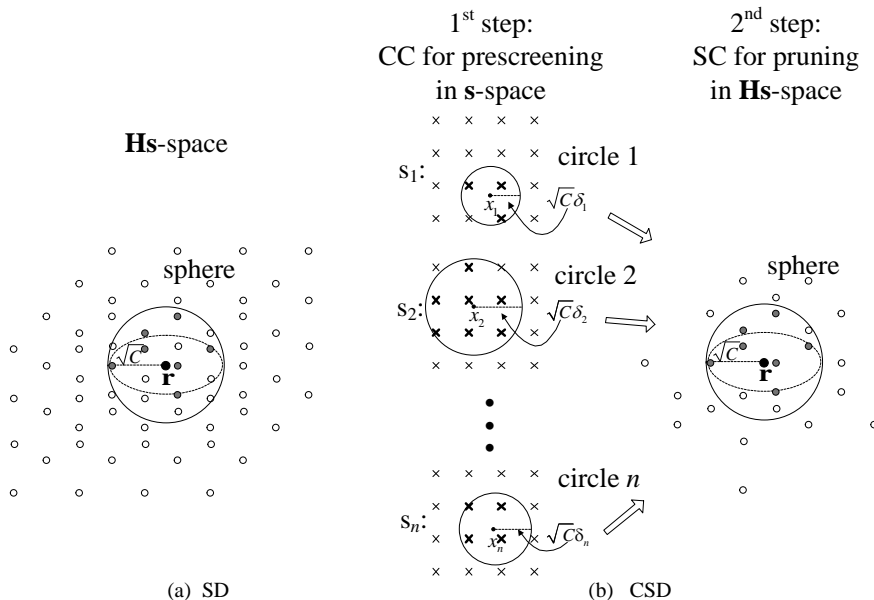


Fig. 1. Geometry of SD and the proposed CSD with 16 QAM constellation (gray-colored points represent those points which are inside the sphere). (a) SD. (b) CSD.

TABLE I

The CSD Algorithm	
Definitions:	$\mathcal{O} = \{o_1, o_2, \dots, o_L\}$ (the constellation set), $\delta^2 = [\delta_1^2, \delta_2^2, \dots, \delta_n^2]^T$, k (the current level), C (squared radius), $\mathbf{s} = [s_1, s_2, \dots, s_n]^T$ (the currently visited node), d_k (the PED for the current level), $\hat{\mathbf{s}}$ (ML solution).
Inputs:	\mathcal{O} , δ^2 , C_0 , \mathbf{r} , \mathbf{x} , and \mathbf{R} .
Outputs:	$\hat{\mathbf{s}}$.
Step 0: (C-metric)	Compute C-metrics $\Delta_k(s_k)$ for $\forall s_k \in \mathcal{O}$ and $\forall k$.
Step 1: (Initialization)	set $C \leftarrow C_0$, $k \leftarrow n$, $found \leftarrow 0$.
Step 2: (Initialization of the node for a visit)	for $k = 1:n$ set $I_k = 1$ and $s_k = o_1$.
end	Go to step 4
Step 3: (Next node)	if $I_k \neq L$, set $I_k = I_k + 1$ and $s_k = o_{I_k}$, and go to step 4 else go to step 8
Step 4: (CC test *)	if $\Delta_k(s_k) \leq C \cdot \delta_k^2$, go to step 5 else go to step 3
Step 5: (PED) Compute PED $d_k(s_{k:n})$.	
Step 6: (SC test)	if $d_k(s_{k:n}) < C$ go to step 7 else go to step 3
Step 7: (Forward)	if $k = 1$, $C \leftarrow d_k$ (radius updates), $\hat{\mathbf{s}} \leftarrow \mathbf{s}$, $found \leftarrow 1$, go to step 3. else $k \leftarrow k - 1$, go to step 4
Step 8: (Backward)	if $k = n$ (root node) if $found = 0$, $C_0 \leftarrow C_0 + \Delta C_0$ (radius increase), go to step 2 else exit else for $k' = 1:k$ set $I_{k'} = 1$ and $s_{k'} = o_1$. end $k \leftarrow k + 1$, go to step 3

* $C \cdot \delta_k^2$ is not computed for every CC test. It is computed only when the radius C is updated in Step 7.

This provides the benefit of making more nodes be pruned by the efficient CC tests. The CC becomes stricter whenever the search proceeds and reaches a leaf in the tree, where the radius is reduced to the metric for the leaf. Now that the siblings are sorted by the C-metric, stricter CC tests are performed for those siblings with larger C-metrics which have higher chances of being pruned by the CC. This results more candidates to be eliminated before the SC tests. It will be later shown in Sec. VIII that CSD with the circular enumeration (C-CSD) provides considerable complexity reduction to CSD; it also outperforms the SE-SD.

V. PROPOSED PREDICT-AND-CHANGE STRATEGY

Pruning of a single node at a level of a tree amounts to pruning of its whole underlying sub-tree whose size in the number of nodes is exponential to the number of levels left to reach to the leaf level. It is thus efficient in all SDs, including CSD, to prune nodes at higher levels of the tree. Aiming to increase pruning at higher levels of the tree, we take the approach which *i*) predicts the pruning potential of each symbol s_i of \mathbf{s} for $1 \leq i \leq n$, and *ii*) reorganizes the tree so that the symbols with larger pruning potentials to be placed at higher levels of the tree; the CSD search is performed on this tree. In the following subsection, we develop the idea of pruning potential for each symbol and provide a method to calculate it.

A. Symbol pruning potential

The pruning potential we propose to use here is an upper bound on the number of constellation points that can be pruned by the CC test. The largest number of constellation points that can be pruned by the CC test is calculated by using the strictest CC test which prunes as many constellation points as possible but without pruning the ML solution. Of course, this requires the identification of the ML solution or $d(\mathbf{s}_{\text{ML}})$ which are not available until the CSD search is completed. That is why we instead obtain an upper bound. This can be calculated by using the concept of the minimum circles. Note that so far there is no computationally feasible way to predict a non-trivial upper bound on the number of nodes that can be pruned by the SC test prior to a SD search.

Definition 2: (the minimum circles) The minimum circles (MCs) are the n smallest circles *i*) which are centered at x_1, x_2, \dots, x_n , respectively, *ii*) the proportion of whose radii is $\delta_1 : \delta_2 : \dots : \delta_n$, and *iii*) each of which contains at least a single constellation point inside it. The MCs are the geometric view of the strictest CC test which is satisfied by at least one of $\mathbf{s} \in \mathcal{O}^n$.

Let us denote the set of constellation points inside the i^{th} circle of the MCs by $\mathcal{O}_i^{\text{MC}}$. $\mathcal{O}_i^{\text{MC}}$ can be identified by using the following proposition.

Proposition 3: A constellation point $s_i \in \mathcal{O}$ belongs to $\mathcal{O}_i^{\text{MC}}$ if and only if

$$\Delta_i(s_i) \leq C_{\min} \cdot \delta_i^2, \quad (8)$$

where $C_{\min} := \max_i \left\{ \delta_i^{-2} \min_{s_i \in \mathcal{O}} \Delta_i(s_i) \right\}$.

Proof: The minimum radius for each circle of the MCs is $\delta_i^{-2} \min_{s_i \in \mathcal{O}} \Delta_i(s_i)$ for $1 \leq i \leq n$. The maximum radius is selected to guarantee that all the circles in MCs contain at least one constellation point. \square

The CC test corresponding to the MCs is stricter than or equal to the strictest CC test which the ML solution satisfies. Note that the ML solution is not guaranteed to be included in the vector points \mathbf{s} which are constituted by the constellation points inside the MCs. Thus, we can obtain the pruning potential P_i , an upper bound on the number of constellation points that can be pruned by the CC test, for each symbol at $1 \leq i \leq n$, as follows

$$P_i := L - |\mathcal{O}_i^{\text{MC}}| \quad (9)$$

$$= \left| \left\{ s_i \mid \Delta_i(s_i) > C_{\min} \delta_i^2, s_i \in \mathcal{O} \right\} \right|.$$

B. Predict-And-Change (PAC) strategy

The pruning potential P_i for each symbol s_i for $i = 1, 2, \dots, n$ in \mathbf{s} can be obtained using (9) once the C-metrics are computed for a received signal \mathbf{r} . Using the pruning potentials, the symbols in \mathbf{s} and the columns of \mathbf{H} are reordered. We propose that the n symbols in \mathbf{s} are placed from the root ($k = n$) to the leaf ($k = 1$) of the tree in non-increasing order with respect to their pruning potentials. The corresponding n columns in \mathbf{H} are also reordered accordingly. The reordered \mathbf{s}' and \mathbf{H}' are constituted as follows,

$$\mathbf{s}' = [s_{i_1}, s_{i_2}, \dots, s_{i_n}]^T \text{ and } \mathbf{H}' = [\mathbf{H}_{-i_1}, \mathbf{H}_{-i_2}, \dots, \mathbf{H}_{-i_n}] \quad (10)$$

$$\text{s. t. } P_{i_1} \leq P_{i_2} \leq \dots \leq P_{i_n}.$$

Then, the reorganized search tree is created by doing the QR decomposition on \mathbf{H}' . The CSD search is performed on the reorganized tree; the reorganization is done right before Step 1 in TABLE I.

Different from the SD ordering schemes which consider only \mathbf{H} [1], [2], PAC exploits the information of the received signal \mathbf{r} . Note that this becomes possible in PAC because it is based on the pruning potentials which are available prior to the formation of a tree. It is shown in Fig. 2 that the complexity reduction in CSD which employs the PAC strategy (PAC-CCSD) is substantial compared to that with the conventional ordering (PINV-CCSD); this conventional ordering places the symbol with a smaller inverse channel norm $\|\mathbf{H}_{i,-}^\dagger\|^2$ at a higher level of the tree.

C. A modified PAC

QR decomposition in SD needs to be performed whenever the channel \mathbf{H} changes. For PAC strategy, it requires n factorial QR decompositions per channel change since there exist n factorial different reordered channel matrices \mathbf{H}' . This may not be any problem in terms of computational overhead when the channel changes slowly (quasi-static channel), but otherwise, it may become problematic.

We here provide a variation of PAC and reduce the computational overhead of QR decompositions. This reduces the overhead per channel change significantly while the benefit of PAC per channel use is still kept large. For the variation, we consider a subset of \mathbf{H}' , only n different reordered channel matrices \mathbf{H}' out of the total n factorial of them. The reorganization is modified as follows,

$$\mathbf{s}' = [s_{i_1}, s_{i_2}, \dots, s_{i_n}]^T \text{ and } \mathbf{H}' = [\mathbf{H}_{-i_1}, \mathbf{H}_{-i_2}, \dots, \mathbf{H}_{-i_n}] \quad (11)$$

$$\text{s. t. } i_n = \arg \max_{i_k} \{P_{i_k}\} \text{ and } \delta_{i_1}^2 \geq \delta_{i_2}^2 \geq \dots \geq \delta_{i_{n-1}}^2.$$

The root of the tree, $k = n$, is placed by the symbol with the largest pruning potential. The other levels of the tree from $k = n-1$ to $k = 1$ are placed by the remaining symbols in non-decreasing order with respect to their δ_i^2 .

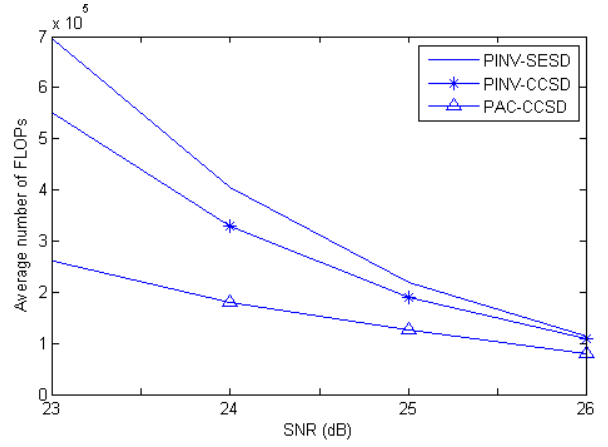


Fig. 2. Ordering benefits: complexity of PAC-CCSD, PINV-CCSD, and PINV-SESD for 10×10 MIMO systems with 64 QAM.

Although the pruning potentials are utilized only for the symbol to be placed at the root and δ_i^2 is utilized instead for the other levels of the tree, the benefit of PAC is reduced only slightly (Sec. VIII). The intuition for this is that *i*) we use the pruning potential at the level with the best pruning efficiency and *ii*) for other levels, we still consider pruning potentials but in the average sense since a smaller δ_i^2 indicates a larger P_i in the case where the information on $\Delta_i(s_i)$ and C_{\min} is not available (refer to (9)).

It is now possible with these n QR decompositions to give a comparable complexity reduction to that of the n factorial QR decompositions. In addition, we provide in the following subsection an efficient way of performing the n QR decompositions and the computational overhead is reduced to be less than that of two QR decompositions. With the modification and the efficient QR decompositions, we can say that the problem of PAC with the computational overhead is resolved.

For a clear description on the benefit of the modified PAC, we provide a comparison of complexities for the conventional PINV ordering based SE-SD (PINV-SESD), the PAC based C-CSD (PAC-CCSD), and the modified PAC based C-CSD (PAC*-CCSD) in TABLE II. They are for $n=m=10$ with (8,24,32) star 64 QAM. We split the complexities into those for channel rate processing and those for symbol rate processing. This is because the overhead is incurred in the channel rate preprocessing stage and the benefit is obtained in the symbol rate detection stage. This separately shows the gains and the losses of the PAC-CCSD and the PAC*-CCSD compared to PINV-SESD. We also provide the number of channel uses, N_{CH} , for PAC-CCSD and PAC*-CCSD to be net beneficial. It is

TABLE II

Ordering	Channel-rate preprocessing	Per-Symbol detection	N_{CH}
PINV	$C_{\text{QR}} \sim 5.33 \times 10^3$	$C_{\text{PINV-SESD}} \sim 4.33 \times 10^5$	
PAC	$n! C_{\text{QR}} \sim 1.94 \times 10^{10}$	$C_{\text{PAC-CCSD}} \sim 2.14 \times 10^5$	106619
PAC*	$2 C_{\text{QR}} \sim 1.07 \times 10^4$	$C_{\text{PAC*-CCSD}} \sim 2.51 \times 10^5$	1

C_{QR} refers to the FLOPs for the QR decomposition of a complex-valued matrix [16]. Symbol rate processing computational complexities are from the numerical simulation for $n=m=10$ with (8,24,32) star 64 QAM at the SNR of 24 dB (Fig. 4).

found that PAC-CCSD is the most advantageous in symbol rate detection. However, the net benefit to PINV-SESD is obtained in a very restricted condition where $N_{\text{CH}} > 106619$. It is seen that PAC*-CCSD provides a comparable symbol rate detection complexity while its overhead is significantly reduced. As a result, it is found that the proposed PAC*-CCSD always has a net gain to PINV-SESD.

D. Computational overhead for the modified PAC

We assume that Givens rotation, a well-known unitary transform based method, is used for QR decompositions. We do not consider the Gram Schmidt method here for it requires costly operations, such as square root operations and divisions, and it is not numerically stable; see [17] and references therein.

There are mainly two kinds of operations in Givens rotation, rotation and cancellation. Rotation makes a complex valued element turned into a real valued one. Cancellation makes a real valued element annihilated. For each columns of \mathbf{H}' , rotations and cancellations are performed. For the k^{th} column, $\mathbf{H}'_{:,k}$, the elements $H'_{i,k}$ for $k \leq i \leq n$ are rotated and become real valued. The complexity for rotations for the column is $(n-k+1) \cdot C_{\text{rot}}(n-k+1)$ where $C_{\text{rot}}(n-k+1)$ is the computational complexity for a rotation for the k^{th} column. Note that the single rotation for $H'_{i,k}$ amounts to the rotations for the $n-k+1$ elements, $H'_{i,j}$ for $k \leq j \leq n$, since the operations corresponding to the rotation for $H'_{i,k}$ are performed to those elements $H'_{i,j}$ where $k+1 \leq j \leq n$.

Once the rotations are done, the elements $H'_{i,k}$ for $k+1 \leq i \leq n$ which now have turned into real valued ones are annihilated and zero valued. The complexity for cancellations for the column is $(n-k) \cdot C_{\text{can}}(n-k+1)$ where $C_{\text{can}}(n-k+1)$ is the computational complexity for a cancellation for the k^{th} column. Note also that the single cancellation for $H'_{i,k}$ amounts to the cancellation for the $n-k+1$ elements for the same reason with the rotations. Now, the diagonal element $H'_{k,k}$ becomes real valued and the lower diagonal elements becomes zero. By performing the rotations and the cancellations from the 1st to the n^{th} column of \mathbf{H}' , a QR decomposition is done. The cost of this operation is $C_{\text{QR}} = \sum_{k=1}^n (n-k+1) \cdot C_{\text{rot}}(n-k+1) + (n-k) \cdot C_{\text{can}}(n-k+1)$.

There are n QR decompositions required for the proposed variation of PAC. This may be done by performing n separate QR decompositions with a cost of nC_{QR} . Fortunately, this can be done very efficiently (with less than $2C_{\text{QR}}$) by *i*) performing a QR decomposition on \mathbf{H}' whose columns are ordered by δ_i^2 and *ii*) deriving remaining $n-1$ QR decompositions from the one in step '*i*'. For the explanation of the method, we assume that $\delta_1^2 \geq \delta_2^2 \geq \dots \geq \delta_n^2$. We use \mathbf{H}'_n to denote the reorganized \mathbf{H} by (11). For step '*i*', in this case, $i_n = n$, a separate QR

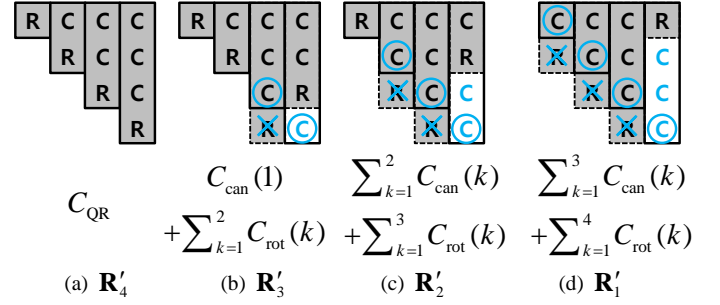


Fig. 3. Efficient QR decompositions for $n=4$. Symbols R and C in the matrix are meant to imply that the element in the pertinent position is either real or complex valued respectively. The circles and crosses represent the rotations and the cancellations, respectively. Gray colored rectangles are the columns taken from the upper triangular matrix in (a).

decomposition is performed. And for the remaining QR decompositions, the upper triangular matrix \mathbf{R}'_n obtained in step '*i*' is used instead of \mathbf{H}' . For the case of $i_n = i$, the i^{th} column of \mathbf{R}'_n is placed at the right-most position, and the columns from the $i+1^{\text{th}}$ to the n^{th} positions are shifted to the left by one. Then, the QR decomposition for $i_n = i$ can be done with only $C_i = \sum_{k=i}^n C_{\text{rot}}(n-k+1) + \sum_{k=i}^{n-1} C_{\text{can}}(n-k+1)$. This is done for $i_n = 1, 2, \dots, n-1$. Then, the net computational overhead for the n QR decompositions becomes $C_{\text{net}} = C_{\text{QR}} + \sum_{k=1}^n k \cdot C_{\text{rot}}(n-k+1) + (k-1) \cdot C_{\text{can}}(n-k+2) < 2C_{\text{QR}}$ ². An example for the QR decompositions is given in Fig. 3.

VI. EXTENSION TO LIST SPHERE DETECTION

In the iterative detection and decoding (IDD) system, the detector and the decoder exchange soft information repeatedly and improve the quality of their outputs. If messages are encoded with a channel code, and interleaved before they are mapped to modulation symbols \mathbf{s} , it is possible to achieve near channel capacity by employing IDD in the MIMO receiver [9], [18], [19].

In this section, we focus on the detection in IDD and consider an application of CSD for it. The channel code and the corresponding decoding operations can be any of those for well performing channel codes, such as low-density parity-check (LDPC) and turbo codes; we do not specify them in this paper. In the detection, a posteriori log-likelihood ratio (AP-LLR) on the all the coded bits that constitute \mathbf{s} are computed for soft information exchange [9]. Note that the coded bits here mean the coded and interleaved bits. The AP-LLR of the k^{th} coded bit, c_k , is given by

$$L(c_k | \mathbf{r}) = \ln \frac{P(c_k = 1 | \mathbf{r})}{P(c_k = 0 | \mathbf{r})} = \ln \frac{\sum_{\mathbf{s} \in S_k^1} P(\mathbf{r} | \mathbf{s}) P(\mathbf{s})}{\sum_{\mathbf{s} \in S_k^0} P(\mathbf{r} | \mathbf{s}) P(\mathbf{s})}$$

²The computation overhead for \mathbf{Q} matrix generation is not discussed here. However, it is still $C_{\text{net}} < 2C_{\text{QR}}$ for the computational overhead for \mathbf{Q} matrix generation is proportional to the number of rotations and cancellations.

$$\begin{aligned}
&= L(c_k) + \ln \frac{\sum_{\mathbf{s} \in \mathcal{S}_k^1} P(\mathbf{r} | \mathbf{s}) \prod_{i=1, i \neq k}^{n \log_2 L} P(c_i = c_i(\mathbf{s}))}{\sum_{\mathbf{s} \in \mathcal{S}_k^0} P(\mathbf{r} | \mathbf{s}) \prod_{i=1, i \neq k}^{n \log_2 L} P(c_i = c_i(\mathbf{s}))} \\
&= L(c_k) + \ln \frac{\sum_{\mathbf{s} \in \mathcal{S}_k^1} \varphi(\mathbf{s})}{\sum_{\mathbf{s} \in \mathcal{S}_k^0} \varphi(\mathbf{s})}, \quad (12)
\end{aligned}$$

where $\mathcal{S}_k^1 = \{\mathbf{s} | c_k(\mathbf{s}) = 1\}$, $\mathcal{S}_k^0 = \{\mathbf{s} | c_k(\mathbf{s}) = 0\}$, $c_i(\mathbf{s})$ is the i^{th} coded bit of symbol \mathbf{s} , $L(c_k) = \ln \frac{P(c_k=1)}{P(c_k=0)}$, and

$$\varphi(\mathbf{s}) = \exp\left(\frac{-1}{2\sigma^2} d(\mathbf{s})\right) \exp\left(\sum_{i=1, i \neq k}^{n \log_2 L} \ln P(c_i = c_i(\mathbf{s}))\right). \quad (13)$$

The direct AP-LLR in (12) without approximations requires the exhaustive Euclidean distance calculations of $d(\mathbf{s})$ for all the candidates \mathbf{s} , which are the most complex operations in the calculations. The computation is infeasible for the systems with large n and/or L .

SD algorithms which are called the list SD (LSD) [9], [18], [19] address this problem by searching a small set of candidate symbols over which the AP-LLRs are estimated. The approximate AP-LLR is given by

$$L(c_k | \mathbf{r}) \approx L(c_k) + \ln \left(\frac{\sum_{\mathbf{s} \in \mathcal{L} \cap \mathcal{S}_k^1} \varphi(\mathbf{s})}{\sum_{\mathbf{s} \in \mathcal{L} \cap \mathcal{S}_k^0} \varphi(\mathbf{s})} \right), \quad (14)$$

where \mathcal{L} is the chosen small set of the promising candidates. Hochwald and Brink proposed the N -best LSD which chooses candidates \mathbf{s} with N smallest $d(\mathbf{s})$ and stores them into \mathcal{L} [9]. A probabilistic tree pruning method for an approximation of the N -best LSD is proposed in [18]. Hochwald and Brink also presented a max-log approximation of AP-LLR for further size reduction on \mathcal{L} [9]. In [19], a modified max-log approximation which considers only the symbols \mathbf{s}_{ML} and all its binary complements are presented. Note that, in LSDs, small numbers of candidates \mathbf{s} with small $d(\mathbf{s})$ are sought for \mathcal{L} . The ground for this is that the significant contribution to the value of $\varphi(\mathbf{s})$ comes from \mathbf{s} with small $d(\mathbf{s})$, and the AP-LLR estimation over the list approximates the direct one. In this sense, given that $|\mathcal{L}| = N$, the N -best LSD is the optimal for AP-LLR approximations.

We here show that the use of CSD techniques are beneficial not only to ML search but also to the AP-LLR computation. The AP-LLR computation in (14) is largely made of evaluation of $\varphi(\mathbf{s})$ for $\mathbf{s} \in \mathcal{L}$. The main operations for $\varphi(\mathbf{s})$ in IDD detector is to calculate the $d(\mathbf{s})$ for $\mathbf{s} \in \mathcal{L}$; note that the values for $L(c_k)$ in (12) and $\ln P(c_i = c_i(\mathbf{s}))$ in (13) are given by the decoder. The $d(\mathbf{s})$ for $\mathbf{s} \in \mathcal{L}$ is obtained in the process of LSD, thus, the detection complexity in IDD is mostly from the LSD operations. Therefore, the computational complexity of LSD almost determines the complexity of the IDD detection.

We extend the CSD scheme to LSD and reduce the complexity of N -best LSD. We employ the CC for prescreening candidates in LSD. We modify the CSD procedure so that the radius update is performed only when the CSD finds at least N candidates inside the sphere. We found that the proposed list CSD (LCSD) reduces the complexity of the LSD for MIMO systems with general 2D signal constellations without

TABLE III

The N -Best List CSD Algorithm: Step 7'	
Step 7' : (Forward)	
if $k = 1$, Insert(\mathcal{L}, \mathbf{s}).	
if $ \mathcal{L} = N$, $C \leftarrow d_{\max}^{\mathcal{L}}$ (radius updates), $\hat{\mathcal{L}} \leftarrow \mathcal{L}$, $\text{found} \leftarrow 1$.	
go to step 3.	
else $k \leftarrow k - 1$, go to step 4	
* Insert(\mathcal{L}, \mathbf{s}): insert \mathbf{s} into the set \mathcal{L} ,	
$d_{\max}^{\mathcal{L}}$: the current maximum $d(\mathbf{s})$ among those of $\mathbf{s} \in \mathcal{L}$.	

compromising the N -best optimality. It will be shown in Section VIII that the considerable complexity reduction is obtained in N -best list search using proposed LCSDs. The modification to CSD algorithm for N -best list search is provided in TABLE III.

VII. COMPLEXITY ANALYSIS

In this section, we aim to analyze the complexity of the proposed CSD in Sec. IV-B. We use FLOPs as a measure of complexity. A lower bound on the expected FLOPs as function of n and SNR is analyzed. On the one hand, a lower bound analysis could be undesirable if the bound is not tight. But, on the other hand, the final expression of the lower bound could be simple enough to give a clear insight. To rip the benefit of the latter, we take the lower bound approach in this paper. We found that the lower bound can still give useful information on the additional complexity reduction behavior of CSD with respect to CV-SD. There are other lower bound analyses in the literature, those by Jaldén and Ottersten [20] and Shim and Kang [14]. Hassibi and Vikalo provide the exact expected complexity [21]. But, their final results are not easy to interpret for their complex expressions as they include integrations.

A. Overview of CSD complexity

Let the number of nodes which satisfy the SC at the k^{th} level of the tree be called N_k^{sc} , i.e.,

$$N_k^{\text{sc}} := \left| \left\{ \mathbf{s}_{k:n} \in \mathcal{O}^{n-k+1} \mid d_k(\mathbf{s}_{k:n}) \leq C \right\} \right|, \quad (15)$$

where $d_k(\mathbf{s}_{k:n}) = \sum_{i=k}^n \left| \mathbf{R}_{i,i:n} (\bar{\mathbf{s}}_{i:n} - \mathbf{s}_{i:n}) + \tilde{\mathbf{v}}_i \right|^2$ denotes the PED of $\mathbf{s}_{k:n}$ and $\tilde{\mathbf{v}} := \mathbf{Q}^* \mathbf{v}$ is a random vector whose entries are i.i.d. CSCG random variables, and the number of nodes which satisfy the CC at the k^{th} level of the tree be called N_k^{cc} ,

$$N_k^{\text{cc}} := \left| \left\{ s_k \in \mathcal{O} \mid \tilde{\Delta}_k(s_k) \leq C \right\} \right|, \quad (16)$$

where $\tilde{\Delta}_k(s_k) := \delta_k^{-2} |x_k - s_k|^2 = \delta_k^{-2} \left| \bar{s}_k - s_k + \mathbf{R}_{k,-}^\dagger \tilde{\mathbf{v}} \right|^2$ denotes the C-metric of s_k divided by δ_k^2 .

In CV-SD, the complexity is given by

$$\mathcal{C}_{\text{SD-C}} = \sum_{k=1}^n (6(n-k) + 9L) N_{k+1}^{\text{sc}}, \quad (17)$$

where $N_{n+1}^{\text{sc}} = 1$. Recall the L -expansion property (Sec. III-B). That is, for the identification of N_k^{sc} nodes at the k^{th} level of the

tree, $L \cdot N_{k+1}^{sc}$ PED computations are required; this results in high complexity in the CV-SD; note the L in $9L$.

In CSD, the high complexity problem of the CV-SD is alleviated. Among all the possible children nodes of each surviving node at the previous level, $k+1$, note that there are L children nodes, only those children nodes, $N_k^{cc} \leq L$ of them, that satisfy CC are passed for PED computations. The complexity for CSD is given by

$$\mathcal{C}_{\text{CSD-C}} = \sum_{k=1}^n 6L + (6(n-k) + 9N_k^{cc}) N_{k+1}^{sc}. \quad (18)$$

Here, $6L$ is the FLOPs for the CC tests at each level.

We analyze the complexity of CSD in the following subsection.

B. CSD complexity

The expected complexity of CSD over \mathbf{H} , $\bar{\mathbf{s}}$, and \mathbf{v} is

$$\mathbb{E}[\mathcal{C}_{\text{CSD-C}}] = \sum_{k=1}^n 6L + 6(n-k) \cdot \mathbb{E}[N_{k+1}^{sc}] + 9 \cdot \mathbb{E}[N_k^{cc} N_{k+1}^{sc}]. \quad (19)$$

For simplicity of the derivation, we assume $m=n$ in the sequel. But, it can be easily seen that the result for the general case $m>n$ remains the same. We also use the SNR $\rho = \frac{m}{\sigma^2}$ and the radius $C = \alpha n \sigma^2$ where α is determined so that a solution is found with a high probability, $1 - \varepsilon = 0.99$, in the sphere [21]. Since the complexity is averaged over $\bar{\mathbf{s}}$, we consider $\tilde{\Delta}_k$ and d_k as functions of not only \mathbf{s} but also $\bar{\mathbf{s}}$.

Lemma 4: $\mathbb{E}[N_k^{cc} N_{k+1}^{sc}]$ is lower bounded by

$$\mathbb{E}[N_k^{cc} N_{k+1}^{sc}] \geq \frac{1}{L^{n-k+1}} \sum_{\mathbf{s}_{k:n}} \sum_{\bar{\mathbf{s}}_{k:n}} \Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C) \Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C). \quad (20)$$

Proof: See Appendix A. \square

We compute $\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C)$ first, and then $\Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C)$.

Lemma 5: The probability $\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C)$ is lower bounded by

$$\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C) \geq \left(1 - \frac{1}{\alpha n} \left(\left| \bar{s}_k - s_k \right|^2 \rho + 1 \right) \right)_+. \quad (21)$$

Proof: See Appendix B. \square

Now, we compute a lower bound on $\Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C)$.

Lemma 6: The probability $\Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C)$ is lower bounded by

$$\Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C) \geq \left(1 - \frac{n-k}{\alpha n} \left(\left\| \bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n} \right\|^2 \frac{\rho}{n} + 1 \right) \right)_+. \quad (22)$$

Proof: See Appendix C. \square

Finally, we obtain a lower bound on the expected complexity

of CSD.

Theorem 7: The expected complexity $\mathbb{E}[\mathcal{C}_{\text{CSD-C}}]$ is lower bounded by

$$\mathbb{E}[\mathcal{C}_{\text{CSD-C}}] \geq \sum_{k=1}^n 6L + 6(n-k)(1 - \beta_s)_+ + 9L^{n-k+1}(1 - \beta_s)_+(1 - \beta_c)_+, \quad (23)$$

where

$$\beta_c := \frac{1}{\alpha n} \left(\mathbb{E} \left[\left| \bar{s}_k - s_k \right|^2 \right] \rho + 1 \right) \quad (24)$$

and $\beta_s := \frac{n-k}{\alpha n} \left(\mathbb{E} \left[\left| \bar{s}_k - s_k \right|^2 \right] \frac{n-k}{n} \rho + 1 \right)$ are the complexity reduction factors which are determined by the system parameters. Here, $\mathbb{E} \left[\left| \bar{s}_i - s_i \right|^2 \right]$, the average intra-constellation squared distance, is determined when the constellation is decided; for example, it is 2 for QAMs and PSKs, and 1.8163 for (8,24,32) 64 star QAM.

Proof: See Appendix D. \square

A lower bound on the expected complexity of the CV-SD can be easily obtained by using a similar procedure, but with (17) and Lemma 6.

Theorem 8: The expected complexity $\mathbb{E}[\mathcal{C}_{\text{SD-C}}]$ is lower bounded by

$$\mathbb{E}[\mathcal{C}_{\text{SD-C}}] \geq \sum_{k=1}^n 6(n-k)(1 - \beta_s)_+ + 9L^{n-k+1}(1 - \beta_s)_+. \quad (25)$$

Here, we see that the additional constraint CC in CSD results in an additional multiplicative factor $(1 - \beta_c)_+$, in the last term of the CSD complexity in (23), compared to the CV-SD complexity in (25). Among the last term in (25), β_c portion of them are excluded in (23); $(\cdot)_+$ and $6L$ are ignored for easy evaluation. We expect that this gives CSD a complexity reduction relative to the CV-SD.

We note from (24) that the additional complexity reduction factor β_c increases as *i*) SNR increases and/or *ii*) n decreases. That is, it is expected that the additional complexity reduction in CSD to the CV-SD become more as SNR increases and/or n decreases. This finding exactly matches to the complexity reduction behavior of CSD which is observed in the simulations. Here, the complexity reduction behavior with respect to n may not be attractive. But, this problem disappears when PAC is utilized; it performs better as n increases (Sec. VIII).

VIII. SIMULATION RESULTS

A. Setup

In this section, we show the complexity reduction capability of the proposed CSDs through system simulations. We compare the proposed CSDs (CSD in Sec. IV-B, C-CSD in Sec. IV-C, and PAC-CCSD in Sec. V) with the conventional CV-SDs (SD, SE-SD, and PINV-SESD) which are also directly applicable to general complex valued constellations. We also compare the

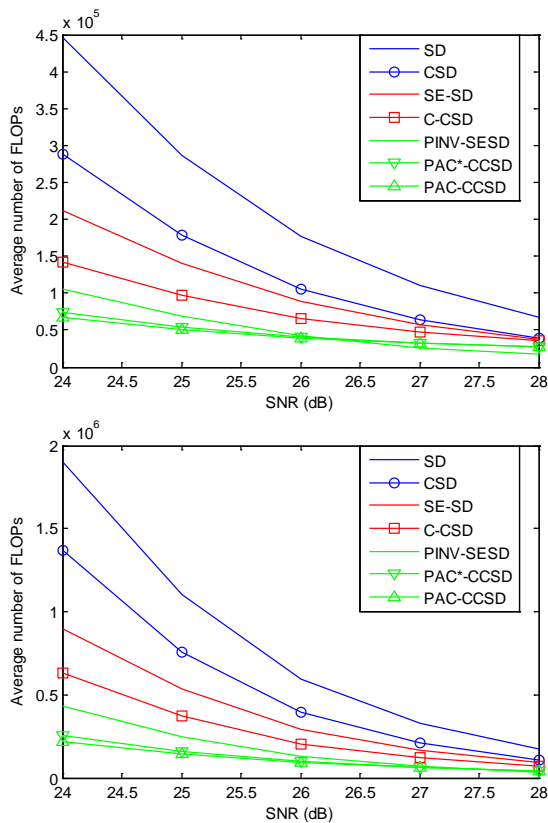


Fig. 4. Complexities of CV-SDs and the proposed CSDs for 8×8 and 10×10 MIMO systems with $(8,24,32)$ star 64 QAM.

complexities of the proposed LCSs in Sec. VI to those of the conventional LSDs for N -best list search. We considered star QAM, rectangular QAM, and PSK in simulations. However, we present the result for star QAM only. The patterns of the complexity reduction of the proposed CSDs in rectangular QAM and PSK are almost the same to those in star QAM. Note that the usage of CSD is not limited to these three constellations, and is applicable to any arbitrary complex valued constellation. We use 8×8 and 10×10 MIMO systems with $(4, 24, 32)$ star 64 QAM for the ring ratios of $(2, 3)$ [5].³ The initial radius is set to $C_0 = \alpha n \sigma^2$. We employ the average FLOPs as a metric for complexity; they are averaged over 10^4 runs of channels in each SNR value. The SNR range is determined by considering the dynamic range of the FLOPs so that the FLOPs at the minimum considered SNR and those at the maximum SNR are reasonably far from each other. If they are too apart, the FLOPs at the maximum SNR all look overlapped and not distinguishable.

B. Results

Fig. 4 plots the average numbers of FLOPs for the CV-SDs and the proposed CSDs in MIMO systems with $(8,24,32)$ star 64 QAM. The proposed CSDs reduce the complexities of the CV-SDs considerably. We observe that CSD, C-CSD, and PAC-CCSD outperform SD, SE-SD, and PINV-SESD by 35%, 37%, 41%, 42%, 43%, by 33%, 30%, 27%, 17%, 3%, and by

³ The ring ratios indicate the ratios of the minimum ring amplitude to the other ring amplitude.

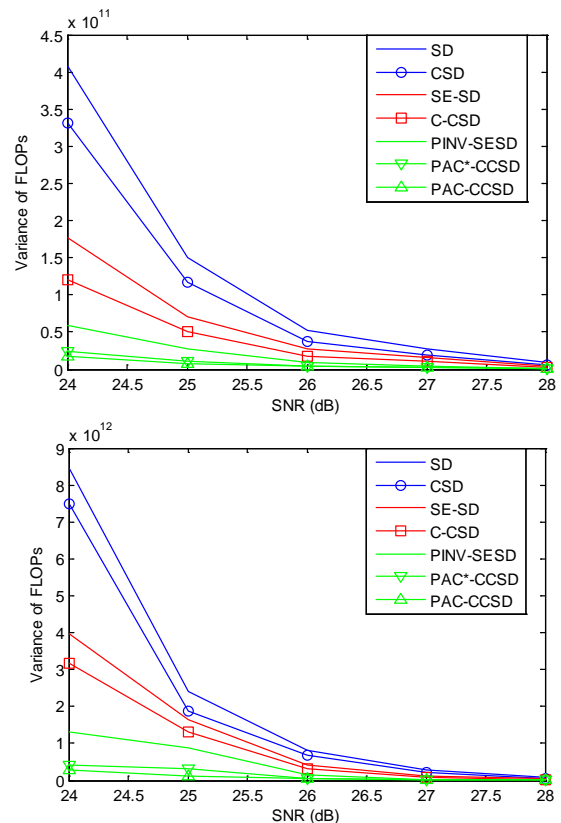


Fig. 5. Complexity variances of CV-SDs and the proposed CSDs for 8×8 and 10×10 MIMO systems with $(8,24,32)$ star 64 QAM.

37%, 26%, 6%, -18%, -37%,⁴ respectively, at SNR (dB) of 24, 25, 26, 27, 28 in 8×8 MIMO systems. For 10×10 MIMO systems, CSD, C-CSD, and PAC-CCSD reduce the complexities of SD, SE-SD, and PINV-SESD by 28%, 31%, 33%, 36%, 40%, by 30%, 31%, 29%, 27%, 21%, and by 50%, 43%, 30%, 13%, -14%, respectively.

As discussed in Section VII-B, the complexity reduction factor of CSD increases as SNR increases and/or n decreases. Interestingly, the complexity reduction factor of PAC-CCSD behaves in an opposite way to that of the CSD; it increases as SNR decreases and n increases. There are significant complexity reductions in the proposed CSDs in almost all the SNR regions (up to 43% by CSD at 28 dB and up to 37% by PAC-CCSD at 24 dB in 8×8 MIMO systems and up to 40% by CSD at 28 dB and up to 50% by PAC-CCSD at 24 dB in 10×10 MIMO systems). Note that these complexity reductions in CSDs are obtained without compromising its ML performance; of course, they all achieved the ML performance in simulations. It was found that PAC-CCSD has a higher complexity than PINV-SESD does at SNRs of 27 dB and 28 dB in a 8×8 MIMO system and at SNR of 28 dB in a 10×10 MIMO system. Still, it is not a big issue since both exhibit low complexities in this SNR

⁴ The minus sign means that the complexity of the proposed method is more than that of the conventional method. The corresponding numbers are the increased amounts of complexity compared to those of the conventional method.

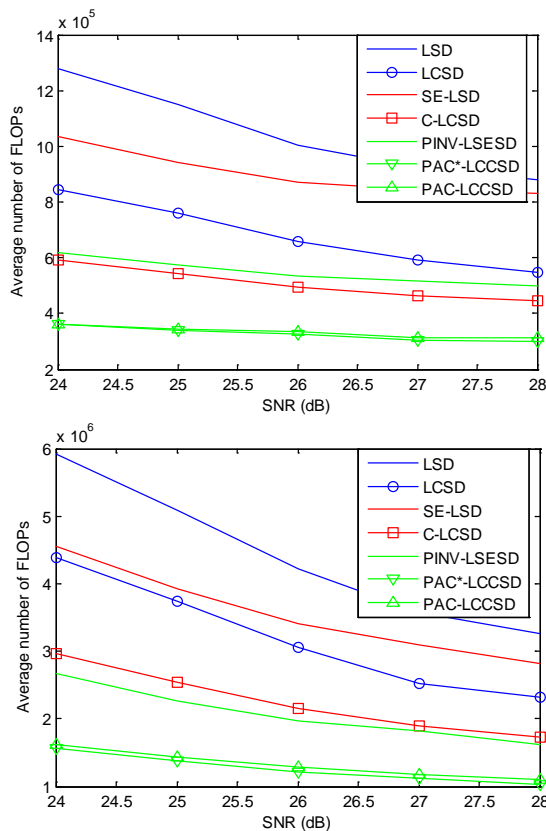


Fig. 6. Complexities of CV-LSDs and the proposed LCSDs for 8×8 and 10×10 MIMO systems with (8,24,32) star 64 QAM.

region.

The proposed CSDs also reduce the complexity variances of the CV-SDs (Fig. 5). Note that in Fig. 5 variance reductions in the proposed CSDs are significant throughout the entire SNR range considered. For example, the PAC-CCSD scheme provides reductions up to 71% in a 8×8 MIMO system and up to 87% in a 10×10 MIMO system.

We also consider N -best list detection. The patterns of the complexity reduction of the proposed LCSDs are similar to those of CSDs but with more margins. It is observed that the proposed LCSDs perform better than LSDs by large margins. For 64 QAM, LCSD, C-LCSD, and PAC-LCCSD outperform LSD, SE-LSD, and PINV-LSESD in terms of average number of FLOPs by 34%, 34%, 34%, 36%, 37%, by 43%, 43%, 43%, 45%, 47%, and by 41%, 40%, 37%, 39%, 37%, respectively, in a 8×8 system with $N=20$, by 26%, 27%, 27%, 29%, 29%, by 35%, 35%, 37%, 39%, 39%, and by 39%, 37%, 35%, 36%, 32%, respectively, in 10×10 system with $N=30$ at SNR (dB) of 24, 25, 26, 27, 28 (Fig. 6). In terms of complexity variance, it is shown in Fig. 7 that the variance reductions by LCSDs are significant (up to 57% in a 8×8 MIMO system and up to 51% in a 10×10 MIMO system) as they are in CSDs.

IX. CONCLUSION

In this paper, we have proposed CSD, a low complexity CV-SD, for general 2D constellations. CSD uses the simple

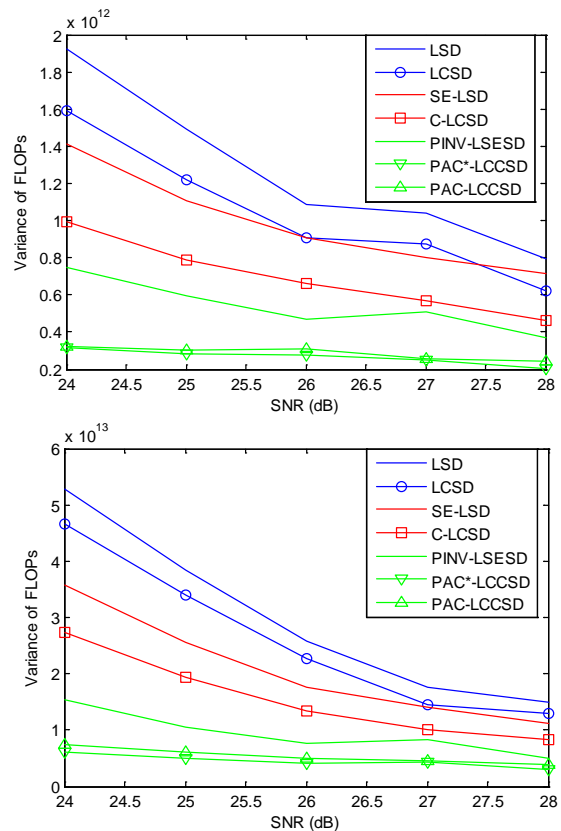


Fig. 7. Complexity variances of CV-LSDs and the proposed LCSDs for 8×8 and 10×10 MIMO systems with (8,24,32) star 64 QAM.

circular constraint (CC) for prescreening candidates and pruning some of them even before executing the SC tests. Simulations have shown that CSD yields a large reduction in the number of FLOPs. We also propose a further complexity reduction strategy, the Predict-And-Change (PAC). PAC also provides a further considerable complexity reduction. Thanks to the proposed methods, the CSD becomes surely a good candidate for a general 2D constellation low complexity MIMO detector. It was also shown that the proposed methods are beneficial in soft output SD schemes. With the proposed CSD, it becomes possible to decode signals with any integer data rate (not only for $L = 2^i$, $i = 1, 2, \dots$) and arbitrary shape of 2D constellations which may be optimal for target applications with low complexity while achieving the optimal error rate performance as CSD is compatible to general 2D constellation. A large number of 2D constellations can be handled in this single CSD algorithm without any additional constellation-dependent functionality.

APPENDIX

A. Proof of Lemma 4

$$\begin{aligned}
 & \mathbb{E}[N_k^{cc} N_{k+1}^{sc}] \\
 &= \mathbb{E} \left[\sum_{\tilde{s}_k} I\{\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C\} \sum_{\tilde{s}_{k+1:n}} I\{d_{k+1}(s_{k+1:n}, \bar{s}_{k+1:n}) \leq C\} \right] \quad (26)
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{s_{k:n}} \Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C, d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C) \\
&= \frac{1}{L^{n-k+1}} \sum_{s_{k:n}} \sum_{\bar{s}_{k:n}} \Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C, d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C),
\end{aligned}$$

where $I\{\cdot\}$ is the indicator function which is 1 for the condition inside the bracket is true, otherwise 0, and (a) is from $\Pr(\bar{\mathbf{s}}_{k:n}) = 1/L^{n-k+1}$.

Before proceeding further, we introduce a definition and a lemma which are useful for further simplification of (26).

Definition 9: (non-negative correlation) Two random events \mathcal{A} and \mathcal{B} are said to be *non-negatively correlated* if $\text{cov}(I\{\mathcal{A}\}, I\{\mathcal{B}\}) \geq 0$. We use the tilde symbol \sim between two events, i.e., $\mathcal{A} \sim \mathcal{B}$, to imply that the two are *non-negatively correlated*.

Lemma 10: $\mathcal{A} \sim \mathcal{B}$ if and only if $\Pr(\mathcal{A}\mathcal{B}) \geq \Pr(\mathcal{A})\Pr(\mathcal{B})$.

Proof: This can be easily seen by Def. 9. \square

Here we aim to show that $I\{d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C\} \sim I\{\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C\}$. To this end, we first show that $I\{d_{k+1} \leq C\} \sim I\{d_1 \leq C\}$, and then $I\{d_1 \leq C\} \sim I\{\tilde{\Delta}_k \leq C\}$, and thus $I\{d_{k+1} \leq C\} \sim I\{\tilde{\Delta}_k \leq C\}$. Before we move on, note the inequalities $d_{k+1} \leq d_1$ and $\tilde{\Delta}_k \leq d_1$ from (3) and (5). Now, we show that $I\{d_{k+1} \leq C\} \sim I\{d_1 \leq C\}$ by showing the following, $\Pr(d_{k+1} \leq C, d_1 \leq C) = \Pr(d_1 \leq C) \geq \Pr(d_{k+1} \leq C)\Pr(d_1 \leq C)$. Thus, the first is shown. Now, for the second, we show that $\Pr(d_1 \leq C, \tilde{\Delta}_k \leq C) = \Pr(d_1 \leq C) \geq \Pr(d_1 \leq C)\Pr(\tilde{\Delta}_k \leq C)$, thus $I\{d_1 \leq C\} \sim I\{\tilde{\Delta}_k \leq C\}$. Therefore, $I\{d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C\} \sim I\{\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C\}$; this is also verified through extensive simulations.

Now, return to the discussion of (26). Let $\mathcal{A} := \{\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C\}$ and $\mathcal{B} := \{d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C\}$. Using Lemma 10, the joint probability $\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C, d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C)$ is lower bounded by $\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C)\Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C)$.

B. Proof of Lemma 5

$E[\tilde{\Delta}_k(s_k, \bar{s}_k)]$ is expressed as follows

$$\begin{aligned}
&E\left[\frac{|\bar{s}_k - s_k + \mathbf{R}_{k,-}^{-1} \tilde{\mathbf{v}}|^2}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] \\
&= \frac{|\bar{s}_k - s_k|^2}{E[\|\mathbf{R}_{k,-}^{-1}\|^2]} + E\left[\frac{\left|\sum_{i=1}^n R_{k,j}^{-1} \cdot \tilde{v}_i\right|^2}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] + 2(\bar{s}_k - s_k)^* E\left[\frac{\sum_{i=1}^n R_{k,j}^{-1} \cdot \tilde{v}_i}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right].
\end{aligned}$$

The first term is upper bounded as follows,

$$|\bar{s}_k - s_k|^2 / E[\|\mathbf{R}_{k,-}^{-1}\|^2] \stackrel{(a)}{\leq} |\bar{s}_k - s_k|^2 E[\|\mathbf{R}_{k,-}^{-1}\|^2] \stackrel{(b)}{=} |\bar{s}_k - s_k|^2 \cdot n,$$

where (a) is from $E[\|\mathbf{R}_{k,-}^{-1}\|^2] E[\|\mathbf{R}_{k,-}^{-1}\|^2] \geq E[\|\mathbf{R}_{k,-}^{-1} \mathbf{R}_{k,-}^{-1}\|] = 1$ and (b) is from $\|\mathbf{R}_{k,-}^{-1}\|^2 = \|\mathbf{Q}^* \mathbf{H}_{k,-}\|^2 = \|\mathbf{H}_{k,-}\|^2$.

The second term becomes σ^2 as follows,

$$\begin{aligned}
&E\left[\frac{\left|\sum_{i=1}^n R_{k,j}^{-1} \cdot \tilde{v}_i\right|^2}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] \\
&\stackrel{(c)}{=} \sum_{i=1}^n E\left[\frac{|R_{k,i}^{-1}|^2}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] E[|\tilde{v}_i|^2] + \sum_{i=1, j=1, i \neq j}^n E\left[\frac{R_{k,i}^{-1*} R_{k,j}^{-1}}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] E[\tilde{v}_i^*] E[\tilde{v}_j] \\
&= \sigma^2,
\end{aligned}$$

where (c) is from the independence of $\tilde{\mathbf{v}}$ and \mathbf{R}^{-1} .

The third term becomes zero as follows

$$E\left[\frac{\sum_{i=1}^n R_{k,j}^{-1} \cdot \tilde{v}_i}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] \stackrel{(d)}{=} \sum_{i=1}^n E\left[\frac{R_{k,i}^{-1}}{\|\mathbf{R}_{k,-}^{-1}\|^2}\right] E[\tilde{v}_i] = 0,$$

where (d) also comes from the independence of $\tilde{\mathbf{v}}$ and \mathbf{R}^{-1} .

Therefore, $E[\tilde{\Delta}_k(s_k, \bar{s}_k)] \leq \sigma^2 \left(|\bar{s}_k - s_k|^2 \rho + 1\right)$.

Now, $\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C)$ is lower bounded by as follows,

$$\Pr(\tilde{\Delta}_k(s_k, \bar{s}_k) \leq C) \stackrel{(a)}{\geq} \left(1 - \frac{1}{\alpha n} \left(|\bar{s}_k - s_k|^2 \rho + 1\right)\right)_+,$$

where (a) is from the Markov inequality and $(\cdot)_+$ is to make sure the probability to be nonnegative.

C. Proof of Lemma 6

$E[d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n})]$ is expressed as follows

$$\begin{aligned}
&E[d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n})] \\
&= \sum_{i=k+1}^n E\left[\left|\sum_{j=i}^n R_{i,j}(\bar{s}_j - s_j) + \tilde{v}_i\right|^2\right] \\
&\stackrel{(a)}{=} \sum_{i=k+1}^n \sum_{j=i}^n |\bar{s}_j - s_j|^2 \left(E[\|\mathbf{R}_{-j}\|^2] - \sum_{i=1}^k E[|R_{i,j}|^2]\right) + (n-k)\sigma^2 \\
&\stackrel{(b)}{=} (n-k) \|\bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n}\|^2 + (n-k)\sigma^2,
\end{aligned}$$

where (a) comes from the independence between $\tilde{\mathbf{v}}$ and \mathbf{R} , and the fact that any off-diagonal element of \mathbf{R} has mean zero [21] and (b) comes from $\|\mathbf{R}_{-j}\|^2 = \|\mathbf{H}_{-j}\|^2$.

Finally, using the Markov inequality,

$$\Pr(d_{k+1}(\mathbf{s}_{k+1:n}, \bar{\mathbf{s}}_{k+1:n}) \leq C) \geq \left(1 - \frac{n-k}{\alpha n} \left(\|\bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n}\|^2 \frac{\rho}{n} + 1\right)\right)_+.$$

D. Proof of Theorem 7

$$\begin{aligned}
E[N_k^{cc} N_{k+1}^{sc}] &\stackrel{(a)}{\geq} (1/L^{n-k+1}) \sum_{s_k} \sum_{\bar{s}_k} \left(1 - \frac{1}{\alpha n} \left(\|\bar{s}_k - s_k\|^2 \rho + 1 \right) \right)_+ \\
&\quad \cdot \sum_{\mathbf{s}_{k+1:n}} \sum_{\bar{\mathbf{s}}_{k+1:n}} \left(1 - \frac{n-k}{\alpha n} \left(\|\bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n}\|^2 \frac{\rho}{n} + 1 \right) \right)_+ \\
&\stackrel{(b)}{=} L^{n-k+1} \cdot \left(1 - \frac{1}{\alpha n} \left(E \left[\|\bar{s}_k - s_k\|^2 \right] \rho + 1 \right) \right)_+ \\
&\quad \cdot \left(1 - \frac{n-k}{\alpha n} \left(E \left[\|\bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n}\|^2 \right] \frac{n-k}{n} \rho + 1 \right) \right)_+,
\end{aligned}$$

where (a) is from Lemma 4, 5 and 6, (b) is from

$$\left(\sum_k a_k \right)_+ \leq \sum_k (a_k)_+, \quad \sum_{\bar{\mathbf{s}}_{k+1:n}} \sum_{\mathbf{s}_{k+1:n}} 1 = L^{2(n-k)}, \quad \text{and}$$

$$E \left[\|\bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n}\|^2 \right] = \frac{1}{L^{2(n-k)}} \sum_{\bar{\mathbf{s}}_{k+1:n}} \sum_{\mathbf{s}_{k+1:n}} \|\bar{\mathbf{s}}_{k+1:n} - \mathbf{s}_{k+1:n}\|^2.$$

$E[N_{k+1}^{sc}]$ can be easily derived using a similar procedure.

REFERENCES

- [1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. on Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.
- [2] M. O. Damen, H. El Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Trans. on Inf. Theory*, vol. 49, no. 10, pp. 2389–2402, Oct. 2003.
- [3] D. Pham, K. R. Pattipati, P. K. Willett, and J. Luo, "An Improved Complex Sphere Decoder for V-BLAST Systems," *IEEE Signal Processing Letters*, vol. 11, no. 9, pp. 748–751, Sep. 2004.
- [4] R. S. Mozos, and M. J. F.-G. Garcia, "Efficient Complex Sphere Decoding for MC-CDMA Systems," *IEEE Trans. on Wireless Communications*, vol. 5, no. 11, pp. 2992–2996, Nov. 2006.
- [5] R. Kobayashi, T. Kawamura, N. Miki, and M. Sawahashi, "Throughput Comparisons of Star 32/64 QAM Schemes Based on Mutual Information Considering Cubic Metric," *International Conference on Wireless Communications and Signal Processing (WCSP)*, Nanjing, China, Nov. 2011.
- [6] G. D. Forney and G. Ungerboeck, "Modulation and Coding for Linear Gaussian Channels," *IEEE Trans. on Inf. Theory*, vol. 44, no. 6, pp. 2284–2415, Oct. 1998.
- [7] I. Nevat, G. W. Peters, and J. Yuan, "Detection of Gaussian Constellations in MIMO Systems Under Imperfect CSI," *IEEE Trans. on Communications*, vol. 58, no. 4, pp. 1151–1160, Apr. 2010.
- [8] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bölcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 7, pp. 1566–1577, July 2005.
- [9] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. on Communications*, vol. 51, no. 3, pp. 389–399, Mar. 2003.
- [10] C. Hess, M. Wenk, A. Burg, P. Luethi, C. Studer, N. Felber, and W. Fichtner, "Reduced-Complexity MIMO Detector with Close-to ML Error Rate Performance," *Proc. 17th ACM Great Lakes symposium on VLSI*, Italy, Mar. 2007.
- [11] M. Wenk, L. Bruderer, A. Burg, C. Studer, "Area-and Throughput-Optimized VLSI Architecture of Sphere Decoding," *18th IEEE/IFIP VLSI System on Chip Conference*, Spain, Sep. 2010.
- [12] M. Stojnic, H. Vikalo, and B. Hassibi, "Speeding up the Sphere Decoder With H^∞ and SDP inspired Lower Bounds," *IEEE Trans. on Signal Processing*, vol. 56, no. 2, pp. 712–726, Feb. 2008.
- [13] S. Barik and H. Vikalo, "Sparsity-Aware Sphere Decoding: Algorithms

and Complexity Analysis," *IEEE Trans. on Signal Processing*, vol. 62, no. 9, pp. 2212–2225, May. 2014.

- [14] B. Shim and I. Kang, "Sphere decoding with a probabilistic tree pruning," *IEEE Trans. on Signal Processing*, vol. 56, no. 10, pp. 4867–4878, Oct. 2008.
- [15] T. P. Ren, Y. L. Guan, C. Yuen, and E. Y. Zhang, "Block-Orthogonal Space-Time Code Structure and Its Impact on QRDM Decoding Complexity Reduction", *IEEE J. Sel. Topics in Signal Processing*, vol. 5, no. 8, pp.1438–1450, 2011.
- [16] M. Arakawa, "Computational Workloads for Commonly Used Signal Processing Kernels," Office of the Secretary of Defense under Air Force Contract FA8721-05-C-0002, Project Report SPR-9, 2006.
- [17] L. Ma, K. Dickson, J. McAllister, and J. McCanny, "QR Decomposition-Based Matrix Inversion for High Performance Embedded MIMO Receivers," *IEEE Trans. on Signal Processing*, vol. 59, no. 4, pp. 1858–1867, Apr. 2011.
- [18] J. Lee, B. Shim, and I. Kang, "Soft-Input Soft-Output List Sphere Detection with a Probabilistic Radius Tightening," *IEEE Trans. on Wireless Communications*, vol. 11, no. 8, pp. 2848–2857, Aug. 2012.
- [19] C. Studer, A. Burg and H. Bölcskei, "Soft-input soft-output single tree-search sphere decoding", *IEEE Trans. Inf. Theory*, vol. 56, pp. 4827–4842, 2010.
- [20] J. Jaldén and B. Ottersten, "On the complexity of Sphere Decoding in Digital Communications," *IEEE Trans. on Signal Processing*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.
- [21] B. Hassibi and H. Vikalo, "On the Sphere-Decoding Algorithm I. Expected Complexity," *IEEE Trans. on Signal Processing*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.



Hwanchol Jang received the B.S. degree from Handong Global University (HGU), Pohang, Korea, in 2006, the M.S. and Ph.D. degrees from Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, in 2008 and 2015, respectively, all in electrical engineering. He is currently a Postdoctoral fellow at Advanced Photonics Research Institute (APRI), GIST. His current research interests include multidimensional and statistical signal processing and their applications to multiple input multiple output communications, compressed sensing, and various imaging modalities such as photoacoustic imaging and optical imaging and focusing inside/through turbid media.



Saeid Nooshabadi (M'01-SM'07) received the M.Tech and PhD degrees in electrical engineering from the India Institute of Technology, Delhi, India, in 1986 and 1992, respectively. Currently, he is the professor of Computer Systems Engineering with Department of Electrical & Computer Engineering and the Department of Computer Science, Michigan Technological University, Michigan. Prior to his current appointment he has held multiple academic and research positions. His last two appointments were with the Gwangju Institute of Science and Technology, Republic of Korea (2007 to 2010), and with the University of New South Wales, Sydney, Australia (2000 to 2007). His research interests include VLSI information processing and low-power embedded processors.

Kiseon Kim: Information available in IEEE Trans on Vehicular Technology, Vol. 65, No 1, pp. 333-347, 2016.



Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He then moved to HRL Laboratories, LLC, Malibu, CA, USA, and worked there as a Research Staff Member from 1999 to 2002. In 2002, he was appointed an Assistant Professor with the University of Pittsburgh, Pittsburgh, PA, USA, where he stayed until 2008. In 2009, he then moved to the School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Korea, where he is currently affiliated. His general areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He has served as a member of technical program committees for several IEEE conferences, including the IEEE International Conference on Communications and the IEEE Global Communications Conference (Globecom). He served as the Lead Guest Editor for the European Association for Signal Processing Journal on Wireless Communications and Networking in 2010–2011. He has served as an Area Editor for the AEU International Journal of Electronics and Communications since January 2013. His research efforts have been recognized with prestigious national awards, including the Top 100 National Research and Development Award from the Korean Ministry of Science, ICT, and Future Planning in 2012, the Top 50 Achievements of Fundamental Researches Award from the National Research Foundation of Korea in 2013, and the Science/Engineer of the Month (January 2014) from the National Research Foundation of Korea. In March 2015, he was appointed as the Dean of Research at GIST.

QUERY FORM

SREP	
Manuscript ID	[Art. Id: srep32355]

Journal: SREP

Author:- The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections at the appropriate positions in the text.

Query No.	Description	Author's Response
	<p>Author surnames have been highlighted – please check these carefully and indicate if the first name or surname have been marked up incorrectly. Please note that this will affect indexing of your article, such as in PubMed.</p> <p>Please check and ensure that the licence type at the end of the paper matches the version on the Licence to Publish form submitted.</p>	

SCIENTIFIC REPORTS

OPEN

Formation of oxygen vacancies and Ti^{3+} state in TiO_2 thin film and enhanced optical properties by air plasma treatment

Received: 03 May 2016
Accepted: 03 August 2016
Published: xx xx xxxx

Bandna Bharti¹, Santosh Kumar², Heung-No Lee³ & Rajesh Kumar¹

This is the first time we report that simply air plasma treatment can also enhance the optical absorbance and absorption region of titanium oxide (TiO_2) films, while keeping them transparent. TiO_2 thin films having moderate doping of Fe and Co exhibit significant enhancement in the aforementioned optical properties upon air plasma treatment. The moderate doping could facilitate the formation of charge trap centers or avoid the formation of charge recombination centers. Variation in surface species viz. Ti^{3+} , Ti^{4+} , O^{2-} , oxygen vacancies, OH group and optical properties was studied using X-ray photon spectroscopy (XPS) and UV-Vis spectroscopy. The air plasma treatment caused enhanced optical absorbance and optical absorption region as revealed by the formation of Ti^{3+} and oxygen vacancies in the band gap of TiO_2 films. The samples were treated in plasma with varying treatment time from 0 to 60 seconds. With the increasing treatment time, Ti^{3+} and oxygen vacancies increased in the Fe and Co doped TiO_2 films leading to increased absorbance; however, the increase in optical absorption region/red shift (from 3.22 to 3.00 eV) was observed in Fe doped TiO_2 films, on the contrary Co doped TiO_2 films exhibited blue shift (from 3.36 to 3.62 eV) due to Burstein Moss shift.

Among various metal oxide semiconductors, TiO_2 is considered as a prime candidate due to its many peculiar properties^{1,2} for diverse applications. It is the most suitable candidate for photocatalytic applications due to its biological and chemical inertness, strong oxidizing power, non-toxicity and long term stabilization against photo and chemical corrosion³. The films of TiO_2 have valuable applications in LEDs, gas sensors, heat reflectors, transparent electrodes, thin film photo-anode to develop new photovoltaic, photo-electrochemical cells, solar cells and water splitting⁴⁻¹⁰. In anodic applications, it is a preferred material because of its low density/molar mass and structural integrity over many charge and discharge cycles¹¹. However, the efficiency of pure TiO_2 is substantially low because of its wide band gap and fast recombination of photo-generated electrons and holes. The key issue to improve the performance of TiO_2 relies on efficient light harvesting, including the increase of its photo-efficiency and expansion of photo-response region, and to ensure efficient number of photo-generated electrons and holes reaching to the surface before their recombination. In order to meet these desired performances the bands structure modification of TiO_2 is preferred.

Generally, three fundamental approaches are implemented for band structure modification viz. doping with metallic/non-metallic elements or co-doping of metallic and non-metallic elements^{1,12-14}, modification via introducing defects such as oxygen vacancies and Ti^{3+} in the band gap^{15,16}, and surface modification by treatment methods^{11,17-19}. In metallic doping, among the range of dopants such as Ni, Mn, Cr, Cu, Fe etc.^{3,20-23}, the Fe is found most suitable due to its half filled electronic configuration. Similarly, from non-metallic dopants S, C, F, N etc.²⁴⁻²⁷, the N is preferred. In the case of metallic dopants, there are some contradictory reports that show disadvantages of thermal and chemical instability of TiO_2 . Also, their high doping although enhances the band gap but at the same time reduces optical/photocatalytic activity because of increasing carrier recombination centers²⁸⁻³¹. What is the mechanism of observed photo-response of doped/modified TiO_2 ; it is still a question, however a

¹Jaypee University of Information Technology, Wagnaghat, Solan-173234, H.P., India. ²School of Materials Science and Engineering, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro Buk-gu, Gwangju, 61005, South Korea. ³Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro Buk-gu, Gwangju, 61005, South Korea. Correspondence and requests for materials should be addressed to R.K. (email: rajesh.kumar@juit.ac.in)

generally accepted concern states that the photo absorption of a material is explained better by introducing the defects in the lattice of TiO_2 . For example, Ti^{3+} and oxygen vacancies³² create trap centers, rather than the recombination centers unlike the high doping case, and results in the variation of band gap of pristine TiO_2 .

On the other hand, surface modification methods including surface hydrogenation³³, vacuum activation³² and plasma treatment³⁴ are also practiced. In the hydrogenation method, the surface of TiO_2 is terminated with hydrogen leading to an enhanced photocatalytic activity³⁵ in visible region; however, it is still unknown that how does the hydrogenation modify a surface to enhance its optical performance (photocatalytic activity)³⁶. The drawback of the hydrogenation method is that it requires high temperature and the obtained TiO_2 sample/film are black³⁵, which makes the films unable for many optoelectronic applications, such as a transparent electrode in optoelectronic devices. Both the vacuum activation and plasma treatment methods create highly stable Ti^{3+} and oxygen vacancies^{32,34}. In vacuum activation method, the sample may exhibit higher absorption intensity but it appears brown in color³⁵, that makes it unable for transparent electrode applications. Finally, in case of plasma treatment methods, generally hydrogen gas is used to create oxygen Ti^{3+} and vacancies in TiO_2 , but it is always avoidable to use such a hazardous and expensive gas. Except hydrogen there are few reports on the use of argon³⁷, oxygen³⁸ and nitrogen plasma³⁹ for surface modification of TiO_2 . We know that the implementation of gas in the treatment chamber may be hazardous and cost effective; therefore, it is always required to avoid the use of hazardous gas, and to implement a simple and low cost approach to meet the requirements. In this regard, treatment by air plasma may be an effective approach. However, to the best of our knowledge there is no report on the application of air plasma for the surface modification of TiO_2 film.

In this report, the band structure modification of thin transparent films of TiO_2 was done by implementing simply the air plasma and thus creating Ti^{3+} and oxygen vacancies in TiO_2 films. The effect of air plasma treatment was studied in conjunction with metallic doping. First, Fe and Co doped TiO_2 thin films were formed on glass substrate, which were subsequently treated in air plasma. Considering the drawback of high metallic doping (formation of recombination centers), in this study, a moderate amount of dopants were used to enhance the optical properties of TiO_2 thin film and thereafter the air plasma was applied to enhance them further. The moderate amount of metallic dopant not only favors the separation of electrons and holes but also narrows the band gap of TiO_2 ³. We observed that simultaneous effect of the joint approaches increases photo absorbance as well as extends photo response region of the films towards both the visible and UV spectrum. The doped films of TiO_2 were treated in plasma with varying treatment time. The moderate doping of Fe and Co elements reduces band gap minutely in both the cases, but when treated with air plasma a significant change in the optical properties was observed due to the formation of Ti^{3+} and oxygen vacancies in the band gap.

Results and Discussion

After fabricating, the thin films of pure TiO_2 , Fe and Co doped TiO_2 were treated in air plasma for 0, 10, 30 and 60 seconds, which were analyzed for surface morphology and crystal structure variations using SEM (see Supplementary Information; Figure S1) and XRD. Here we show XRD pattern of doped thin films for extreme treatment time 0 and 60 seconds (for XRD spectra of samples treated at other treatment time, please see Supplementary Information; Figure S2). Figure 1(a,b) represents XRD pattern of Fe doped, and Fig. 1(c,d) represents XRD patterns of Co doped TiO_2 thin films for 0 (untreated) and 60 seconds of plasma treatment time. Since there is no detection of Fe and Co signals, it indicates that all the Fe and Co ions in the respective samples get incorporated into the structure of TiO_2 by replacing some of Ti ion, and occupying the interstitial sites⁴⁰.

Absence of sharp peak in XRD patterns represents amorphous phase of TiO_2 thin films⁴¹. After plasma treatment 2θ angle and FWHM of the peaks remain almost unchanged, indicating negligible effect on the film structure. XRD indicates that plasma treatment does not create any change in the crystal structure of Fe and Co doped TiO_2 thin films. The obtained low signal-to-noise ratio in the above XRD spectra is due to the low crystallinity of the films and small crystallite size; such observations have been reported by others⁴².

The presence of atomic percentage of the dopants in TiO_2 thin films was detected by EDX signals (see Supplementary Information; Figure S3). The EDX of Fe doped TiO_2 film shows the atomic percentage of Fe, Ti and O as 1.66%, 12.93% and 85.41%, respectively, which closely matches to the stoichiometry of elements in $\text{Ti}_{0.95}\text{Fe}_{0.05}\text{O}_2$. Similarly, in case of Co doped TiO_2 , the obtained atomic percentage of Co, Ti and O in EDX are 1.33%, 23.33% and 75.35%, respectively, which confirms the stoichiometry of elements of $\text{Ti}_{0.95}\text{Co}_{0.05}\text{O}_2$ thin film.

Variation in optical properties of TiO_2 thin films by doping and subsequent air plasma treatment was analyzed by UV-Vis spectrophotometer. The change in absorption edge and corresponding band gap is mentioned in Table 1. Pure TiO_2 film (undoped and untreated) showed absorption edge at 367 nm and band gap 3.37 eV, whereas Fe doped TiO_2 film showed a shift in the absorption edge to 385 nm, with a decreasing in the band gap to 3.22 eV. Similarly, Co doping shifts the absorption edge from 367 nm to 369 nm with a reduction in the band gap to 3.36 eV. The observed red shift in absorption edge and narrowing band gap in both dopants cases is similar to other reports on metallic doping³. In both the cases, samples were doped with a moderate (5%) concentration of Fe and Co forming $\text{Ti}_{0.95}\text{Fe}_{0.05}\text{O}_2$ and $\text{Ti}_{0.95}\text{Co}_{0.05}\text{O}_2$, respectively. We could have tuned the optical properties further by increasing the dopant concentration but that would form recombination centers²⁸; therefore, to avoid the formation of recombination centers, a further tuning in the optical properties was done by treating these moderately doped TiO_2 films in air plasma. The films were treated in air plasma for treatment time (0, 10, 30 and 60 seconds), and investigated for the shift in absorption edge and band gap variation. With increasing treatment time, the absorption edge of Fe doped TiO_2 films shifts continuously from 385 nm (for 0 seconds treatment time) to 413 nm (for 60 seconds treatment time), with a corresponding band gap change from 3.22 eV to 3.00 eV, showing a significant increase in the absorption region. In case of Co doped TiO_2 films, the absorption edge shifts from 369 nm to 342 nm (for 60 seconds treatment time) with a corresponding band gap change from 3.36 to 3.62 eV, which shows an increase in the optical band gap/UV absorption region probably due to the Burstein-Moss effect⁴³ explained latter.

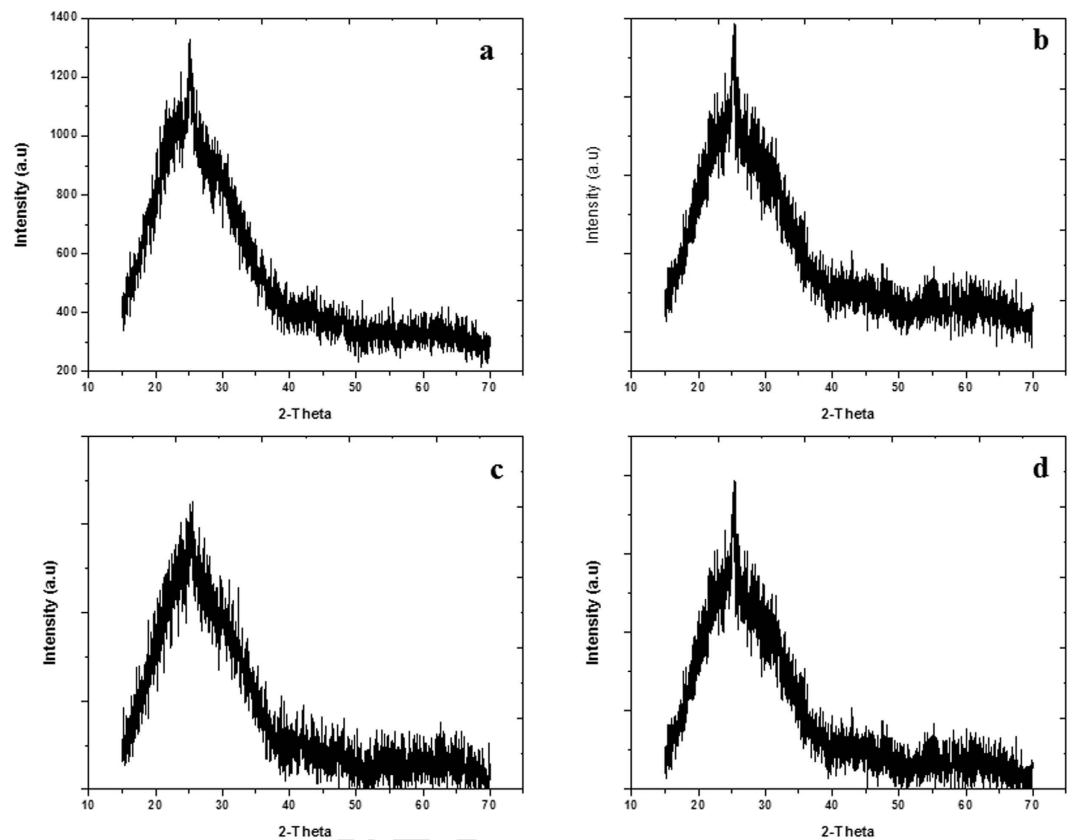


Figure 1. X-ray diffraction spectra of (a) Fe doped/untreated TiO₂ film; plasma treatment time 0 second, (b) Fe doped/treated TiO₂ film, plasma treatment time 60 second, (c) Co doped/untreated TiO₂ film, plasma treatment time 0 second and (d) Co doped/treated TiO₂ film; plasma treatment time 60 second.

Plasma treatment time [sec]	Absorption edge [nm]		Band gap [eV]	
	Fe doping	Co doping	Fe doping	Co doping
0	385	369	3.22	3.36 by doping
10	396	360	3.13	3.44
30	402	345	3.08	3.59 by plasma treatment
60	413	342	3.00	3.62

Table 1. Variation in absorption edge and band gap of Fe and Co doped TiO₂ thin films with plasma treatment time.

From the Table, it is observed that the change in optical properties of TiO₂ films appears at two levels; first by the doping of Fe and Co, and then by plasma treatment. However, here it should be noted that the change in the band gap due to the doping is smaller as compared to the subsequent band gap change by plasma treatment. While discussing the effect of doping on the change of band gap, we know that the reduction may take place due to either by the increasing grain size of highly crystalline sample⁴⁴ or the formation of electronic energy levels within energy band gap⁴⁵. In our study, since the XRD results showed the samples to be amorphous, thus the first reason can be discarded. Therefore, Fe³⁺ and Co²⁺ ions substitute Ti⁴⁺ ions in TiO₂ matrix and cause a change in the band gap by forming their mid gap energy levels in the respective samples along with the formation of Ti³⁺ and oxygen vacancies. The electronic transition from valance band to dopant level and then from dopant level to conduction band, and/or from valance band to oxygen level and then from oxygen level to Ti³⁺ level/dopant level effectively cause a red shift in the absorption edge, showing reduced band gap^{46–48}. In many cases, the localized level of t_{2g} state of the doping element even lies in the middle of band gap (in case of, Cr, Mn or Fe as the doping materials), and at the top of the valance band (when Co is used as a dopant)⁴⁹. Next, the variation in the absorption edge/band gap with plasma treatment time is due to the increase of Ti³⁺ and oxygen vacancies, detailed discussion is given under XPS studies in the following section.

Figure 2 shows variation in the absorption spectra of Fe doped TiO₂ thin film treated for 60 seconds of time (Fig. 2(b)) with respect to untreated one (Fig. 2(a)) (to see the increase in the absorption edge and reduction in

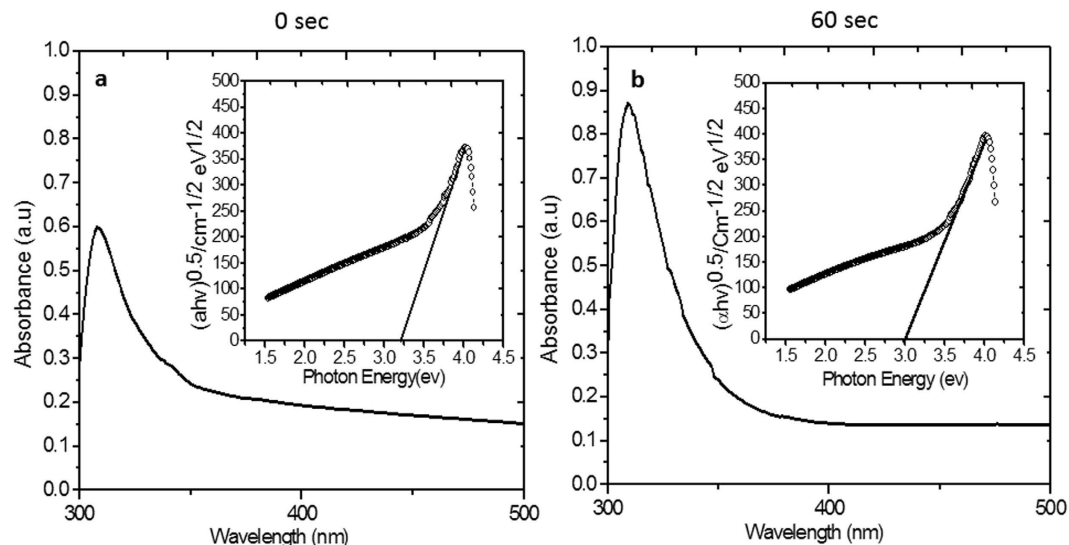


Figure 2. Optical absorption spectra and Tauc plot $((\alpha h\nu)^{1/2}$ versus $h\nu$ plot) in the inset for (a) Fe doped/untreated TiO_2 film; plasma treatment time 0 second and (b) Fe doped/treated TiO_2 film; plasma treatment time 60 second.

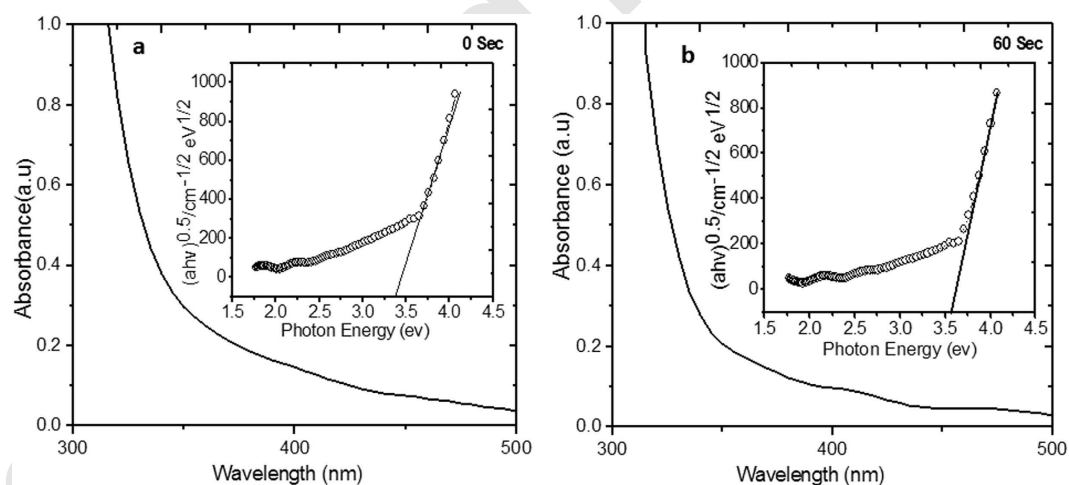


Figure 3. Optical absorption spectra and Tauc plot $((\alpha h\nu)^{1/2}$ versus $h\nu$ plot) in the inset for (a) Co doped/untreated TiO_2 film; plasma treatment time 0 second and (b) Co doped/treated TiO_2 film; plasma treatment time 60 second.

band gap, please refer to Supplementary Information; Figure S4). There is a continuous change in the absorbance, absorption edge and band gap of the films with plasma treatment time. The absorbance of the film increased from 60% (untreated film) to 87% (treated for 60 seconds) along with a red shift in the absorption edge and band gap narrowing by 0.22 eV (Tauc plot shown in the inset of Fig. 2(b)). The band gap and absorption edge were estimated using the following equations⁵⁰:

$$(\alpha h\nu)^{1/2} = C(h\nu - E_g) \quad (1)$$

$$E_{eV} = hc/\lambda \quad (2)$$

where α is absorption coefficient and E_g is band gap energy.

Similarly, the variation in absorption spectra of Co doped TiO_2 thin film treated for 0 and 60 seconds is shown in Fig. 3(a,b) (details of other samples is given in Supplementary Information; Figure S5). In this case, doping shows a red shift due to the presence of Co levels in the energy gap of TiO_2 , whereas after plasma treatment the film shows continuous blue shift with increasing treatment time. This overall shift (due to treatment in plasma for 60 seconds) in the band gap is 0.26 eV. The observed blue shift can be explained by Burstein-Moss effect⁴³, resulted

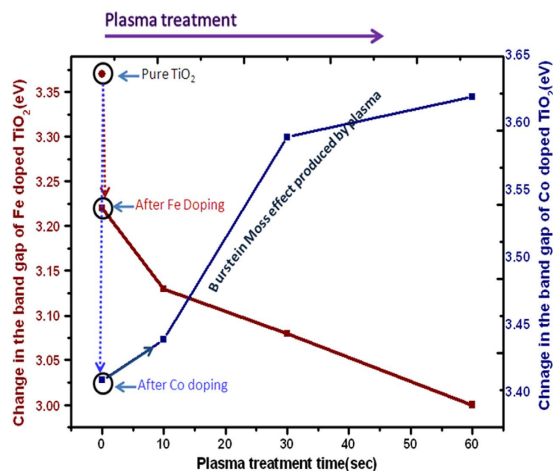


Figure 4. Plots for variation of optical band gap of Fe and Co doped TiO₂ thin film with plasma treatment time.

by the change in the position of Fermi level into the conduction band. General equation representing enhancement in the band gap energy is given by:

$$\Delta E_g^{BM} = \frac{\hbar^2 K_F^2}{2} \left[\frac{1}{m_e^*} + \frac{1}{m_h^*} \right] \quad (3)$$

where m_h^* and m_e^* are the effective mass of hole and electron in the respective bands, and K_F is Fermi wave vector. In our case, the shift of Fermi level into the conduction band leads to the energy band widening. Absorption edge shifts to shorter wavelength region due to the increase in the carrier concentration, which is discussed in XPS studies section.

The overall variation in the absorption edge and band gap of TiO₂ thin film due to the doping (Fe and Co) and air plasma treatment is plotted in Fig. 4. In the plasma treatment region, a remarkable change in the band gap values can be observed with treatment time.

XPS study. In order to understand the mechanism resulting the change in the band gap of Fe and Co doped TiO₂ films with plasma treatment time, the films were investigated by XPS. The XPS being surface sensitive technique provides information about the change in chemical state of film constituting species. Here, the variation in the chemical state of elements 'O' and 'Ti' with plasma treatment time was analyzed in detail to correlate it with the observed variations in the band gap of the films. Figure 5(a,b) show XPS survey spectra of untreated and plasma treated Fe and Co doped TiO₂ thin films, respectively. In these spectra, Cls is probably an instrumental impurity. The intensities of O1s and Ti2p peaks increase with the increasing plasma treatment time, indicating an increase in these states with treatment time.

Figure 6(a) shows high resolution XPS spectrum of pure TiO₂ film. In this spectrum, the doublet 'Ti2p_{3/2}' (binding energy 458.6 eV) and Ti2p_{1/2} (binding energy 464.4 eV) arises from spin orbit-splitting. These peaks are consistent with Ti⁴⁺ in TiO₂ lattice⁵¹. Also, the shoulder Ti2p_{1/2} at binding energy 460.2 eV is corresponding to Ti³⁺⁵² in Ti₂O₃. This indicates that both TiO₂ and Ti₂O₃ are formed in the film (Without deconvolution, the XPS spectra are shown in Supplementary Figure S6). After doping with Fe, the high resolution XPS spectrum (Fig. 6(b)) shows a slight shift in the position along with a variation in the area of the original peaks. The peaks in the Fe doped samples are now located at binding energies 458.4 (Ti2p_{3/2}), 464.3 eV (Ti2p_{1/2}) and 459.0 eV (Ti2p_{1/2}), respectively (see Supplementary Information; Table S1). The shift in the position of these peaks indicates an influence of Fe addition on the electronic state of Ti element; probably some of the Ti ions get substituted with Fe ions in the lattices. After doping, the area of Ti³⁺ peak increased by 81% and that of the peak Ti⁴⁺ decreased by 19%. The increase in the area of Ti³⁺ peak indicates that either Ti₂O₃ is formed in large amount or some mixed oxide structure with Fe (having oxidation state Ti³⁺) is formed after doping. Meanwhile, the decreasing area of Ti⁴⁺ indicates a reduction of TiO₂ in the sample, and probably formation of Ti-O-Fe structure in the TiO₂ lattice through the substitution of transition metal ions. Observed shift in the peaks also indicates interaction between Ti and Fe atoms and an overlapping of their 3d orbital⁵³. This causes an electronic excitation from Fe to Ti in the optical absorption experiment, which shows a reduction in the band gap of Fe doped TiO₂ film (as observed in the optical analysis).

After doping, the film was treated in air plasma. In the XPS results, only the sample which was treated for 60 seconds in plasma is demonstrated. The XPS shows a further increase in the peak corresponding to Ti³⁺ at 459.0 eV (Fig. 6(c)) and a decrease in the peak area of Ti⁴⁺. The change in stoichiometry was estimated by the change in the area of relative peaks. The peak area of Ti³⁺ increases by 20% and that of Ti⁴⁺ decreases by 12%. The increase in the peak area of Ti³⁺ indicates that after plasma treatment there is removal of oxygen from the lattice, which shows a relative increase in Ti³⁺ in the XPS spectrum. On the other hand decreasing peak area of Ti⁴⁺ is

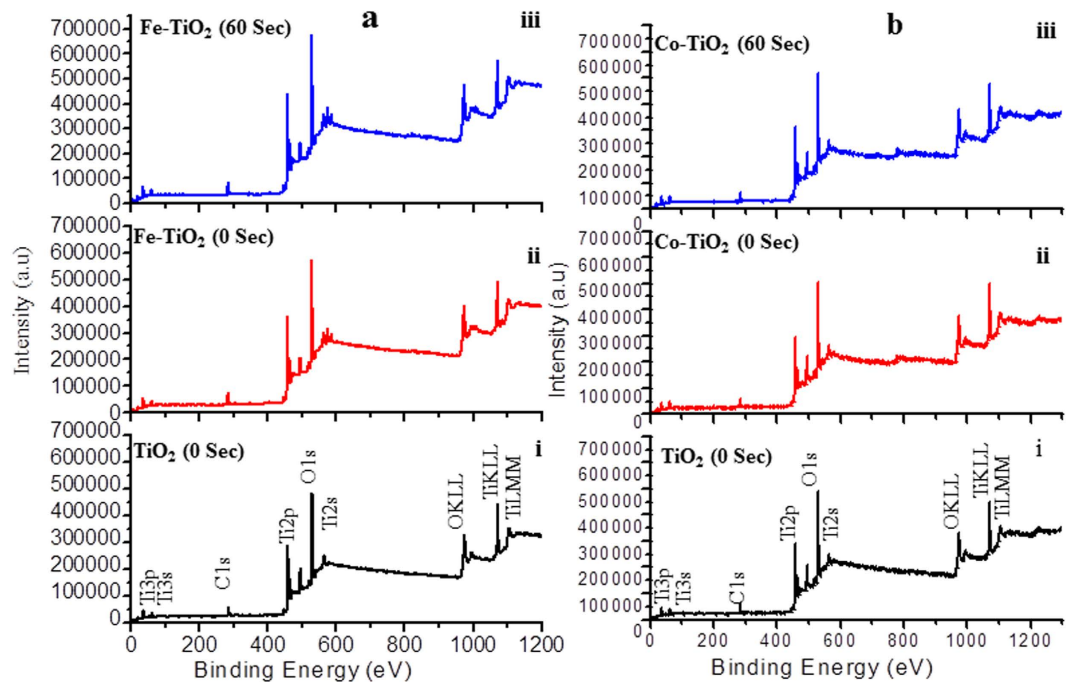


Figure 5. XPS survey spectra in a(i) pure TiO₂ film indicating all the peaks of elements present in the sample, here the appeared carbon peak is instrumental impurity, a(ii) Fe doped/untreated TiO₂ film; plasma treatment time 0 seconds, a(iii) Fe doped/treated TiO₂; plasma treatment time 60 seconds, b(i) pure TiO₂ film which is similar to a(i), and b(ii) Co doped/untreated TiO₂ film; plasma treatment time 0 seconds, b(iii) Co doped/treated TiO₂ film; plasma treatment time 60 seconds.

inferred due to the reaction of Ti⁴⁺ with electrons coming either from plasma or due to the formation of oxygen vacancies in the surface layer generated by the plasma treatment⁴¹. Now, as observed in optical analysis, the band gap of Fe doped films (3.22 eV) decreased to 3.00 eV (for 60 seconds of treatment time), this is correlated with the increasing carrier/electrons density due to plasma treatment. As we know that in the doped samples, the possible reasons of red shift/decreasing band gap is the introduction of donor states in the energy gap (here oxygen vacancies and Ti³⁺, Table 1). In the present case, the band gap decreases further with increasing treatment time, while the concentration of the dopant was kept constant, which is due to the change in the surface states of the constituents i.e. Ti element and oxygen vacancies.

Next, the O1s spectrum of pure TiO₂ thin film is shown in Fig. 6(d), which is fitted with three peaks. The peaks at binding energies 529.9 eV, 530.3 eV and 531.3 eV are attributed to lattice oxygen, Ti₂O₃ and non-lattice oxygen^{54,55}. Similarly, for the doped sample, O1s spectrum of Fe doped TiO₂ thin film fitted with two peaks is shown in Fig. 6(e). In this spectrum, only two peaks at binding energies 529.8 eV, and 531.9 eV are observed which are attributed to lattice oxygen and surface adsorbed OH group, whereas the peak 530.3 eV corresponding to Ti₂O₃, disappears. This indicates that in the doping process TiO₂ is formed along with some mixed oxide. Again, the change in stoichiometry was estimated by the change in area of relative peaks. In case of Fe doped TiO₂ film, the area of the peak at 529.7 increases by 64% and that of the peak at 531.5 eV increases by 54%.

After plasma treatment, the binding energy of lattice oxygen (O in TiO₂) shifts slightly from 529.8 eV to 529.7 eV (Fig. 6(f)), whereas its area increases by 35%. Also, the area of the peak at 531.5 eV (non-lattice oxygen/OH) increases by 15% (see Supplementary Information; Table S1). The increase in the area of non-lattice oxygen indicates the formation of oxygen vacancies in the lattice. This result is analogous to the XPS spectrum of Ti2p (Fig. 6(c)).

Fe doping results in a minor shift in the binding energy, indicating that Fe ions are better dispersed in the substitutional sites of TiO₂ lattice and produce more mixed oxide structure, probably Fe-O-Ti. Figure 7(a) shows high resolution XPS spectrum (for Fe2p_{3/2}) of Fe doped TiO₂ film. After plasma treatment, the high resolution XPS spectrum of Fe2p_{3/2} is shown in Fig. 7(b). These spectra are fitted with Gauss-peak shapes as shown in Fig. 7(c,d). The deconvoluted XPS spectrum of Fe2p_{3/2} (Fig. 7(c,d)) contains main peaks at 710.1 eV and 724.6.1 eV corresponding to Fe2p_{3/2} and Fe2p_{1/2}, respectively (see Supplementary Information; Table S2). The appearance of these peaks supports the presence of Fe in Fe³⁺ ionic state⁵⁵. Further, after plasma treatment the shift in the binding energy of Fe2p_{3/2} from 710.1 eV to 711.3 eV also indicates the presence of Fe³⁺ species, irrespective of the particular oxide (i.e., Fe₂O₃, Fe₃O₄, and FeOOH). Shake up satellite at 716.9 eV also supports that Fe is presented in Fe³⁺ state (oxide)⁵⁶. These shake-up satellites are associated with Fe3d-O2p hybridization. Thus XPS analysis confirmed that Fe ions are doped into TiO₂ matrix in the form of Fe-O-Ti. From the XPS analysis, we confirmed that by increasing the plasma treatment time the concentration of Ti³⁺ and oxygen vacancies also increases.

The Co doped samples after treating in plasma show adverse effect on the band gap of the doped TiO₂ film. In this case, band gap increases with the increasing treatment time as observed in optical studies. To investigate this

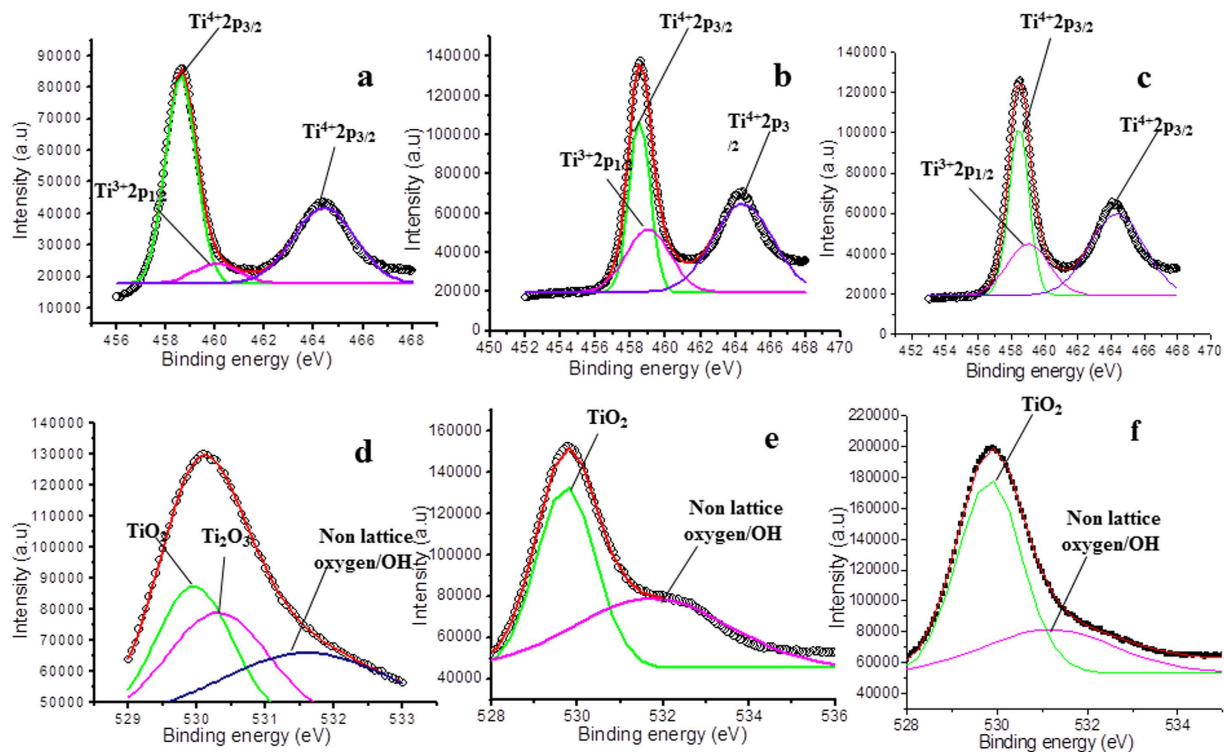


Figure 6. High resolution XPS spectra of Ti2p and O1s in (a) pure/untreated TiO₂ film, (b) Fe doped/untreated TiO₂ film; plasma treatment time 0 second, (c) Fe doped/treated TiO₂ film; plasma treatment time 60 seconds, (d) O1s for pure/untreated TiO₂ film, (e) O1s for Fe doped/untreated TiO₂ film; plasma treatment time 0 second, and (f) O1s for Fe doped/treated TiO₂ film; plasma treatment time 60 seconds.

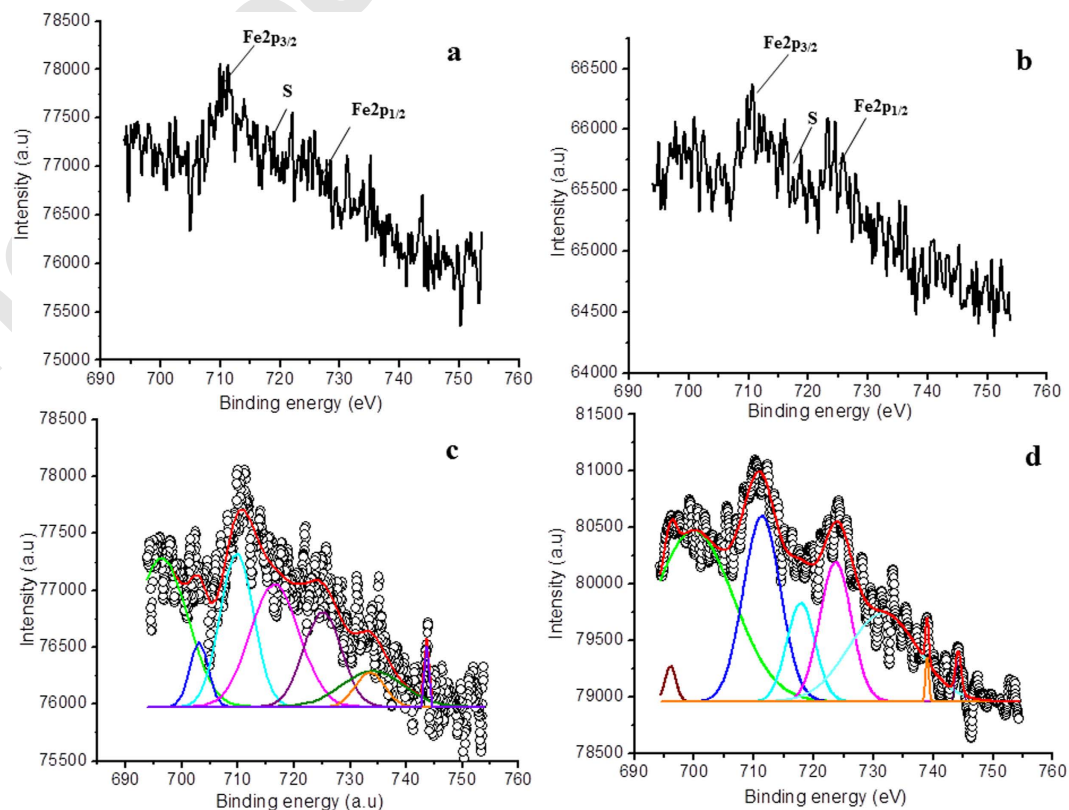


Figure 7. High resolution XPS spectra of Fe2p in (a) Fe doped/untreated TiO₂ film; plasma treatment time 0 second, (b) Fe doped/treated TiO₂ film; plasma treatment time 60 seconds, (c,d) are Gaussian fit of (a,b).

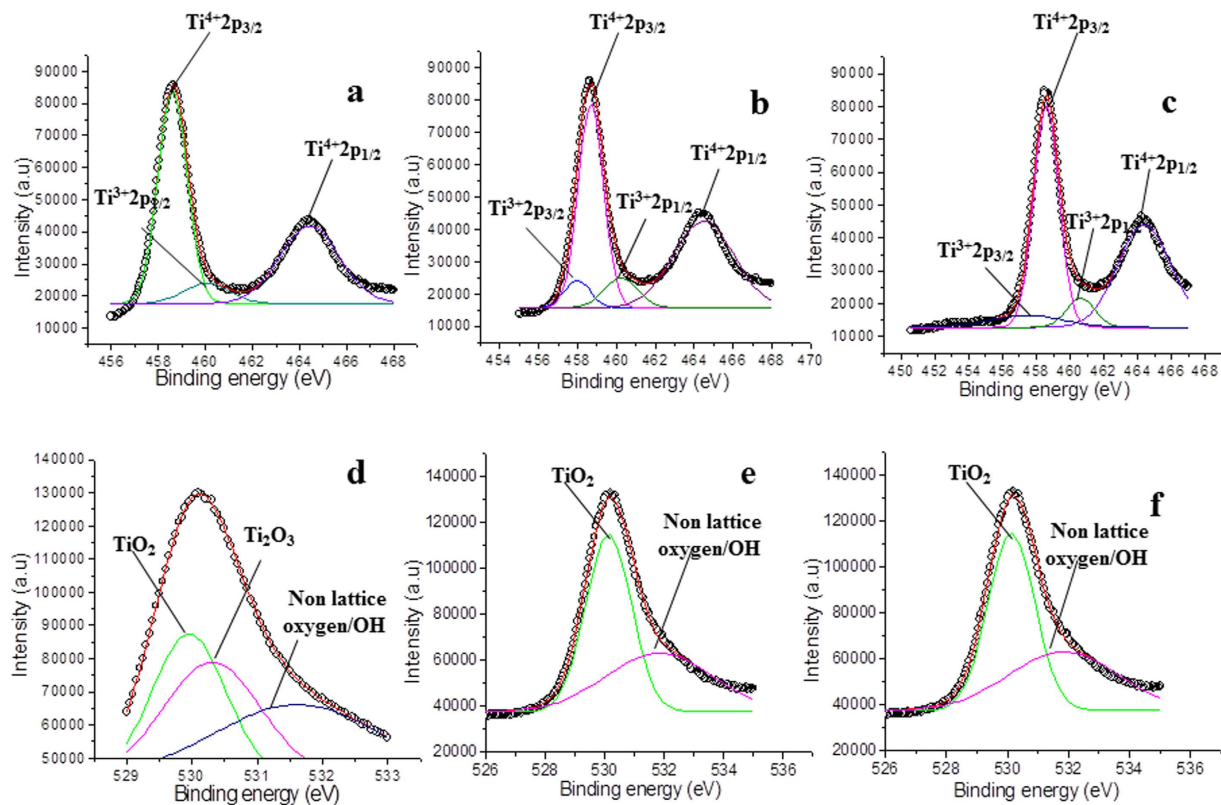


Figure 8. High resolution XPS spectra of Ti2p and O1s in (a) pure/untreated TiO₂ film, (b) Co doped/untreated TiO₂ film; plasma treatment time 0 second, (c) Co doped/treated TiO₂ film; plasma treatment time 60 seconds, (d) O1s for pure/untreated TiO₂ film, (e) O1s for Co doped/untreated TiO₂ film; plasma treatment time 0 second, and (f) O1s for Co doped/treated TiO₂ film; plasma treatment time 60 seconds.

divergent behavior, the samples were analyzed via XPS, Fig. 8 shows high resolution spectra. Figure 8(a) shows the XPS spectrum of pure TiO₂, and Fig. 8(b) shows XPS for Co doped sample. As discussed above in the case of Fe doped sample, the XPS of pure TiO₂ is also fitted with three peaks corresponding to titanium dioxide (Ti⁴⁺) and titanium sub oxide (Ti³⁺) in Ti2p_{1/2} and Ti2p_{3/2}, respectively. These peaks are fitted as Ti⁴⁺2p_{1/2} at 464.4 eV, Ti⁴⁺2p_{3/2} at 458.6 eV, and Ti³⁺2p_{3/2} at 460.2 eV. The line separation between Ti2p_{1/2} and Ti2p_{3/2} is 5.8 eV, which is consistent with the standard binding energy of TiO₂⁵¹. However, in this case the Ti2p spectrum (Fig. 8(b)) is fitted with four peaks as 464.4 for Ti⁴⁺2p_{1/2}, 458.6 eV for Ti⁴⁺2p_{3/2}, 460.4 for Ti³⁺2p_{3/2} and 457.9 eV for Ti³⁺2p_{1/2}⁵⁷, respectively (see Supplementary Information; Table S1). In comparison to the pure TiO₂, the area of Ti³⁺ peak in Co doped TiO₂ increases by 26%, while that of the peak Ti⁴⁺ decreases by 7%, indicating a reduction in the formation of TiO₂, which is similar to the case of Fe doped samples.

After the plasma treatment (Fig. 8(c)), binding energies of the mentioned peaks are shifted slightly to the positions such as 464.3 eV (Ti⁴⁺2p_{1/2}), 458.5 eV (Ti⁴⁺2p_{3/2}), 460.6 eV (Ti³⁺2p_{3/2}) and 457.4 eV (Ti³⁺2p_{1/2}), respectively. The change in stoichiometry was estimated by the change in peak area of respective peaks.

After plasma treatment, while investigating for peak area, we observed that the peak area of Ti³⁺ increases by 30%, whereas the peak area of Ti⁴⁺ decreases by 12%. Again, this is expected due to the reaction of Ti⁴⁺ with the electrons coming either from plasma or due to the formation of oxygen vacancies in the surface layer by the plasma treatment. Further, the high resolution O1s XPS spectrum obtained for Co doped sample is shown in Fig. 8(d-f). The spectrum is fitted with three peaks i.e. 529.9 eV, 530.3 eV and 531.6 eV that correspond to lattice oxygen of TiO₂, oxygen in Ti₂O₃ and non-lattice oxygen, respectively.

The change in stoichiometry was estimated by change in the peak area of relative peaks. With the doping of Co, the lattice oxygen (corresponding to TiO₂) peak at 529.9 shifts to the position 530.3 eV, and the area of the peaks at 530.3 eV and 531.6 eV increases by 51% and 24%, respectively. The original peak at 530.3 eV (Fig. 8(d)) corresponding to Ti₂O₃ disappears after doping (Fig. 8(e)), which is due to the formation of mixed oxide structure. Further, with the increasing treatment time, the areas of the peaks at 530.3 eV and 531.6 eV ((Fig. 8(f)) also increases by 24% and 25%, respectively. (To explain in a more quantitative manner we have tabulated all the data in a table by comparing all the peaks at different plasma treatments time see Supplementary Information; Table S1).

Next, Fig. 9(a) corresponds to high-resolution XPS spectra of Co2p region of Co doped TiO₂ thin films and Fig. 9(b) shows high-resolution XPS spectra with plasma treatment. Figure 9(c,d) represent deconvoluted XPS spectra of doped TiO₂ and plasma treated TiO₂ thin films, respectively. The core level binding energies of peaks Co2p_{1/2} and Co2p_{3/2} are 796.9 eV and 781.0 eV, respectively. The satellite peaks at 787 eV and 802 eV reveal high spin Co(II) state with complex transitions⁵⁸. These results are an indication that Co does not precipitate as

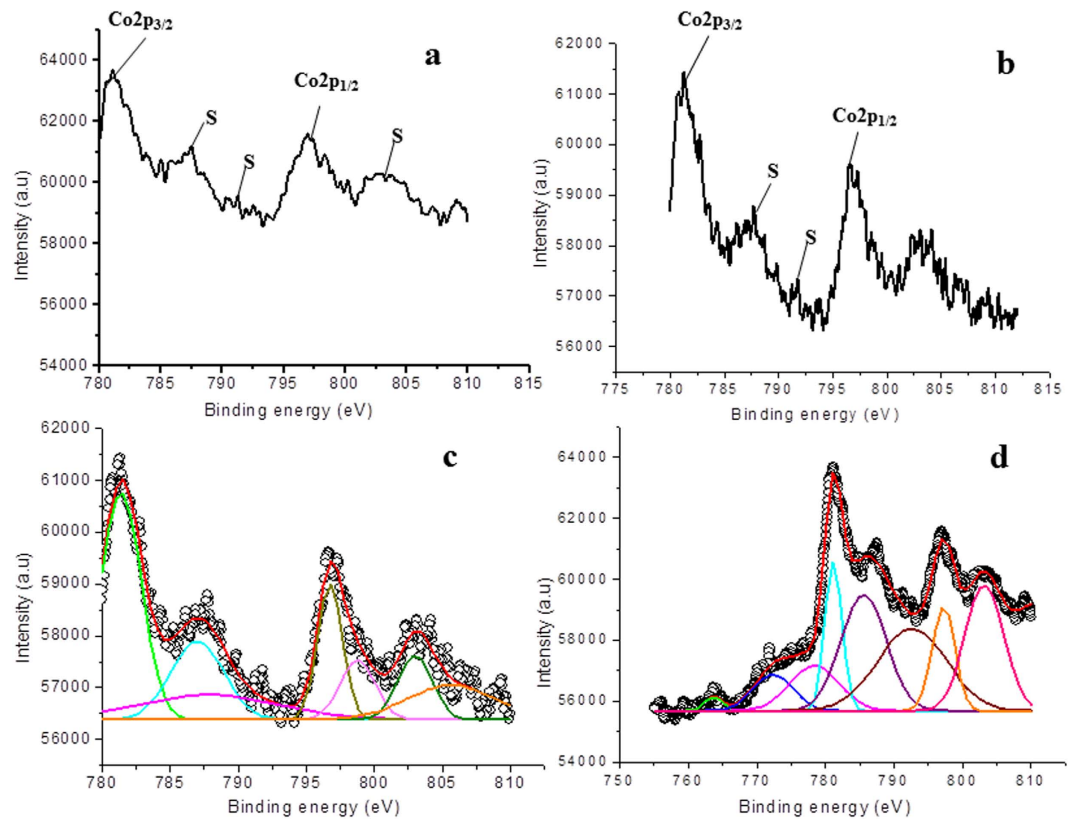


Figure 9. High resolution XPS spectra of Co2p in (a) Co doped/untreated TiO₂ film; plasma treatment time 0 second, (b) Co doped/treated TiO₂ film; plasma treatment time 60 seconds, (c,d) are Gaussian fit of (a,b).

metallic Co on the film surface. After plasma treatment, the satellites peaks shifts slightly to the 785.3 eV and 802.3 eV. Also, the binding energies of Co2p_{1/2} and Co2p_{3/2} are shifted to 796.6 eV and 781.2 eV, respectively (see Supplementary Information; Table S1). These spectra are typical of compounds containing high-spin Co²⁺ ions^{59,60}, revealing the presence of CoO(Co²⁺), CoTiO₃ (Co²⁺), Co₂O₃ (Co³⁺) or mixed valence Co₃O₄ (Co²⁺ and Co³⁺) in the surface. The presence of strong satellites indicates that Co atoms in the doped TiO₂ film are in 2+ oxidation state, referring the possible formation of CoO or CoTiO₃ inside the film.

Now we discuss the probable reason of band gap narrowing in TiO₂ film with Fe doping, and widening in the case of Co doping after plasma treatment. As reported, the iron dopant acts as an acceptor impurity in TiO₂ lattice⁶¹. Thus when the TiO₂ film is doped with Fe, the acceptor levels of Fe along with oxygen vacancies are created in the band gap of TiO₂⁶². In our case, as discussed above Ti³⁺ is also formed which creates energy level in the band gap, contributing to the reduction of band gap. Next, when this Fe doped TiO₂ film was treated in air plasma, the Ti³⁺ levels and oxygen vacancies increases further with the treatment time, whereas no change in the dopant levels occurs as the dopant concentration was kept constant. The increase in Ti³⁺ levels and oxygen vacancies would further reduce the band gap of Fe doped TiO₂ film. In case of Co doping, there is a formation of Co acceptor levels along with Ti³⁺ and oxygen vacancies levels in the band gap which reduces the band gap of Co doped TiO₂ film. But when the film was treated with plasma we observed continuous widening in the band gap with treatment time. The observed increase in the band gap can be explained by Burstein-Moss effect. The probable reason for Burstein-Moss shift in this case is that with the treatment time the Ti³⁺ levels and oxygen vacancies increases more as compared to Fe doped case. By plasma treatment for 60 seconds the Ti³⁺ increases by 20%, oxygen vacancies increases by 15% in case of Fe doped TiO₂, whereas Co doped TiO₂ Ti³⁺ increases 30%, oxygen vacancies increases 25%. These created levels donate more electrons and thus shift the Fermi level to the conduction band, which increases the band gap of Co doped TiO₂ film. The exact reason for this divergent behavior is unclear as of now but the most appropriate reason seems to us is, the on-site coulomb interaction/repulsion that are occurring only in case of Co doped TiO₂ films⁶³. When Co²⁺ ion substitutes Ti⁴⁺ ions, the imbalance positive charge inside the lattice is compensated by the formation of oxygen vacancies located near Co ion. The formation of oxygen vacancies is equivalent to the addition of two electrons per Co ion^{64,65}. The oxygen vacancies produced in case of Co doped TiO₂ thin films are higher as compared to Fe doped TiO₂ films as observed by XPS. Suppose both Fe and Co doped films increase by same values of Ti³⁺ levels and oxygen vacancies, but due to Columbian interactions, which are only in case Co doped TiO₂^{64,65}, the optical transition results in the blue shift of the absorption spectra. The proposed mechanism for both the Fe and Co doped TiO₂ is illustrated in Fig. 10.

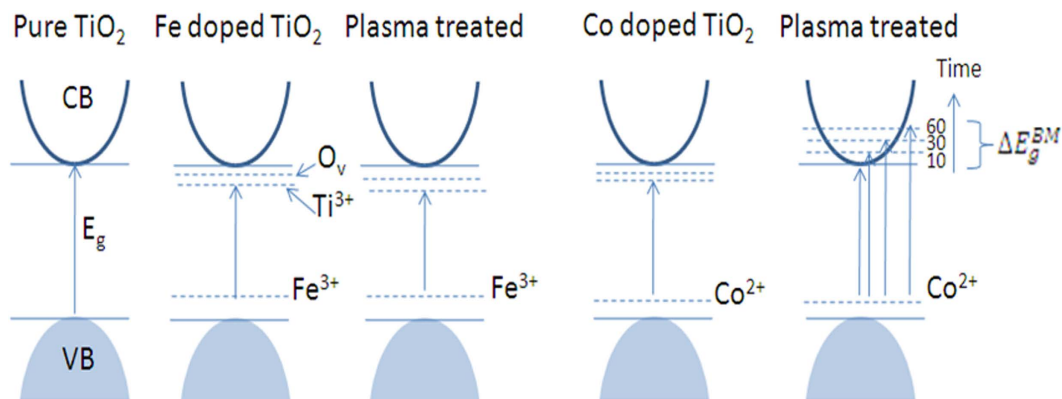


Figure 10. Schematic diagram of the energy levels of (a) pure/untreated TiO_2 films, (b) Fe doped/untreated TiO_2 film, (c) Fe doped/treated TiO_2 ; for 60 seconds of treatment time, (d) Co doped/untreated TiO_2 film, (e) Co doped/treated TiO_2 film; for 10, 30 and 60 seconds of treatment time, indicating Burstein Moss effect. (O_v represents oxygen vacancies).

Conclusion

Treatment by air plasma leads to significant change in the optical properties of TiO_2 thin films. Unlike other treatment methods, in this approach the transparency of TiO_2 thin film remains invariant. The charge separation centers i.e. oxygen vacancies and Ti^{3+} is created with the doping of metallic Fe and Co elements; however, they are significantly enhanced by the air plasma treatment. In Fe doped TiO_2 thin film, the formation of oxygen vacancies and Ti^{3+} causes enhances absorbance and red shift due to the formation of energy levels in the band gap, whereas in Co doped TiO_2 the Burstein-Moss shift is effective to make blue shift in the absorption spectra. Conclusively, we can say that the joint approaches i.e. low level/moderate doping and safe and low cost air plasma treatment resulted in enhanced optical properties of transparent TiO_2 thin films, making them efficient candidate for transparent electrode applications.

Experimental Methods

Thin films of TiO_2 , Fe doped TiO_2 ($\text{Ti}_{0.95}\text{Fe}_{0.05}\text{O}_2$) and Co doped TiO_2 ($\text{Ti}_{0.95}\text{Co}_{0.05}\text{O}_2$) were fabricated on glass substrate using dip-coating method. Titanium (IV) isopropoxide (TTIP, $\text{Ti}[\text{OCH}(\text{CH}_3)_2]_4$, 97%, Aldrich) was used as precursor solution. First of all triethanolamine $\text{C}_6\text{H}_{15}\text{NO}_3$, a stabilization agent was dissolved in $\text{C}_2\text{H}_5\text{OH}$, which resulted in a colorless solution. In this solution, the precursor solution $\text{Ti}[\text{OCH}(\text{CH}_3)_2]_4$ was added dropwise to form a pale yellow solution with a continuous stirring. To avoid the precipitation of TiO_2 , $\text{C}_2\text{H}_5\text{OH}$ and H_2O was added in a ratio 9:1. Now during the sol gel synthesis solutions of ferric nitrate ($\text{Fe}(\text{NO}_3)_3 \cdot 9\text{H}_2\text{O}$), and cobalt nitrate ($\text{Co}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$) were added separately as the dopant in TiO_2 . These solutions were stirred for two hours and allowed for ageing overnight. Then glass substrates cleaned with H_2O , detergent, $\text{C}_3\text{H}_8\text{O}$ and $\text{C}_2\text{H}_5\text{OH}$ were coated with the aged solution. Coated films were dried and annealed at 400°C to form transparent thin films. The fabricated films were treated in air plasma, generated in a vacuum coating unit (Hindhivac model: 12A4D), for varying treatment time; 0, 10, 30, and 60 seconds, respectively. The air plasma was generated at reduced pressure of 10^{-3} mbar in the vacuum chamber. During the treatment process the applied bias voltage was 30 volts with a power of 22.7 watt. After treating in plasma, the samples were analyzed for optical, structural, morphological and surface properties.

Materials Characterization. The optical (absorbance, shift in absorption edge and band gap) properties of the films were studied by UV-Vis spectrophotometer (Perkin-Elmer Lambda 750). The band gap of Fe and Co doped thin films was calculated by using the absorbance spectra by plotting $(\alpha h\nu)^{1/2}$ against $h\nu$, where $h\nu$ being incident photon energy. Surface morphology was studied using scanning electron microscopy (SEM), and elemental confirmation was done using energy dispersive X-ray (EDX). The structural analysis of the samples was done using X-ray diffractometer (XRD) (company name Rigaku, with $\text{Cu K}\alpha$ radiation, $\lambda = 1.5406 \text{ \AA}$), and to observe the effect of plasma treatment on surfaces states, X-ray photoelectron spectroscopy (XPS: VG Multilab 2000, Thermo electron corporation, UK) studies were performed.

References

1. Hamal, D. B. *et al.* A multifunctional biocide/sporicide and photocatalyst based on titanium dioxide (TiO_2) codoped with silver, carbon, and sulfur. *Langmuir* **26**, 2805–2810 (2010).
2. O'Regan, B. & Grätzel, M. A. Low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. *Nature* **353**, 737–740 (1991).
3. Zhou, M., Yu, J. & Cheng, B. Effects of Fe-doping on the photocatalytic activity of mesoporous TiO_2 powders prepared by an ultrasonic method. *J. Hazard. Mater.* **137**, 1838–1847 (2006).
4. Lin, Y., Zhou, S., Liu, X., Sheehan, S. & Wang, D. $\text{TiO}_2/\text{TiSi}_2$ Heterostructures for High-Efficiency Photoelectrochemical H_2O Splitting. *J. Am. Chem. Soc.* **131**, 2772–2773 (2009).
5. Sauvage, F. *et al.* Dye-Sensitized Solar Cells Employing a Single Film of Mesoporous TiO_2 Beads Achieve Power Conversion Efficiencies Over 10%. *ACS Nano*. **4**, 4420–4425 (2010).
6. Zhang, Y. *et al.* Synthesis and characterization of TiO_2 nanotubes for humidity sensing. *Appl. Surf. Sci.* **254**, 5545–5547 (2008).

7. Lee, K. S., Lim, J. W., Kim, H. K., Alford, T. L. & Jabbour, G. E. Transparent conductive electrodes of mixed TiO_{2-x}-indium tin oxide for organic photovoltaics. *Appl. Phys. Lett.* **100**, 213302 (2012).
8. Rabaste, S. *et al.* Sol-gel fabrication of thick multilayers applied to Bragg reflectors and microcavities. *Thin Solid Films* **416**, 242–247 (2002).
9. Liao, Y. *et al.* New Mechanistic Insight of Low Temperature Crystallization of Anodic TiO₂ Nanotube Array in Water. *Cryst. Growth Des.* **16**, 1786–1791 (2016).
10. Gültekin, A. Effect of Au Nanoparticles Doping on the Properties of TiO₂ Thin Films. *Mater. Sci.* **20**, 10–14 (2014).
11. Shin, J., Joo, J. H., Samuelis, D. & Maier, J. Oxygen-Deficient TiO_{2-δ} nanoparticles via hydrogen reduction for high capability Lithium batteries. *Chem. Mater.* **24**, 543–551 (2012).
12. Peng, B. *et al.* General synthesis and optical properties of monodisperse multifunctional metal-ion-doped TiO₂ hollow particles. *J. Phys. Chem. C* **113**, 20240–20245 (2009).
13. Asahi, R., Mikawa, T., Ohwaki, T., Aoki, K. & Taga, Y. Visible Light Photocatalysis in Nitrogen-Doped Titanium Oxides. *Science*. **293**, 269–271 (2001).
14. Kurtoglu, M. E., Longenbach, T., Sohlberg, K. & Gogotsi, Y. Strong Coupling of Cr and N in Cr -N-doped TiO₂ and Its Effect on Photocatalytic Activity. *J. Phys. Chem. C* **115**, 17392–17399 (2011).
15. Amano, F., Nakata, M., Yamamoto, A. & Tanaka, T. Effect of Ti³⁺ Ions and Conduction Band Electrons on Photocatalytic and Photoelectrochemical Activity of Rutile Titania for Water Oxidation. *J. Phys. Chem. C* **120**, 6467–6474 (2016).
16. Chen, C. S. *et al.* Effect of Ti³⁺ on TiO₂-supported Cu catalysts used for CO oxidation. *Langmuir* **28**, 9996–10006 (2012).
17. Liu, H. *et al.* The enhancement of TiO₂ photocatalytic activity by hydrogen thermal treatment. *Chemosphere* **50**, 39–46 (2003).
18. Nakamura, I., Sugihara, S. & Takeuchi, K. Mechanism for NO Photooxidation over the Oxygen-Deficient TiO₂ Powder under Visible Light Irradiation. *Chem. Lett.* **29**, 1276–1277 (2000).
19. Zhang, Z. K., Bai, M. L., Guo, D. Z., Hou, S. M. & Zhang, G. M. Plasma-electrolysis synthesis of TiO₂ nano/microspheres with optical absorption extended into the infra-red region. *Chem. Commun.* **47**, 8439–8441 (2011).
20. Jing, D., Zhang, Y. & Guo, L. Study on the synthesis of Ni doped mesoporous TiO₂ and its photocatalytic activity for hydrogen evolution in aqueous methanol solution. *Chem. Phys. Lett.* **415**, 74–78 (2005).
21. Gracia, F., Holgado, J. P., Caballero, A. & Gonzalez-Elipe, A. R. Structural, optical, and photoelectrochemical properties of Mn²⁺-TiO₂ model thin film photocatalysts. *J. Phys. Chem. B* **108**, 17466–17476 (2004).
22. Zhu, J. *et al.* Hydrothermal doping method for preparation of Cr³⁺-TiO₂ photocatalysts with concentration gradient distribution of Cr³⁺. *Appl. Catal. B Environ.* **62**, 329–335 (2006).
23. Colón, G., Maicu, M., Hidalgo, M. C. & Navío, J. A. Cu-doped TiO₂ systems with improved photocatalytic activity. *Appl. Catal. B Environ.* **67**, 41–51 (2006).
24. Yang, K., Dai, Y. & Huang, B. Understanding photocatalytic activity of S- and P-doped TiO₂ under visible light from first-principles. *J. Phys. Chem. C* **111**, 18985–18994 (2007).
25. Ren, W. *et al.* Low temperature preparation and visible light photocatalytic activity of mesoporous carbon-doped crystalline TiO₂. *Appl. Catal. B Environ.* **69**, 138–144 (2007).
26. Yu, J. C., Yu, J., Ho, W., Jiang, Z. & Zhang, L. Effects of F-doping on the photocatalytic activity and microstructures of nanocrystalline TiO₂ powders. *Chem. Mater.* **14**, 3808–3816 (2002).
27. Sato, S., Nakamura, R. & Abe, S. Visible-light sensitization of TiO₂ photocatalysts by wet-method N doping. *Appl. Catal. A Gen.* **284**, 131–137 (2005).
28. Xing, M., Zhang, J., Chen, F. & Tian, B. An economic method to prepare vacuum activated photocatalysts with high photo-activities and photosensitivities. *Chem. Commun.* **47**, 4947–4949 (2011).
29. Pan, X., Yang, M. Q., Fu, X., Zhang, N. & Xu, Y. J. Defective TiO₂ with oxygen vacancies: synthesis, properties and photocatalytic applications. *Nanoscale* **5**, 3601–3614 (2013).
30. Pan, X., Yang, M. Q. & Xu, Y. J. Morphology control, defect engineering and photoactivity tuning of ZnO crystals by graphene oxide—a unique 2D macromolecular surfactant. *Phys. Chem. Chem. Phys.* **16**, 125589–125599 (2014).
31. Zhang, N., Yang, M. Q., Liu, S., Sun, Y. & Xu, Y. J. Waltzing with the versatile platform of graphene to synthesize composite photocatalysts. *Chem. Rev.* **115**, 10307–10377 (2015).
32. Liu, Y., Wang, J., Yang, P. & Matras-Postolek, K. Self-modification of TiO₂ one-dimensional nano-materials by Ti³⁺ and oxygen vacancy using Ti₂O₃ as precursor. *RSC Adv.* **5**, 61657–61663 (2015).
33. Lu, X. *et al.* Hydrogenated TiO₂ Nanotube Arrays for Supercapacitors. *Nano Lett.* **12**, 1690–1696 (2012).
34. Konstantakou, M. *et al.* Influence of Fluorine Plasma Treatment of TiO₂ Films on the Behavior of Dye Solar Cells Employing the Co (II)/(III) Redox Couple. *J. Phys. Chem. C* **118**, 16760–16775 (2014).
35. Lu, H. *et al.* Safe and facile hydrogenation of commercial Degussa P25 at room temperature with enhanced photocatalytic activity. *RSC Adv.* **4**, 1128–1132 (2014).
36. Zheng, Z. *et al.* Hydrogenated titania: synergy of surface modification and morphology improvement for enhanced photocatalytic activity. *Chem. Commun.* **48**, 5733–5735 (2012).
37. Heo, C. H., Lee, S. B. & Boo, J. H. Deposition of TiO₂ thin films using RF magnetron sputtering method and study of their surface characteristics. *Thin Solid Films* **475**, 183–188 (2005).
38. Li, Y. & Jang, B. W. L. Investigation of calcination and O₂ plasma treatment effects on TiO₂-supported palladium catalysts. *Ind. Eng. Chem. Res.* **49**, 8433–8438 (2010).
39. Yamada, K. *et al.* Photocatalytic activity of TiO₂ thin films doped with nitrogen using a cathodic magnetron plasma treatment. *Thin Solid Films* **516**, 7560–7564 (2008).
40. Cong, Y., Zhang, J., Chen, F., Anpo, M. & He, D. Preparation, photocatalytic activity, and mechanism of nano-TiO₂ co-doped with nitrogen and iron (III). *J. Phys. Chem. C* **111**, 10618–10623 (2007).
41. Bharti, B., Kumar, S. & Kumar, R. Superhydrophilic TiO₂ thin film by nanometer scale surface roughness and dangling bonds. *Appl. Surf. Sci.* **364**, 51–60 (2016).
42. Li, S., Xu, Q., Uchaker, E., Cao, X. & Cao, G. Comparison of amorphous, pseudo hexagonal and orthorhombic Nb₂O₅ for high-rate lithium ion insertion. *Cryst. Eng. Comm.* **18**, 2532–2540 (2016).
43. Bhachu, D. S. *et al.* Solution processing route to multifunctional titania thin films: Highly conductive and photocatalytically active Nb:TiO₂. *Adv. Funct. Mater.* **24**, 5075–5085 (2014).
44. White, C. W., McHargue, C. J., Sklad, P. S., Boatner, L. A. & Farlow, G. C. Ion implantation and annealing of crystalline oxides. *Mater. Sci. Reports* **4**, 41–146 (1989).
45. Sathish, M., Viswanathan, B., Viswanath, R. P. & Gopinath, C. S. Synthesis, Characterization, Electronic Structure, and Photocatalytic Activity of Nitrogen-Doped {TiO₂} Nanocatalyst. *Chem. Mater.* **17**, 6349–6353 (2005).
46. George, S. *et al.* Role of Fe doping in tuning the band gap of TiO₂ for photo-oxidation induced cytotoxicity paradigm. *J. Am. Chem. Soc.* **133**, 11270–11278 (2011).
47. Shwetharani, R., Fernando, C. A. N. & Balakrishna, G. R. Excellent hydrogen evolution by a multi approach via structure-property tailoring of titania. *RSC Adv.* **5**, 39122–39130 (2015).
48. Moser, J., Grätzel, M. & Gallay, R. Inhibition of Electron-Hole Recombination in Substitutionally Doped Colloidal Semiconductor Crystallites. *Helv. Chim. Acta.* **70**, 1596–1604 (1987).

49. Umebayashi, T., Yamaki, T., Itoh, H. & Asai, K. Analysis of electronic structures of 3d transition metal-doped TiO₂ based on band calculations. *J. Phys. Chem. Solids* **63**, 1909–1920 (2002).
50. Yang, J. Y. *et al.* Grain size dependence of electrical and optical properties in Nb-doped anatase TiO₂. *Appl. Phys. Lett.* **95**, 213105 (2009).
51. Sanjinés, R. *et al.* Electronic structure of anatase TiO₂ oxide. *J. Appl. Phys.* **75**, 2945–2951 (1994).
52. Bert, I., Mohai, M., Sullivan, J. L. & Saied, S. O. Surface characterisation of plasma-nitrided an XPS study titanium: *Appl. Surf. Sci.* **84**, 357–371 (1995).
53. Wang, E., Yang, W. & Cao, Y. Unique Surface Chemical Species on Indium Doped TiO₂ and Their Effect on the Visible Light Photocatalytic Activity. *J. Phys. Chem. C* **113**, 20912–20917 (2009).
54. Xu, N. *et al.* Characteristics and mechanism of conduction/set process in TiN/ZnO/Pt resistance switching random-access memories. *Appl. Phys. Lett.* **92**, 35–38 (2008).
55. Hsieh, P. T., Chen, Y. C., Kao, K. S. & Wang, C. M. Luminescence mechanism of ZnO thin film investigated by XPS measurement. *Appl. Phys. A Mater. Sci. Process.* **90**, 317–321 (2008).
56. Mekki, A., Holland, D., McConville, C. F. & Salim, M. An XPS study of iron sodium silicate glass surfaces. *J. Non-Cryst. Solids* **208**, 267–276 (1996).
57. Kim, H. J., Kim, J. & Hong, B. Effect of hydrogen plasma treatment on nano-structured TiO₂ films for the enhanced performance of dye-sensitized solar cell. *Appl. Surf. Sci.* **274**, 171–175 (2013).
58. Fu, L. *et al.* Beaded Cobalt Oxide Nanoparticles along Carbon Nanotubes: Towards More Highly Integrated Electronic Devices. *Adv. Mater.* **17**, 217–221 (2005).
59. Tan, B. J., Klabunde, K. J. & Sherwood, P. M. A. XPS studies of solvated metal atom dispersed catalysts. Evidence for layered cobalt-manganese particles on alumina and silica. *J. Am. Chem. Soc.* **113**, 855–861 (1991).
60. Brik, Y., Kacimi, M., Ziyad, M. & Bozon-Verduraz, F. Titania-Supported Cobalt and Cobalt-Phosphorus Catalysts: Characterization and Performances in Ethane Oxidative Dehydrogenation. *J. Catal.* **202**, 118–128 (2001).
61. Radeka, M., Rekas, M. & Zakrzewcka, K. Electrical and optical properties of undoped and Fe doped TiO₂ single nanocrystal. *Solid State Phenom.* **39**, 113–116 (1994).
62. Zhang, J. *et al.* Synthesis, surface morphology, and photoluminescence properties of anatase iron-doped titanium dioxide nanocrystalline films. *Phys. Chem. Chem. Phys.* **13**, 13096–13105 (2011).
63. Simpson, J. R. *et al.* Optical band-edge shift of anatase Ti_{1-x}Co_xO_{2-δ}. *Phys. Rev. B* **69**, 193205 (2004).
64. Kong, L. G. *et al.* Oxygen-vacancies-related room-temperature ferromagnetism in polycrystalline bulk Co-doped TiO₂. *Electrochem. Solid State Lett.* **9**, G1–G3 (2006).
65. Anisimov, V. I. *et al.* The role of transition metal impurities and oxygen vacancies in the formation of ferromagnetism in Co-doped TiO₂. *J. Phys. Condens. Matter* **18**, 1695–1704 (2006).

Acknowledgements

This work was supported by research grant for Nanotechnology Lab of Jaypee University of Information Technology, also by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2015R1A2A1A05001826).

Author Contributions

B.B. fabricated and characterized the entire sample and wrote the manuscripts. S.K. carried out XPS studies of the samples. H.L. helped in the revision of the manuscript. R.K. supervised the work, reviewed and corrected the manuscript. All the authors participated in the discussion and commented on the paper.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Bharti, B. *et al.* Formation of oxygen vacancies and Ti³⁺ state in TiO₂ thin film and enhanced optical properties by air plasma treatment. *Sci. Rep.* **6**, 32355; doi: 10.1038/srep32355 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016

Spatially Concatenated Channel-Network Code for Underwater Wireless Sensor Networks

Zafar Iqbal and Heung-No Lee, *Senior Member, IEEE*

Abstract—Underwater environment monitoring is an important application of wireless sensor networks (WSNs). However, WSNs face challenges, such as erroneous communication, ensuring the lifetime and robustness of the network, and cost constraints. The underwater acoustic channel (UAC) is highly frequency-selective, and the channel response changes over time because of variations in the channel conditions. Therefore, designing a cooperative coded orthogonal frequency division multiplexing (COFDM) system that is suitable for the doubly selective UAC and has reduced power consumption is very challenging. We propose a cooperative spatial-domain coding scheme combined with the low-density parity-check-coded OFDM system, called spatially concatenated channel-network code, for underwater acoustic WSNs. The designed underwater acoustic WSN exhibits a significant advantage regarding the required number of sensors, bit error rate (BER), and power consumption over the non-cooperative COFDM communication system. We also analyze sensor deployment schemes and find out the area in which our proposed scheme can be beneficial in terms of reduced power consumption and enhanced BER.

Index Terms—Underwater acoustic communication, OFDM, LDPC, multipath fading, network coding, deployment, WSN.

I. INTRODUCTION

UNDERWATER acoustic communication has widespread applications, such as the monitoring of underwater environments, military/oceanic surveillance, underwater navigation, the observation of radiation leaks, and the exploration of underwater resources. Most of these applications require sophisticated underwater wireless sensor networks (WSNs), for which researchers have attempted to design reliable and robust underwater communication systems [1], [2].

The underwater acoustic channel (UAC) is time-varying because of the changes in the temperature, geometry of the channel, roughness of the sea surface, and spatial position caused by the sea current. Acoustic waves are considered as the major carrier in underwater communications because of their low attenuation characteristics [3], but the limited bandwidth and time-varying response of the UAC makes it difficult to obtain accurate channel state information (CSI)

at the transmitter and/or receiver. Furthermore, the multipath delay spread due to the reflections at the sea surface and bottom causes inter-symbol interference (ISI) and frequency-selective fading. Hence, these factors degrade the system performance [4], [5].

To overcome such performance falloffs, coded orthogonal frequency division multiplexing (COFDM) systems have been proposed that employ low-density parity-check (LDPC) codes [6]–[8], Reed-Solomon codes [9], and adaptive modulation and coding (AMC) [10] in OFDM systems for UACs. LDPC-coded and OFDM-based underwater acoustic communication has been well investigated in [6]–[8] and [11]–[13], respectively. However, we observe that a non-cooperative LDPC-COFDM system exhibits significant performance degradation in the presence of random fading. Also, because point-to-point systems are vulnerable to long-term deep fading, Doppler spread, and shadow zones [1], [14], [15], the interest in designing cooperative communication systems with network coding, has recently increased.

Therefore, we aim to develop a cooperative network communication scheme for underwater WSNs that resolves the aforementioned challenges of the UAC and reduces the power consumption in order to enhance the lifetime of the overall network. The envisioned network is a WSN wherein multiple sensors inside a shallow body of water cooperate while transmitting to a buoy on the water surface. To resolve the aforementioned problems, we propose a cooperation scheme that enables a group of transmitting sensors to form a network code over the spatial domain and is suitable for time- and frequency-selective UACs. The proposed cooperation scheme provides a considerable transmit power saving compared with the conventional non-cooperative scheme. Our study shows that the network coding benefit is sufficiently large to offset the increase in the power consumption due to the cooperation among the sensors in the network and yields an overall benefit of ~ 11 dB.

A. Related Works and Contribution of This Paper

After the idea of network coding was proposed in [16] and its application to linear network coding in [17], different strategies for cooperation, using network coding, have been proposed. These schemes may be classified based on the network types assumed, such as the single-source single-relay network [18]–[23], single-source multiple-relay network [19], [24]–[29], multiple-source single-relay network [30], [31], and multiple-source multiple-relay network [19], [32]–[35].

Manuscript received December 3, 2015; revised May 13, 2016; accepted July 12, 2016. Date of publication July 21, 2016; date of current version September 14, 2016. This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government Ministry of Science, ICT and Future Planning (MSIP) (NRF-2015R1A2A1A05001826). The associate editor coordinating the review of this paper and approving it for publication was A. Ghayeb. (*Corresponding author: Heung-No Lee.*)

The authors are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: zafar@gist.ac.kr; heungno@gist.ac.kr).

Digital Object Identifier 10.1109/TCOMM.2016.2593746

For multiple-source multiple-relay networks with a single destination where CSI is not readily available for relay assignment, adaptive network coded cooperation (ANCC) was proposed in [32]. In this scheme, each relay randomly selects a small number of correctly decoded messages from all the source nodes to generate a parity-check message in the cooperation phase. This leads to the formation of a graph code at the destination and a belief propagation decoding algorithm is used for decoding. However, for the decode-and-forward relaying scheme, detection errors at the relays should be taken into consideration for performance analysis and code design. While [32] assumes a set of relays that can successfully decode the received messages, [35] considers the possibility of unsuccessful decoding at the relays, making the scheme more realistic. In addition, [35] also assigns fixed relays to each source node and therefore, the relays do not have the overhead of sending an extra bit-map field to the destination to inform it about the underlying connections in the graph code.

More recently, [31] proposed a two-user and single-relay bilayer spatially-coupled LDPC (SC-LDPC) scheme for correlated sources. The system uses joint source-channel coding to transmit to the relay as well as to the destination. Correct decoding of the received signal is assumed at the relay and the relay then uses network coding to combine the received data before forwarding it to the destination. The scheme uses a factor-graph-based design of joint source-channel-network decoder at the destination when the sources are correlated. Also, an OFDM based dynamic coded cooperation (DCC) for underwater acoustic channels is presented in [23]. The relay listens until correct decoding of the received signal and then generates either an identical or a different OFDM block from the source and superimposes it on the transmission from the source in the cooperation phase. A delay control mechanism is used at the relay to achieve block-level synchronization between the source and the relay. This scheme requires a powerful relay node with abundant resources, such as a surface buoy, to assist the communication between the source and the destination.

Compared to the above-mentioned works, our design is based on multiple sources with multiple relays that transmit to a single destination. Because practical networks suffer from link failures and topology changes due to randomly fading channels, fixed relay assignment, as proposed in [35], is subject to failure in certain situations. Therefore, instead of the fixed relay assignment, we use random relay selection mechanism. In this scheme, a relay receives data from the neighboring source nodes. Some of these data are selected at random, encoded and transmitted in the relay phase. Our scheme of random relay selection thus, provides more robustness against link/node failures and outages in the underwater sensor network, without the need for a very powerful relay node as is the case in [23]. In our proposed cooperation mechanism, the relay randomly selects a small number of symbols from the data received from its neighbor nodes, without decoding it. It then re-encodes the symbols using an LDGM code, resulting in a concatenated channel-network code. For this channel-network code a joint iterative decoder

is designed and its performance is evaluated using extrinsic information transfer (EXIT) charts and BER simulations. The above-mentioned schemes either use repetition codes, convolutional codes, or some form of block codes in a distributed way, but in our scheme, each node independently encodes the data with an LDPC code and then in the cooperation phase, the nodes concatenate the received LDPC-coded symbols with an LDGM code in a distributed manner. Thus, our scheme combines the power of concatenated coding with adaptive network coding for underwater acoustic communication where CSI is not readily available. It also uses the random relay selection mechanism, resulting in a more practical cooperation scheme.

The contributions of this paper are summarized as follows:

- We consider a doubly-selective channel which was not considered in [32]. The work in [32] does not consider the effects of time- and frequency-selectivity, while our work takes care of time- and frequency-selectivity by using OFDM modulation and the scheme has been applied to underwater acoustic communication, for the first time. In the proposed scheme, the underwater acoustic sensors should take the role of relays for cooperation, but the sensors are limited in power, computational resources, and the challenge of underwater acoustic communication is great. Therefore, investigating the effectiveness of the cooperative coding scheme for underwater communication is very important and has not been addressed in the works discussed above.
- We have removed some unrealistic assumptions considered in [32]. The work in [32] considers network coding part only while assuming that a perfect channel coding has been performed. Our work removes this assumption as we have extended the network code by concatenating it with a channel code and included the effect of propagated error from the channel code to the network code part. The proposed scheme is termed as a spatially concatenated channel-network code (SCCNC) for underwater acoustic communication.
- Our relaying mechanism is different from [32] and other previous works. In the previous works, in the event of unsuccessful decoding, the relay either remains silent or sends its own data to the destination. In our proposed scheme, the relays do not need to decode the received codewords; they only detect the binary symbols. The relays then re-encode randomly selected symbols received from a number of sources and send it to the destination in the second phase. Therefore, the relays do not need to spend power on decoding the received codewords, thus saving time, energy, and hardware resources. This is very critical for underwater acoustic sensor networks, keeping in mind the limited power and computational resources of the sensors.
- In underwater acoustic communications, the sensor nodes require a particularly high power for the transmission and reception. Thus, the power consumption of the overall network is expected to increase, as each node must listen to the neighboring nodes' transmissions in order to realize the cooperation. In Section IV, we present an

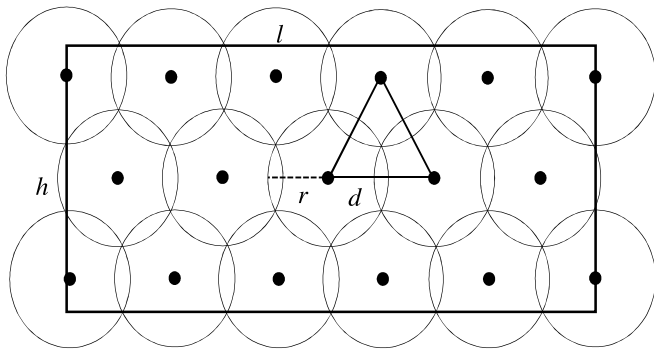


Fig. 1. 2D triangular grid deployment of sensors in an $l \times h$ area.

analysis indicating that the cooperation among sensor nodes significantly increases the power consumption of the network. Energy consumption analysis of both the cooperative and non-cooperative schemes is performed to observe the effects on the battery life of the sensor nodes.

- Random and grid-deployment schemes are considered, and the performance of these schemes is compared based on the BER and cost of network deployment and operation.

The remainder of this paper is organized as follows. We describe the network deployment issues in Section II. The proposed cooperative network coding scheme for underwater WSNs is presented in Section III. The performance analysis of the scheme is discussed in Section IV, and Section V concludes the paper. Some related information regarding the underwater channel is given in Appendixes A and B.

II. DEPLOYMENT OF SENSOR NODES

Sensor deployment is an important issue, especially in underwater WSNs, because the harsh underwater environments pose various challenges for the effective operation and robustness of the network. Sensor deployment addresses the problem of the coverage and connectivity of the network by targeting the minimized power consumption for a prolonged network lifetime.

The underwater WSN can be deployed in two types of communication architectures: two-dimensional (2D), where the sensors are deployed at the bottom of the sea, and three-dimensional (3D), where the sensors float at different depths to cover the entire volume of water [1]. Herein, we consider a static and 2D grid deployment for our WSN, which is relatively easy to deploy and operate. We use the k -coverage parameter to ensure that the target area is almost fully covered. A region is said to be k -covered if every point inside it falls within the sensing range r of at least k sensors. Our deployment target is to achieve l -coverage, as the underwater acoustic sensors are expensive devices, and we wish to minimize the power consumption and cost of operation.

The optimal deployment strategy to cover a 2D rectangular area using the minimum number of sensors involves placing each sensor at a vertex on a grid of equilateral triangles [36], as shown in Fig. 1. To obtain the full coverage, the coverage

ratio η (covered area/target area) should be 1, which can be achieved by adjusting the distance d among the sensors, such that $d = \sqrt{3}r$. This makes the uncovered areas shown in Fig. 1 zero, and the overlapping areas are minimized. Using [36, eq. (3)], we can compute the minimum number of sensors U required to cover a target area $l \times h$ to satisfy a given coverage ratio η as $U(l, h, d, r) = \lceil \frac{l-d}{d} + 1 \rceil \times \lceil \frac{2\sqrt{3}h-6d+4\sqrt{3}r}{3d} + 1 \rceil$. Thus, the minimum number of sensors necessary to provide l -coverage in an area of $100 \times 100 \text{ m}^2$ for $r = 20 \text{ m}$ is 12.

The next step is to estimate the number of redundant sensors required to ensure the robustness of the network to node failures within a pre-determined observation period. We assume that all the nodes have the same failure rate and that the node failures occur according to a Poisson distribution and are independent of each other. Therefore, the number of redundant sensors required to compensate for the Poisson-distributed node failures is given in [36, eq. (18)] as $\sum_{u=0}^{\Delta U} \frac{(\lambda T)^u e^{-\lambda T}}{u!} \geq \Gamma$, where λ is the sensor failure rate, T is the observation time in days, u is the number of sensors that may fail during the time T , and Γ is the probability that no more than ΔU failures occur in the observation time T . For example, with an average of one sensor failure every month ($\lambda = 1/(365/12)$) and a success probability of $\Gamma = 0.95$, there are approximately six sensor failures during a period of three months [36]. Thus, to ensure network connectivity and provide l -coverage in an area of $100 \times 100 \text{ m}^2$ for $r = 20 \text{ m}$ and an observation period of three months we must deploy 18 sensors rather than 12.

Finally, to ensure the connectivity of the network, we use the argument given in [37]: $\Theta(\log U)$ neighbors are necessary and sufficient for a sensor network to be asymptotically connected. This number is proven to be between $0.074 \log U$ and $5.1774 \log U$. Therefore, for a network of 12 or 18 nodes, we select the minimum required number of neighbors as 5.

Although the triangular-grid deployment appears to be a cost-effective solution regarding the number of sensors needed to provide the coverage and connectivity in a given area of interest, it may not be an effective solution for underwater area monitoring when cooperative communication is used to enhance the performance of the network. Moreover, compared with a randomly deployed network, a triangular-grid structure may be expensive to deploy and maintain for a long period of time in an underwater environment. Herein, we compare the effects of random and triangular-grid 2D static deployment strategies employing cooperation among sensor nodes that communicate to a buoy on the sea surface.

III. COOPERATIVE NETWORK-CODED COMMUNICATION

The point-to-point LDPC-COFDM communication system for the UAC has been thoroughly investigated [8], [11], [12], [38]. The results show that COFDM systems perform robustly in UACs designed with simplified channel conditions. Here, we show that a point-to-point COFDM system may encounter problems under the realistic fading conditions that exist in UACs. Moreover, the variations in the positions of the sensors and buoy can significantly change

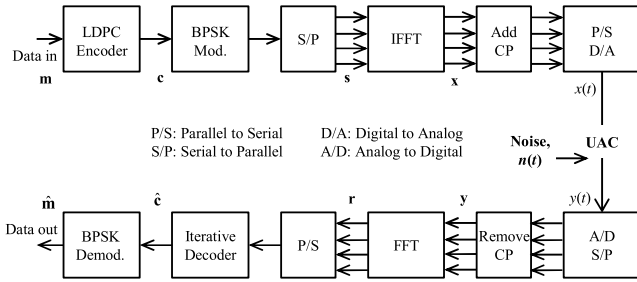


Fig. 2. LDPC-COFDM system block diagram.

TABLE I
OFDM SYSTEM PARAMETERS

Parameter	Value
Carrier frequency: f	7 kHz
Transmission bandwidth: BW	10 kHz
Maximum Doppler spread: $B\tau_{max}$	4.744 Hz
Coherent time: $T_C = 1/B\tau_{max}$	210 ms
Maximum delay spread: τ_{max}	25 ms
Coherent bandwidth: $B_C = 1/\tau_{max}$	40 Hz
Number of sub-carriers: N	256
Sub-carrier bandwidth: $\Delta f = BW/N$	39.0625 Hz
Valid symbol duration: $T_D = 1/\Delta f$	25.6 ms
CP period: $T_{CP} \geq \tau_{max}$	25 ms
OFDM symbol duration: $T_S = T_D + T_{CP}$	50.6 ms

the impulse response of the UAC, yielding a performance variation.

A. LDPC-COFDM System

Fig. 2 shows a block diagram of the suggested COFDM system employing the regular LDPC code [39]. The OFDM system parameters are summarized in Table I. The block size N is kept the same as the number of sub-carriers in the designed OFDM system. It is essential to choose a number of subcarriers that satisfies the conditions to overcome both frequency-selective fading ($\Delta f \leq B_C$) and time-selective fading ($T_S \ll T_C$). Because the Doppler spread increases geometrically as the carrier frequency increases [40], to overcome the time-selective fading, a suitable carrier frequency should be selected for the UAC. Based on our distance assumption of 1000 m, a bandwidth of 10 kHz is chosen, along with a carrier frequency of 7 kHz [4]. To overcome the ISI problem, the cyclic prefix (CP) period is set as 25 ms via an analysis of the impulse response of the modeled channel. Under this setting, the maximum delay spread and coherent time of the channel are approximately 25 ms and 210 ms, respectively [41]. The designed OFDM system can overcome not only frequency-selective fading, $\Delta f < B_C$, but also ISI, $T_{CP} \geq \tau_{max}$, as well as time-selective fading, $T_S \ll T_C$.

An OFDM block of size N is generated by splitting the incoming information into N subcarriers. Therefore, the input data sequence \mathbf{m} is encoded using a regular LDPC ($N = 256$,

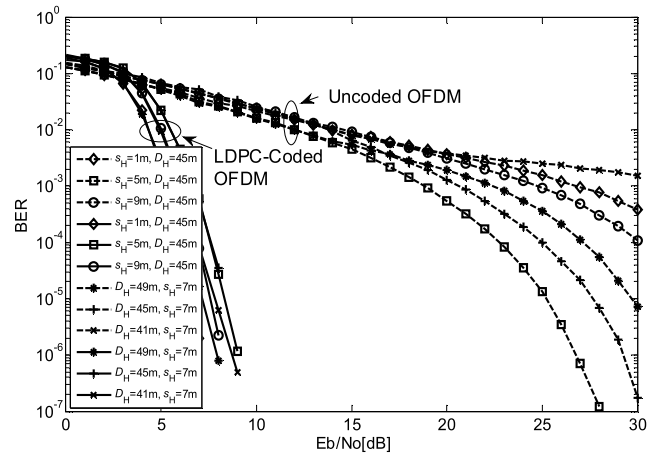


Fig. 3. Performance comparison between LDPC-COFDM and uncoded OFDM systems.

$j = 4, k = 8$) code-generator matrix \mathbf{G}_{LDPC} to generate $\mathbf{c} = \mathbf{m}\mathbf{G}_{LDPC}$, where $\mathbf{c} = [c_1, c_2, c_3, \dots, c_N]$, and the subscripts represent the k^{th} bit of the codeword mapped to the k^{th} subcarrier, i.e., $k = 1, 2, 3, \dots, N$. After the binary phase-shift keying (BPSK) modulation of \mathbf{c} , the resulting sequence, $\mathbf{s} = 2\mathbf{c} - 1$, is converted from serial to parallel form, where $\mathbf{s} = [s_1, s_2, s_3, \dots, s_N]^T$. Then, taking the inverse fast Fourier transform of \mathbf{s} yields $\mathbf{x} = \text{IFFT}_N\{\mathbf{s}\}$, which is transmitted through the UAC in the form of $x(t)$ after the CP is added and a digital-to-analog conversion is performed.

At the receiver, a discrete-time signal $\mathbf{y} = [y_1, y_2, y_3, \dots, y_N]^T$ is obtained by sampling the received signal $y(t)$ after removing the CP. This is then transformed into \mathbf{r} by taking its Fourier transform, i.e., $\mathbf{r} = \text{FFT}_N\{\mathbf{y}\}$, and represented as

$$\mathbf{r} = \sqrt{E_s}\mathbf{H}\mathbf{s} + \mathbf{n}, \quad (1)$$

where \mathbf{n} is an i.i.d. Gaussian noise vector; $\mathbf{n} \sim \mathcal{N}^{N \times 1}(0, \sigma^2)$, \mathbf{r} , \mathbf{s} , and \mathbf{n} are each an $N \times 1$ vector, E_s is the symbol energy, and \mathbf{H} is an $N \times N$ diagonal matrix whose diagonal entries are the transfer-function coefficients ($H_1, H_2, H_3, \dots, H_N$) of the UAC multiplied by the lognormal gain g , as discussed in Appendix A.

Let E_b represent the energy per bit in the transmitted codeword in joules, and N_0 is the noise power spectral density of the AWGN given in (1). To observe the performance of the COFDM system in UACs, we simulated the designed system by setting average random heights of the transmitting sensor (s_H) and receiver (D_H) from the sea bottom. The result is shown in Fig. 3, which compares the BER performance of the LDPC-COFDM system with that of an uncoded OFDM system. The COFDM system may overcome the severe frequency-selective performance falloff observed in the uncoded OFDM system, via the LDPC code. The coded system not only achieves a benefit of ~ 18 dB in E_b/N_0 but also reduces the performance variation due to the channel conditions. We observe a performance variation of ~ 3 dB when the positions of the sensor node and buoy change with respect to the sea bottom. This indicates the randomly changing nature of the UAC in shallow waters, which introduces the need for a

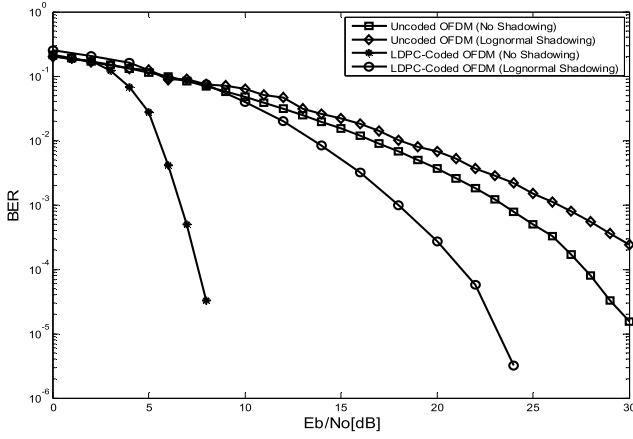


Fig. 4. Effects of lognormal shadowing on uncoded and coded OFDM systems.

more effective communication strategy that considers the time- and frequency-selective fading, along with the other factors described in Appendix A.

B. Cooperative Spatial-Domain Coding

We showed that the LDPC-COFDM system is suitable for time- and frequency-selective channels such as UACs, exhibiting a reasonably robust performance. However, as UACs suffer from long-term large-scale fading, we must observe the performance of the designed system under lognormal fading. Fig. 4 shows the performance of the LDPC-COFDM system under the lognormal shadowing channel model. The results show that although the LDPC code can mitigate the deep frequency-selective fading effect at certain specific subcarriers, it cannot effectively resolve the problem of large-scale fading, exhibiting a degradation of ~ 13 dB in E_b/N_0 at 10^{-4} BER. This effect is so detrimental that the sensors equipped with the LDPC-COFDM system spend on average ~ 13 dB more transmit power to obtain a BER of 10^{-4} than the amount needed with no shadowing.

User cooperation has been particularly beneficial for wireless systems that are subject to independent spatial fading. Thus, we are interested in the possibility of employing a user-cooperation scheme to resolve the detrimental effects of the fading in UACs. We propose the SCCNC scheme, as follows.

C. Design of the SCCNC Scheme

The famous two-phase user-cooperation scheme, which is common in wireless-network coding, [18]–[35], is utilized for our design of the underwater acoustic WSN. This approach is unique in that it aims to simultaneously exploit the diversity benefit from the frequency and spatial domains. The LDPC-coded and OFDM-modulated symbols transmitted by each sensor are relayed by the neighboring sensors, which helps to overcome the frequency-selective fading. Although our proposed system employs the idea of two-phase user cooperation reported by [32], in our scheme, the relays do not need to decode the received symbols, rather the symbols are used in the relay phase without regard to being correct or not.

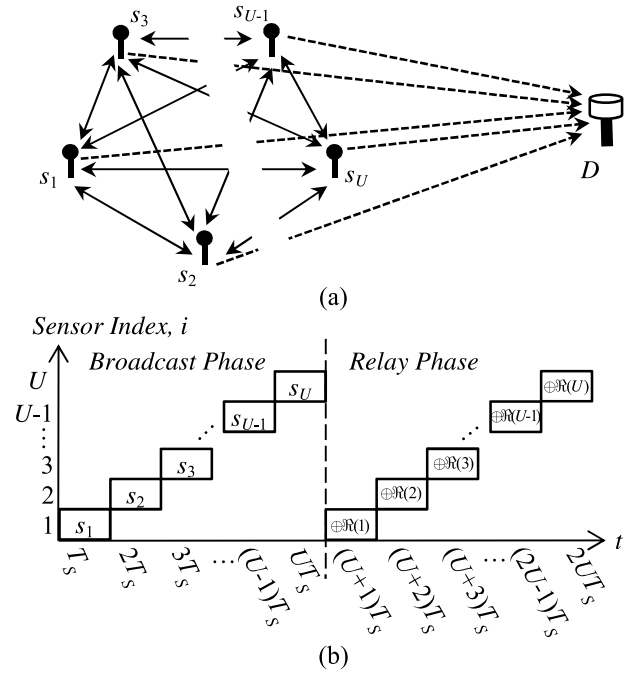


Fig. 5. (a) Spatial representation of the network-cooperation scenario; (b) transmission sequence and time slots for each sensor node.

In this scheme, the SCCNC is formed across the spatial and frequency domain. A joint iterative-decoding algorithm for this cooperative network code is then developed.

Fig. 5(a) depicts the assumed network-cooperation scenario. In this model, U nodes communicate wirelessly to a common destination D via two-phase user cooperation. In each phase, the U nodes transmit BPSK-modulated COFDM symbols using time division multiple access (TDMA). The solid lines in Fig. 5(a) represent the channels between the sensor nodes, and the dashed lines represent the channels between a sensor and the destination. Because of the changing channel conditions, some of the links shown here may be broken at a particular instant of time. The two-phase user cooperation strategy and the decoding algorithm are described as follows.

1) *Broadcast Phase*: Each sensor node transmits to the destination D an N -bit LDPC-COFDM symbol of duration T_S in its assigned time slot, as shown in Fig. 5(b). Let $\mathbf{r}_{1,i,D}$ be the received signal at the destination D , sent from the node i during the first phase. The received signal from the i^{th} node at the destination D is given as follows:

$$\mathbf{r}_{1,i,D} = \sqrt{E_{s1}} \mathbf{H}_{i,D} \mathbf{s}_{i,D} + \mathbf{n}_{i,D}, \quad (2)$$

where E_{s1} is the transmitted symbol energy in the first phase, the index i denotes the transmission from the i^{th} sensor node to the destination D , with $i = 1, 2, 3, \dots, U$. Because we use a TDMA transmission scheme, with the exception of the transmitting node, all of the $U - 1$ other nodes overhear the transmission, $x(t)$, and the received signal $\mathbf{z}_{i,j}$ at the node j is given as

$$\mathbf{z}_{i,j} = \sqrt{E_{s1}} \mathbf{H}_{i,j} \mathbf{s}_{i,j} + \mathbf{n}_{i,j} \quad j \neq i, \quad (3)$$

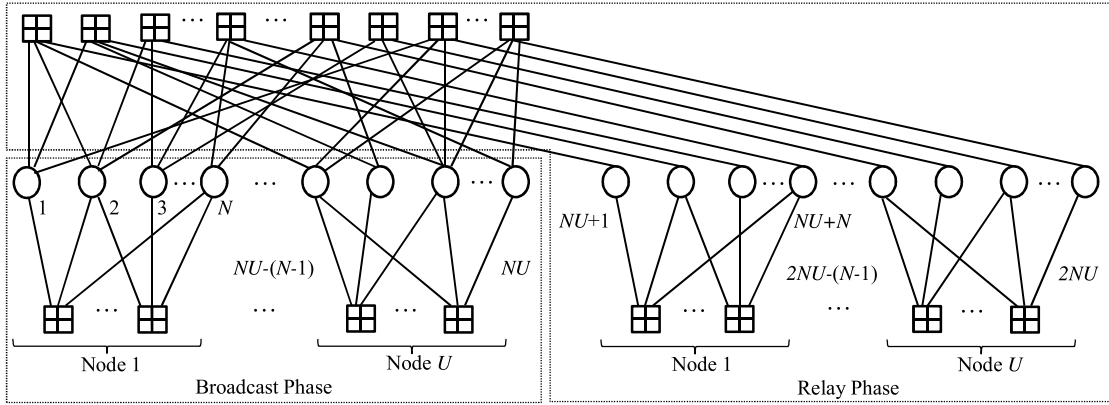


Fig. 6. Example of the SCCNC scheme: each sensor sends an LDPC codeword in the first phase. The spatial-domain checksums are computed and sent during the second phase, which are the LDPC coded symbols received during the first phase.

where j is the index of the receiving node, i is that of the transmitting node, and $j = 1, 2, 3, \dots, U$.

Because of the variation in the channel conditions, not all of the nodes can recover the transmitted codewords. We use a receive-set $\mathfrak{R}(j) \subseteq \{1, 2, \dots, U\}$, which stores the indices of the sensors whose transmissions are received at the node j , where U is the total number of cooperating sensor nodes. The expression $i \in \mathfrak{R}(j)$ indicates that node j has successfully received node i 's broadcasted symbol. Therefore, at the end of this phase, the destination node D has received $\{\mathbf{r}_{1,1}, \mathbf{r}_{1,2}, \dots, \mathbf{r}_{1,U}\}$ symbols, and each sensor node j in the cooperating group has received $\{\mathbf{z}_{1,j}, \mathbf{z}_{2,j}, \dots, \mathbf{z}_{U-1,j}\}$ symbols, as given by (2) and (3), respectively. Assuming that the switching time from one transmitting node to another is negligible, the time taken by U nodes to complete a broadcast phase is UT_S , where T_S is the OFDM symbol duration given in Table I. This phase is similar to that in the traditional COFDM communication system, except that the overhearing nodes in the cooperating group also store the recovered symbols for use in the relay phase. Note that the overhearing nodes do not decode the received symbols, but only store the received binary information.

2) *Relay Phase*: Each node randomly selects a small group of nodes from $\mathfrak{R}(j)$ (5 nodes), computes a checksum over their respective symbols, and forwards the checksum symbol $\oplus \mathfrak{R}(j)$, having length N , to the destination by using the same OFDM parameters in its assigned time slot, as shown in Fig. 5(b). Because the system operates using TDMA, the receive-set satisfies $\mathfrak{R}(j) \subseteq \{1, 2, \dots, U\}$. The spatial-domain code is formed using a code matrix similar to a randomly systematic low-density generator matrix (LDGM) code [42].

The codeword is formed using $\mathbf{G}_{\text{SCCNC}}$, which is the generator matrix for random-cooperation network coding, according to the procedure explained in Section III-A. Because of the random nature of the network code, a small bit field is included in the relay packet so that the destination node knows how the checksum was computed and can perform the message-passing decoding accordingly. Let $\mathbf{r}_{2,j,D}$ be the received SCCNC signal sent from the node j to the destination D during the relay phase. Then, the received signal at the destination D in the relay phase is given as

$$\mathbf{r}_{2,j,D} = \sqrt{E_{s2}} \mathbf{H}_{j,D} \mathbf{s}_{j,D} + \mathbf{n}_{j,D}, \quad (4)$$

where E_{s2} is the transmitted symbol energy in the relay phase, $\mathbf{s}_{j,D}$ is the SCCNC COFDM signal transmitted through the UAC from a sensor node j to the destination D , and $j = 1, 2, 3, \dots, U$. The source-symbols received in the first phase (2) constitute the systematic symbols of the network code, and the relay symbols received in the second phase (4) constitute the parity symbols. Hence, a set of U nodes completes the transmission of one SCCNC network codeword with length $2NU$ by the end of the second phase. The code rate at the destination is the combined code rate of the LDPC code and the network code, which is given as $R_{\text{SCCNC}} = R_{\text{LDPC}} \times R_{\text{LDGM}}$.

Assuming that the switching time from one transmitting node to another is negligible, the time taken by U nodes to complete a relay phase is UT_S . Therefore, the total time taken by U nodes to complete the transmission of an SCCNC symbol is $2UT_S$. The resulting SCCNC graph, as seen by the destination node, is shown in Fig. 6. The circles in Fig. 6 represent the bit nodes, and the squares represent the check nodes in the graph. The figure shows a U -node cooperation scheme, where each node uses a rate 1/4 SCCNC. The broadcast phase bit nodes shown in Fig. 6 represent $\mathbf{r}_{1,i,D}$, and the relay phase bit nodes represent $\mathbf{r}_{2,j,D}$ as defined in (2) and (4), respectively.

D. SCCNC Decoding Algorithm

We propose a joint message-passing decoding algorithm at the destination, whereby extrinsic information is exchanged between the channel code (LDPC) and the spatial-code (LDGM) decoders in every iteration. In this section, we consider imperfect inter-sensor channel condition and try to develop an algorithm incorporating the inter-sensor channel error. For random selection at the relays, probabilistically, each of the links has equal channel condition and the average error probability for a single link is given as

$$\bar{p} = \frac{1}{2} \left(1 - \sqrt{\frac{E_{s1}\gamma}{E_{s1}\gamma + 1}} \right), \quad (5)$$

where $\gamma = \mathbb{E} \left[\frac{g_{i,j}^2}{N_0} \right]$ with $g_{i,j}^2$ as the magnitude square of the lognormal fading coefficients. If each of the relay nodes

chooses L_{deg} of its neighboring nodes' information to form a parity checksum, the corresponding probability of error for each link can be computed as

$$p_e = \sum_{k=1, k \text{ is odd}}^{L_{\text{deg}}} \binom{L_{\text{deg}}}{k} \bar{p}^k (1 - \bar{p})^{L_{\text{deg}} - k} = \frac{1 - (1 - 2\bar{p})^{L_{\text{deg}}}}{2}. \quad (6)$$

The parity-check bits go through two serially concatenated channels; therefore, a modification is needed in the initialization part of the message-passing algorithm to incorporate the inter-sensor channel error in the decoding process. The channel log-likelihood ratio used to initialize the decoding iterations for the parity-check bits is given as

$$\text{LCr}_{j,D} = (1 - p_e) 4E_{s2} \mathbf{r}_{2,j,D} \left| \hat{\mathbf{H}}_{j,D} \right|^2 / N_0 + (p_e) 4E_{s2} \mathbf{r}_{2,j,D} \left| \hat{\mathbf{H}}_{j,D} \right|^2 / N_0. \quad (7)$$

The symbols received in the first phase go through one channel and the initialization for the decoding iterations is done as follows,

$$\text{LCr}_{i,D} = 4E_{s1} \mathbf{r}_{1,i,D} \left| \hat{\mathbf{H}}_{i,D} \right|^2 / N_0. \quad (8)$$

Let $\mathbf{R} = \{\mathbf{r}_{1,i,D}, \mathbf{r}_{2,j,D}\}$ be the received SCCNC signal matrix of size $N \times 2U$, E_s ($R_{\text{SCCNC}} \times E_b$) be the received symbol energy, $\hat{\mathbf{H}}$ is the received estimated channel transfer function, and N_0 is the normalized noise power. Let N_{rs} and N_{cs} , be the number of rows and number of columns, respectively, of the parity-check matrix of the spatial code S. Let N_{rl} and N_{cl} , be the number of rows and number of columns, respectively, of the parity-check matrix of the LDPC code L. We define the messages from the check nodes to the bit nodes of the spatial code and LDPC code as LSr and LLr, respectively. Similarly, the messages from the bit nodes to the check nodes of the spatial code and LDPC code are defined as LSq and LLq, respectively. The number of 1s in each row of the spatial-code parity-check matrix S, called the degree of the code, is S_{deg} , and the number of 1s in each row of the LDPC-code parity-check matrix L is called L_{deg} . Furthermore, we introduce a symmetric function $f(x) := -\log(\tanh(\frac{x}{2})) = \log\left[\frac{e^x + 1}{e^x - 1}\right]$ satisfying $f^{-1}(x) = f(x)$. We also define the functions $\text{find}(\cdot)$, which selects all the non-zero indices from a matrix and stores them into another matrix, and $\text{sgn}(\cdot)$, which selects the sign of the argument.

The SCCNC decoding algorithm, which performs joint iterative decoding over the network code at the destination, is described in Table II. LCr is the combined channel log-likelihood ratio for both phases, used to initialize the decoding iterations. The input data to this decoder is not a vector but a 2D matrix of size $N_{cl} \times N_{cs}$, as each cooperating node sends an LDPC-coded vector signal. The number of iterations (max_iter) can be set according to the desired decoding performance. During an iteration, the decoder calculates the bit-to-check node messages and then the check-to-bit node messages for all the nodes, first for the spatial code and then for the LDPC code as described in Table II.

TABLE II
NETWORK DECODING ALGORITHM

Step	Procedure
Input	$\mathbf{R}_{(N \times 2U)} = [\mathbf{r}_{1,u} : \mathbf{r}_{2,u}]$, E_s/N_0 $\hat{\mathbf{H}}_{(N \times 2U)} = [\hat{\mathbf{H}}_u : \hat{\mathbf{H}}_w]$ $v = 1, 2, \dots, N$, $u = 1, 2, \dots, U$, $w = U + 1, U + 2, \dots, 2U$
Initialize	$\text{LCr}_{v,u} = ((1 - p_e) 4E_{s2} \hat{\mathbf{H}}_{v,u}^2 / N_0) \mathbf{R}_{v,u} + (p_e) 4E_{s2} \hat{\mathbf{H}}_{v,u}^2 / N_0 \mathbf{R}_{v,u}$ $\text{LCr}_{v,w} = (4E_{s1} \hat{\mathbf{H}}_{v,w}^2 / N_0) \mathbf{R}_{v,w}$ $\text{Lp}_{v,p} = [0]$, $p = 1, 2, \dots, 2U$ $\text{LSr}(N_{cs}, N_{rs}, N_{cl}) = [0]$, $\text{LSq}(N_{cs}, N_{rs}, N_{cl}) = [0]$ $\text{LLr}(N_{cl}, N_{rl}, N_{cs}) = [0]$, $\text{LLq}(N_{cl}, N_{rl}, N_{cs}) = [0]$
Iterations	While ($\text{num_iter} < \text{max_iter}$) (i) Calculate the spatial code bit-to-check node messages. Loop: $l = 1$ to N_{cl} , $m = 1$ to N_{cs} $\text{Sc} = \text{find}(S_{\text{all rows}, m})$; $\text{Lc} = \text{find}(L_{\text{all rows}, l})$ $\text{LSq}_{m, \text{Sc}, l} = \text{LCr}_{l,m} + \sum_{m' \neq m, \text{Sc}' \neq \text{Sc}, l' \neq l} \text{LSr}_{m', \text{Sc}', l'} + \sum \text{LLr}_{l, \text{Lc}, m}$ End Loop (ii) Calculate the spatial code check-to-bit node messages. Loop: $l = 1$ to N_{cl} , $m = 1$ to N_{rs} $\text{Sc} = \text{find}(S_{m, \text{all columns}})$; $\text{LSr}_{\text{Sc}, m, l} = \prod_{\text{Sc}' \neq \text{Sc}, m' \neq m, l' \neq l} \text{sgn}(\text{LSq}_{\text{Sc}', m', l'})$ $\times f\left(\sum_{\text{Sc}' \neq \text{Sc}, m' \neq m, l' \neq l} f(\text{LSq}_{\text{Sc}', m', l'})\right) (-1)^{S_{\text{Sc}} + 1}$ End Loop (iii) Calculate the LDPC code bit-to-check node messages. Loop: $l = 1$ to N_{cl} , $m = 1$ to N_{cs} $\text{Sc} = \text{find}(S_{\text{all rows}, m})$; $\text{Lc} = \text{find}(L_{\text{all rows}, l})$; $\text{LLq}_{l, \text{Lc}, m} = \text{LCr}_{l,m} + \sum_{l' \neq l, \text{Lc}' \neq \text{Lc}, m' \neq m} \text{LLr}_{l', \text{Lc}', m'} + \sum \text{LSr}_{m, \text{Sc}, l}$ End Loop (iv) Calculate the LDPC code check-to-bit node messages. Loop: $l = 1$ to N_{cs} , $m = 1$ to N_{rl} $\text{Lc} = \text{find}(L_{m, \text{all columns}})$; $\text{LLr}_{\text{Lc}, m, l} = \prod_{\text{Lc}' \neq \text{Lc}, m' \neq m, l' \neq l} \text{sgn}(\text{LLq}_{\text{Lc}', m', l'})$ $\times f\left(\sum_{\text{Lc}' \neq \text{Lc}, m' \neq m, l' \neq l} f(\text{LLq}_{\text{Lc}', m', l'})\right) (-1)^{L_{\text{deg}}}$ End Loop End While
Result	Calculate the output value at the bit nodes. Loop: $l = 1$ to N_{cl} , $m = 1$ to N_{cs} $\text{Sc} = \text{find}(S_{\text{all rows}, m})$; $\text{Lc} = \text{find}(L_{\text{all rows}, l})$; $\text{Lp}_{l,m} = \text{LCr}_{l,m} + \sum \text{LSr}_{m, \text{Sc}, l} + \sum \text{LLr}_{l, \text{Lc}, m}$ End Loop
Decision	Make a decision based on the values of Lp calculated in the previous step. If $\text{Lp}_{v,p} > 0$, $\hat{c}_{v,p} = 1$; else $\hat{c}_{v,p} = 0$
Output	SCCNC codeword $\hat{\mathbf{c}}$ of size $N \times 2U$.

Finally, the output values are calculated at each node, and a decision of 0 or 1 is made to obtain the SCCNC codeword. The codeword can then be decoded using the corresponding parity-check matrices of $\mathbf{G}_{\text{SCCNC}}$ and then \mathbf{G}_{LDPC} to obtain the message received by each node in the cooperating group.

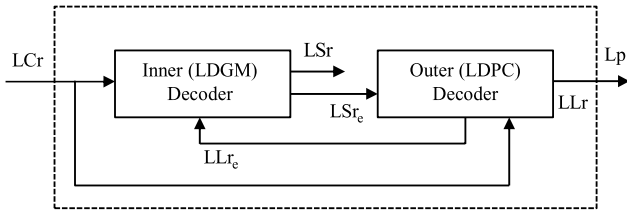


Fig. 7. Iterative decoding procedure for the proposed SCCNC scheme.

1) *EXIT Chart of the SCCNC Decoder*: To determine the characteristics and verify the performance of the joint iterative decoder for the SCCNC, an EXIT chart was used. EXIT charts are used to quantify the extrinsic information exchanged between the constituent decoders in an iterative decoding scheme. The EXIT chart plots two curves, showing the mutual information of the extrinsic log-likelihood ratios with respect to the mutual information of the *a priori* log-likelihood ratios, one for each decoder.

Fig. 7 shows the SCCNC decoding procedure using the LDGM decoder as the inner decoder and the LDPC decoder as the outer decoder. The *a priori* information about the source bits is not shown in the diagram because it is considered to be zero for equiprobable source bits. The LCr, given as $LCr = \{LCr_{j,D}, LCr_{i,D}\}$, represents the channel log-likelihood ratios, and LSe and LLr represent the extrinsic information output from the inner and outer decoders, respectively:

$$\begin{aligned} LSe &= LSr - LCr \\ LLr &= LLr - (LCr + LSe). \end{aligned} \quad (9)$$

The extrinsic information from the inner decoder LSe is used as an *a priori* input to the outer decoder to determine LLr. The new LLr is then used as an *a priori* input to the outer decoder in the next iteration.

The *a priori* input A to a constituent decoder is modeled using an independent Gaussian random variable n_A with a mean of zero and a variance of σ_A^2 . It is given as follows:

$$A = \mu_A \cdot m \cdot H_A + n_A, \quad (10)$$

where $\mu_A = \sigma_A^2/2$ is the mean of the Gaussian-distributed log-likelihood ratios of A , m is the transmitted systematic bit, and H_A is the corresponding frequency-response coefficient of the fading channel.

With the equiprobable source symbols input to the encoder at the transmitter, the bitwise mutual information content of the *a priori* information $I_A = I(M; A)$ and the extrinsic information $I_E = I(M; E)$ are calculated as follows:

$$\begin{aligned} I_A &= \frac{1}{2} \sum_{m=0}^1 \int_{-\infty}^{\infty} p_A(\xi | M = m) \\ &\quad \times \log_2 \frac{2p_A(\xi | M = m)}{p_A(\xi | M = 0) + p_A(\xi | M = 1)} d\xi, \end{aligned} \quad (11)$$

$$\begin{aligned} I_E &= \frac{1}{2} \sum_{m=0}^1 \int_{-\infty}^{\infty} p_E(\xi | M = m) \\ &\quad \times \log_2 \frac{2p_E(\xi | M = m)}{p_E(\xi | M = 0) + p_E(\xi | M = 1)} d\xi, \end{aligned} \quad (12)$$

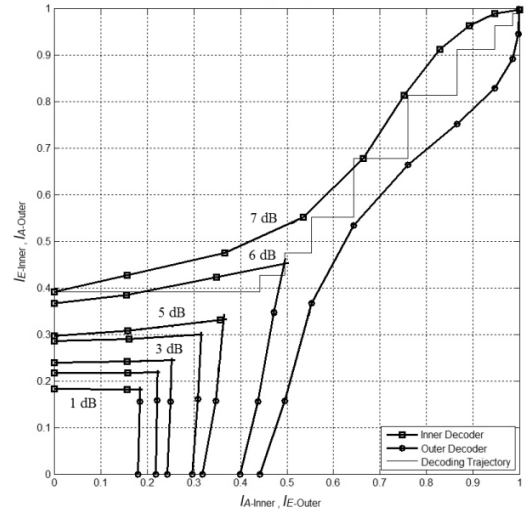


Fig. 8. EXIT chart for the SCCNC decoder for a UAC with SNRs ranging from 1–7 dB.

where M is a random variable representing the bits m of the input symbol \mathbf{m} ; and p_A and p_E are the conditional probability distributions for the *a priori* information and extrinsic information of each decoder, respectively, and are obtained by simulations using histogram measurements. Details about the EXIT-chart procedure and analysis are beyond the scope of this paper; the reader is referred to [43] for more information. Fig. 8 shows the EXIT chart for our proposed SCCNC decoder for a network of randomly deployed 12 nodes. The bitwise mutual information is averaged over the symbols received from all the sensors. We show the proposed decoder's EXIT characteristics for a range of SNRs (1–7 dB) for the UAC. It is observed that the decoder converges at an SNR of ~ 7 dB for the UAC. The decoding trajectory shows that at least 10 iterations are needed for the decoder to converge. The convergence point is also verified by the simulation results in Fig. 11 and Fig. 12, which show a waterfall region near the SNR of 7 dB for a network of randomly deployed 12 nodes. The degradation in the performance of the decoder, compared with that in [43], arises from the harshness of the UAC.

IV. PERFORMANCE ANALYSIS

In this section, we aim to analyze the energy consumption and network coding benefits of the designed network code. The coding gain obtained in the case of channel coding is obvious and well-understood, but in the case of the designed network code, we must consider other factors, such as the energy spent by sensors for receiving and decoding the overheard transmitted symbols, sending the parity check bits, and decoding the network-coded received signal at the destination. We aim to determine whether the network coding gain is sufficiently large to offset the increase in the power consumption for cooperative transmission and network (de)coding operations in the proposed scheme.

A. Energy Consumption of Coop. and Non-Coop. Schemes

With the current technology, an underwater acoustic modem uses an approximate transmit power of 2 W, a receiving

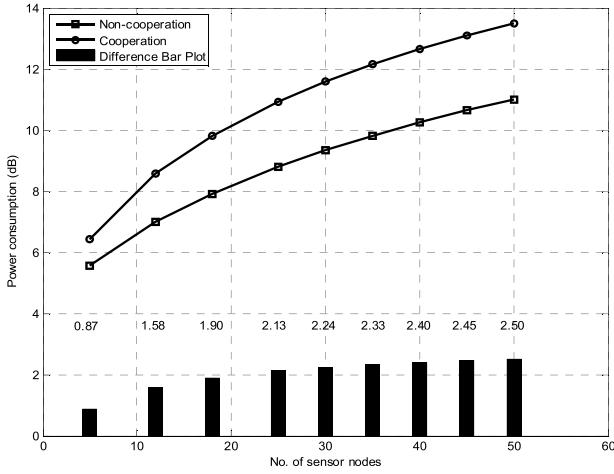


Fig. 9. Power-consumption comparison between cooperation and non-cooperation networks.

power of 0.8 W, and an idle listening power of 0.2 W for communication over a distance of 1000 m [44], [45]. The message-passing decoder power dissipation is shown to be on the order of 500 mW for a throughput of 1 Gbps [46]–[48]. In the proposed cooperation scheme, the power consumption of the message-passing decoder increases by a factor greater than 2 as the length of the codeword doubles. Because the data rate of our proposed scheme is very low, for a throughput of 1 Mbps, we can safely assume the decoding power dissipation to be ~ 0.5 mW in the case of non-cooperation and 1 mW in case of cooperation.

Let E_t , E_r , E_i , E_{dnc} , and E_{dc} , denote the energy consumed by the acoustic modem during the transmit operation by the sensors, receive operation at the sensor/buoy, idle listening by the sensors (no transmit/receive operation), decoding operations at the buoy in the non-cooperative case, and decoding operations at the buoy in the cooperative case, respectively. In the case of non-cooperation, the total energy consumed during one symbol period by the network of U nodes, $E_{s(non-coop)}$, is the sum of the following: the energy of a single transmission by U nodes, $U - 1$ multiplied by the idle listening energy of each node, the energy required for the receive operation, and the decoding energy consumption for U nodes at the destination D . It is given as

$$E_{s(non-coop)} = U (E_t + (U - 1) E_i + E_r + E_{dnc}). \quad (13)$$

In the case of node cooperation, the total energy consumed during one symbol period by the network of U nodes, $E_{s(coop)}$, is the sum of the following: twice the energy of transmission for U nodes, $U - 1$ multiplied by the energies for the reception operations of each node, twice the energy of the receive operation, and the energy of the decoding operations for U nodes in the cooperative case at the destination D . It is given as

$$E_{s(coop)} = U (2E_t + (U - 1) (E_r) + 2E_r + E_{dc}). \quad (14)$$

Using the aforementioned values for E_t , E_r , E_i , E_{dnc} , and E_{dc} , the corresponding power consumption for Eqs. (13) and (14) is plotted in Fig. 9 for a varying number of sensor nodes in the network. The results indicate increases of

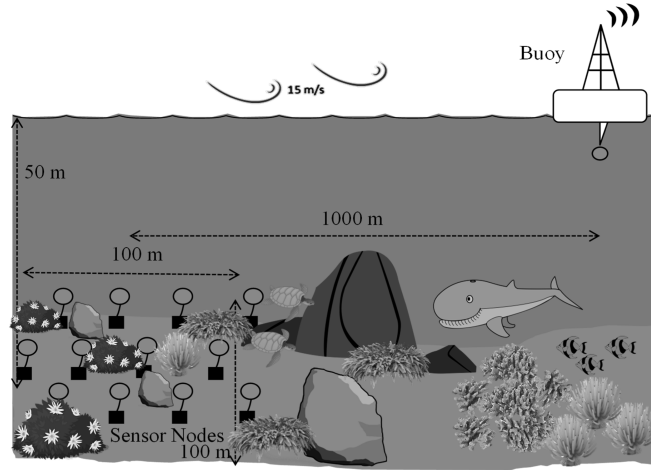


Fig. 10. Underwater WSN scenario (Not to scale).

approximately 1.58, 1.9, and 2.5 dB in the power consumption for $U = 12, 18,$ and 50 , respectively, in the cooperative network. The increase in the power consumption converges to ~ 3 dB for a cooperative network having up to 1000 nodes (not shown here). This shows that for cooperation among a reasonable number of nodes, i.e., $U < 50$, the increase in the power consumption is less than 3 dB.

B. Network Coding Gain

To analyze the BER, we assume that the underwater sensor nodes are distributed at the sea bottom as shown in Fig. 10. The numbers of sensor nodes considered are 12 and 18, according to the calculations done in Section II. The nodes are placed at an average height of 7 m from the sea bottom within the 100×100 m² range and the buoy is placed 5 m below the sea surface. In the case of random deployment, the position of each node is generated randomly uniform for every OFDM symbol transmission, as well as time-varying channel responses between the nodes and buoy. Similarly, for the case of triangular-grid deployment, the position of each node is generated in the form of a triangular grid. Other factors affecting the channel are a maximum sea-surface wind speed of 15 m/s, water depth of 50 m (considering the 44-m average depth of the Korean Western Sea), and distance of 1000 m between the node and buoy. Each node has a transmission range of 1000 m and a data rate of 2.5 kbps. The data-packet size is set as 32 bytes. Each node sends 1 packet of data in the broadcast phase and 1 packet of data in the relay phase towards the buoy.

Fig. 11 shows the performance of the proposed SCCNC scheme for the UAC with the lognormal shadowing model, using random deployment. Here, we simulate two different scenarios, one with a perfect inter-sensor channel (ISC) and another with a realistic underwater channel including the errors induced by the inter-sensor communication in the first phase. The SNR for realistic scenario is chosen to be 10 dB higher than that at the destination, based on the argument given in Appendix B. As shown, the proposed SCCNC scheme exhibits a significant improvement compared with the LDPC-COFDM

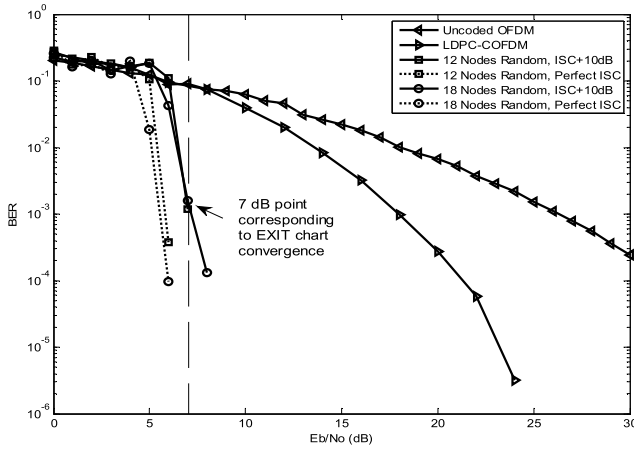


Fig. 11. Performance of the proposed SCCNC scheme compared with the LDPC-COFDM and uncoded OFDM schemes for the UAC.

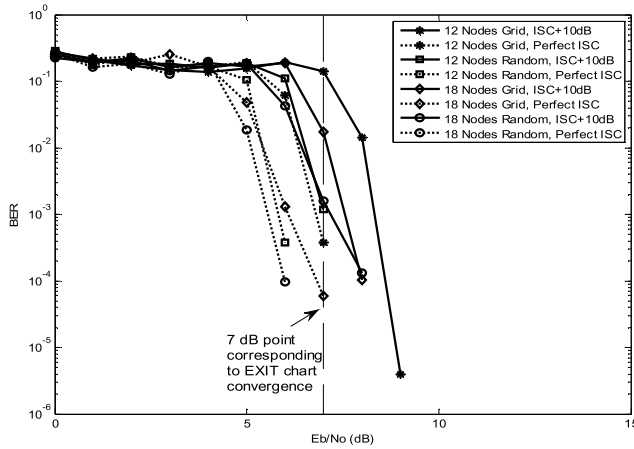


Fig. 12. Performance comparison of random and grid deployment by using the proposed SCCNC scheme for the UAC.

system. For example, at the point where the BER is 10^{-4} , with 18-node cooperation, we obtain a 13-dB benefit compared with the LDPC-COFDM system. We also observe an improvement of ~ 11 dB at the point where the BER is 10^{-3} for a network comprising as few as 12 randomly deployed nodes.

When we compare the performance of the proposed scheme for a realistic channel with perfect ISC, a degradation of ~ 1.5 dB is observed in both the random and grid deployment (Fig. 12) for 12 and 18 nodes cooperation, which is negligible compared to the huge network coding gain of 13 dB. The perfect ISC assumption is equivalent to the scheme proposed in [32] combined with OFDM transmission, as all the symbols are assumed to be correctly received at the relays. Consequently, the codewords formed at the relays contain the information from correctly received symbols. Therefore, we can deduce that ideally, [32] will perform similar to the dotted lines shown in Fig. 11 and Fig. 12 on the underwater acoustic channel. However, in [32], the relay needs to decode the received symbol and decide whether it was correctly received or not, therefore, it spends more power and the hardware is more complex as compared to our proposed scheme. Our results show that without using this complex hardware and

spending more power, we can achieve a similar performance by concatenating the channel and network codes.

The proposed SCCNC scheme for underwater acoustic communication benefits from the spatial diversity offered by the network, along with the frequency-diversity benefit, which is exploited by the LDPC-coded modulation with the OFDM transmission. Considering the additional 1.9 dB of power consumed by the cooperative network (Fig. 9), the designed cooperation scheme saves ~ 11 dB of the transmit-power consumption over the non-cooperative LDPC-COFDM system for a network comprising as few as 18 cooperating sensor nodes deployed within a 100×100 m² area. Using the received SNR curves obtained in Appendix B, the deployment area where our proposed scheme can be beneficial might extend up to 500×500 m², intuitively, which can be exactly determined in a future work.

C. Comparison of Random and Grid Deployment

Fig. 12 shows the BER performance of the random and fixed triangular-grid deployment of sensor nodes over an area of 100×100 m². The grid deployments of both 12 and 18 nodes exhibit slightly higher BERs than the random deployment.

Therefore, we conclude that the random deployment is preferred over the triangular-grid deployment for an underwater acoustic WSN, as the random deployment is easier and cheaper to deploy and maintain over a period of time. Moreover, it exhibits a slightly better performance than the triangular-grid deployment with regard to the BER.

D. Delay and Extended Battery Life

In the case of the non-cooperative network, the throughput of the message-passing decoder at the destination D is $\frac{NU}{UT_s}$ bps (~ 6 kbps), whereas in the case of cooperation, it is $\frac{NU}{2UT_s}$ bps (~ 3 kbps). Thus, the throughput in the case of cooperation is reduced by half, which is expected because the destination must receive all the parity-check symbols from the cooperating sensor nodes before it can start decoding.

We wish to compute the effect on the battery life of the sensor in our proposed scheme. For a sensor battery life of h hours, the total power consumed by a network of U nodes is

$$P_{non-coop} = \frac{E_{s(non-coop)}}{h} = \mathcal{P} \text{ (dB)}, \quad (15)$$

and that for a cooperative network is

$$P_{coop} = \frac{E_{s(coop)}}{h} \cong (\mathcal{P} + 1.9 - 13) \cong (\mathcal{P} - 11) \text{ (dB)}. \quad (16)$$

Eq. (16) incorporates the 1.9-dB increase in the power consumption and the 13-dB network coding gain in our proposed scheme, showing that the scheme consumes ~ 13 times less power than the non-cooperation scheme as given in (15). However, because the time required to transmit the same amount of data is now $2h$ hours, the battery life improves by a factor

of ~ 6.5 overall, increasing to $6.5h$ hours. Because the battery life is a very important factor in the operation/maintenance of underwater WSNs and the delay is not critically important, our proposed scheme is a very good option for low-energy and improved-BER underwater communication networks.

V. CONCLUSION

We discussed the design of a network coding scheme for underwater acoustic communication and networking systems. We found that the non-cooperative LDPC-COFDM communication system mitigates deep frequency-selective fading effects but cannot effectively resolve the problem of shadowing in the UAC. On the other hand, cooperative communication enhanced the BER but significantly increased the power consumption of the network. To solve these problems, we propose a user-cooperation-based network coding scheme called the SCCNC. This scheme is applied to both randomly deployed and triangular-grid networks to facilitate cooperation among the sensors. It greatly enhanced the BER of the network, improving the SNR by ~ 11 dB overall, consuming ~ 13 times less power, and increasing the battery life by a factor of 6.5 compared with the non-cooperative point-to-point LDPC-COFDM system. This benefit can be obtained when the cooperating sensor nodes are deployed within a 100×100 m² area. Our results also show that a random deployment of the underwater acoustic WSN is superior to a triangular-grid deployment with regard to the BER and the deployment cost.

APPENDIX A CHANNEL MODEL

Here, we explain the channel model used for the simulations. We use the geometrical ray-tracing model [5], [38], [49]–[51], to investigate the underwater sound propagation and aim to describe the modeling procedure step-by-step, along with the channel characteristics.

In a UAC, the acoustic waves are reflected at the sea surface and bottom and form a multipath, as shown in Fig. 13(a) [5], [14]. The reflection paths are classified into four types according to the total number of reflections (odd or even) and the first reflection point (surface or bottom). The channel transfer function is a superposition of the transfer functions of each propagation path from the transmitter to the receiver. It is given as

$$H(f, t) = \sum_p H_p(f, t) e^{-j2\pi f \tau_p(t)}, \quad (17)$$

where $H_p(f, t)$ and $\tau_p(t)$ represent the transfer function of the p^{th} path at frequency f and the corresponding delay at time t , respectively. The transfer function of each reflection path is represented as a function of the frequency, number of reflections, and path length. The transfer function of the p^{th} path is given as

$$H_p(f, t) = \frac{V_p}{\sqrt{(C(L_p(t), f))}}, \quad (18)$$

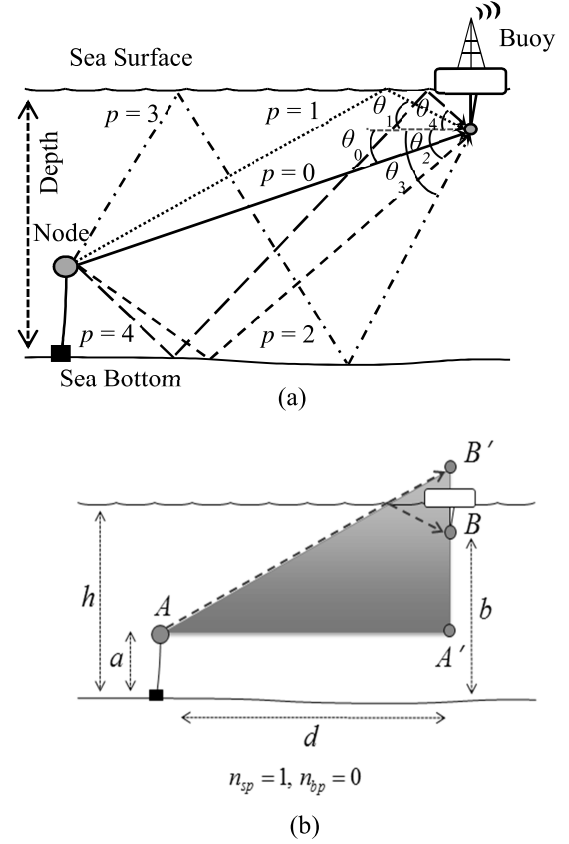


Fig. 13. (a) Geometrical representation of the multipath propagation in the UAC. (b) Example of calculating the reflection-path distance.

where $V_p = v_s^{n_{sp}} v_b^{n_{bp}} (\theta_p)$ is the reflection coefficient, which is the number of times a ray is reflected from the sea surface (n_{sp}) and bottom (n_{bp}), where v_s and v_b are the reflection coefficients at the sea surface and bottom, respectively [15].

Because the single-path loss is a function of the carrier frequency and path length, it is necessary to calculate the length of the reflection path. Similarly, the grazing angle θ_p is an essential factor for calculating the reflection coefficient. We illustrate the proposed method using Fig. 13(b). To calculate the length of the reflection path from A to B, (i) move B to B' against the sea surface; (ii) calculate the length of the baseline d ; (iii) calculate the height of the triangle, which is given by $2h - a - b$, as the distance from the sea surface to point A' is $h - a$ and the distance from the sea surface to B' is $h - b$; and (iv) calculate the distance using the Pythagorean Theorem: $L_p^2 = d^2 + (2h - a - b)^2$. This approach is used to obtain a general equation for the length of the reflection path, which is given as follows:

$$L_p = \sqrt{d^2 + (2h \cdot n_{sp} + \alpha a + \beta b)^2}, \quad (19)$$

where α and β are classification values in accordance with the first reflection point (surface or bottom) and the total number of reflections (odd or even). Specifically, $(\alpha, \beta) = (-1, -1), (+1, +1), (-1, +1)$, and $(+1, -1)$ for the paths having the first reflection on the surface with an odd number of reflections ($p = 1$), the first reflection on the bottom with an odd number of reflections ($p = 2$), the first reflection on

the surface with an even number of reflections ($p = 3$); and the first reflection on the bottom with an even number of reflections ($p = 4$), respectively, as shown in Fig. 13(a). The grazing angle can then be calculated as $\theta_p = \cos^{-1}(d/L_p)$.

In (19), the single-path loss with the distance L_p [m] and carrier frequency f [Hz] is $C(L_p(t), f) = C_0 L_p^\psi(t) \chi(f)^{L_p(t)}$, where C_0 is a constant scaling factor, and ψ is the spreading factor, which ranges between 1 and 2, according to the type of spreading. We set C_0 as 1 and ψ as 1.5, considering practical spreading. $\chi(f)$ is the absorption coefficient, expressed in dB/km, which is defined by Thorp's empirical formula at frequencies above a few hundred Hz as $\chi(f) = \frac{0.11f^2}{1+f^2} + \frac{40f^2}{4100+f^2} + 2.75 \times 10^{-4}f^2 + 0.003$ [15]. The acoustic path loss is then expressed in dB as $10 \log C(L_p(t), f)/C_0 = \psi \cdot 10 \log L_p(t) + L_p(t) \cdot 10 \log \chi(f)$.

We assume that the length of the p^{th} propagation path is

$$L_p(t) = \bar{L}_p + \Delta L_p(t), \quad (20)$$

where \bar{L}_p is the nominal length, and $\Delta L_p(t)$ is the variation in the length $L_p(t)$. The nominal path transfer function for the reference path ($p = 0$) can be written as

$$Q(f) = \frac{1}{\sqrt{C(\bar{L}_0, f)}}. \quad (21)$$

Therefore,

$$H_p(f, t) = \frac{V_p}{\sqrt{\left(\frac{L_p(t)}{\bar{L}_0}\right)^\psi \chi(f)^{L_p(t) - \bar{L}_0}}} Q(f). \quad (22)$$

According to the analysis presented in [52], Eq. (22) is approximated as $H_p(f, t) \approx h_p(t) \cdot Q(f)$, and the path gain is expressed as follows:

$$h_p(t) \approx \bar{h}_p e^{-\zeta_p \Delta L_p(t)/2}, \quad (23)$$

where $\bar{h}_p = \frac{V_p}{\sqrt{\left(\frac{L_p}{\bar{L}_0}\right)^\psi \chi_0^{L_p - \bar{L}_0}}$, $\chi_0 \approx 1$, and $\zeta_p = \chi_0 - 1 + \psi/\bar{L}_p$.

The overall transfer function for the UAC is thus given as

$$H(f, t) = Q(f) \cdot \sum_p h_p(t) e^{-j2\pi f \tau_p(t)}, \quad (24)$$

and taking the inverse Fourier transform of (24), we obtain the following channel impulse response:

$$h(\tau, t) = \sum_p h_p(t) q(\tau - \tau_p(t)). \quad (25)$$

The presence of large rocks, coral reefs, and uneven surfaces causes signal fading in UACs. The signal-strength fading or gain $g(t)$ is a random process in UACs that has been approximated using numerous distribution models, including Ricean, Rayleigh, and lognormal distributions [52]–[54]. We use the lognormal distribution [55] to model the fading effects and thereby make our channel model more realistic, as this distribution is well-known to yield a good fit for the long-term, large-scale fading phenomenon in UACs for shallow water [52]–[54]. The channel gain from a sensor to the buoy

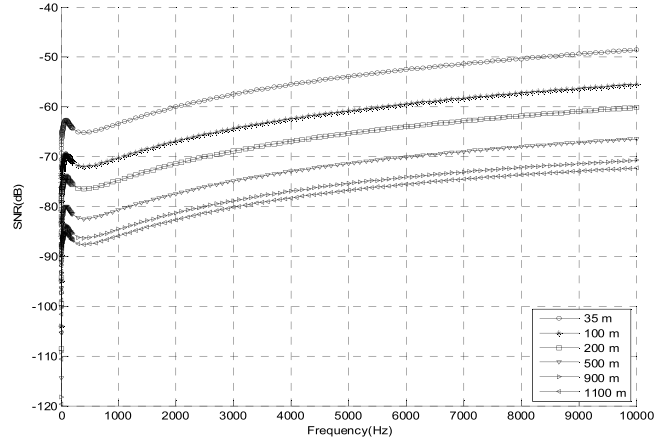


Fig. 14. Received SNR at varying distance from the transmitter.

is modeled as $g(t) \sim \ln \mathcal{N}(\mu, \sigma^2)$, with a mean of 1 and variance of 2, and used to include the fading effect. Here, $g(t)$ is assumed to be independent from one sensor to another. For simplicity, it is assumed to be fixed during each OFDM symbol transmission from a sensor i .

APPENDIX B AMBIENT NOISE AND RECEIVED SNR

The noise in underwater communication is classified as ambient noise and site-specific noise. Site-specific noise exists only in certain areas while ambient noise is always present and can be modelled as Gaussian. It consists of four major factors including turbulence, shipping, waves, and thermal noise. The power spectral density (PSD) of the ambient noise is given as follows,

$$\begin{aligned} 10 \log N_t(f) &= 17 - 30 \log f \\ 10 \log N_s(f) &= 40 + 20(s - 0.5) + 26 \log f \\ &\quad - 60 \log(f + 0.03) \\ 10 \log N_w(f) &= 50 + 7.5w^{1/2} + 20 \log f \\ &\quad - 40 \log(f + 0.4) \\ 10 \log N_{th}(f) &= -15 + 20 \log f \end{aligned} \quad (26)$$

where f is the carrier frequency in kHz, s is the shipping activity factor ranging from 0 to 1 for low and high activity, respectively, and w is the wind speed in m/s [56]. The overall PSD of the ambient noise in dB re μ Pa per Hz, as a function of frequency in kHz is given as,

$$N(f) = N_t(f) + N_s(f) + N_w(f) + N_{th}(f). \quad (27)$$

The SNR observed over a distance L with a transmitted signal power P and carrier frequency f can be evaluated by using the noise PSD $N(f)$ and the signal attenuation $C(L, f)$. The narrow-band SNR is thus given by,

$$SNR(L, f) = \frac{P/C(L, f)}{N(f) \Delta f} \quad (28)$$

where Δf is the receiver noise bandwidth. The frequency-dependent received SNR is plotted in Fig. 14 for a varying transmission distance L , a wind speed $w = 15$ m/s, and

shipping activity factor $s = 0.5$, considering moderate shipping activity. From Fig. 14, we can observe that with the relays located within 100 m distance from the transmitter, the received SNR is ~ 15 dB higher than that of the destination which is at 1000 m distance from the transmitter at a carrier frequency of 7 kHz. This observation is used as a basis for the simulation of our proposed SCCNC scheme. According to our deployment scheme discussed in Section II, the minimum distance between two sensor nodes is 35 m and the maximum distance could be up to 141.5 m in a 100×100 m² area. Looking at Fig. 14, the received SNR difference between a relay at ~ 200 m and destination at ~ 900 m is almost 10 dB. Therefore, considering the worst case scenario, we will use a 10 dB inter-sensor channel SNR, which in ideal case would be up to 15 dB.

REFERENCES

- [1] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: Research challenges," *Ad Hoc Netw.*, vol. 3, no. 3, pp. 257–279, Mar. 2005.
- [2] D. B. Kilfoyle and A. B. Baggeroer, "The state of the art in underwater acoustic telemetry," *IEEE J. Ocean. Eng.*, vol. 25, no. 1, pp. 4–27, Jan. 2000.
- [3] L. Liu, S. Zhou, and J.-H. Cui, "Prospects and problems of wireless communication for underwater sensor networks," *Wireless Commun. Mobile Comput., Underwater Sensor Netw., Archit. Protocols*, vol. 8, no. 8, pp. 977–994, Oct. 2008.
- [4] I. F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks," *ACM SIGBED Rev.*, vol. 1, no. 2, pp. 3–8, Jul. 2004.
- [5] M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 84–89, Jan. 2009.
- [6] L.-Y. Bai, F. Xu, R. Xu, and S.-Y. Zheng, "LDPC application based on CI/OFDM underwater acoustic communication system," in *Proc. Int. Conf. Inf. Sci. Eng. (ICISE)*, Dec. 2009, pp. 2641–2644.
- [7] J. Huang, S. Zhou, and P. Willett, "Nonbinary LDPC coding for multicarrier underwater acoustic communication," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 9, pp. 1684–1696, Dec. 2008.
- [8] B. Li *et al.*, "MIMO-OFDM for high-rate underwater acoustic communications," *IEEE J. Ocean. Eng.*, vol. 34, no. 4, pp. 634–644, Oct. 2009.
- [9] R. Diamant and L. Lampe, "Adaptive error-correction coding scheme for underwater acoustic communication networks," *IEEE J. Ocean. Eng.*, vol. 40, no. 1, pp. 104–114, Jan. 2015.
- [10] L. Wan *et al.*, "Adaptive modulation and coding for underwater acoustic OFDM," *IEEE J. Ocean. Eng.*, vol. 40, no. 2, pp. 327–336, Apr. 2015.
- [11] C. Polprasert, J. A. Ritcey, and M. Stojanovic, "Capacity of OFDM systems over fading underwater acoustic channels," *IEEE J. Ocean. Eng.*, vol. 36, no. 4, pp. 514–524, Oct. 2011.
- [12] K. Pelekanakis and A. B. Baggeroer, "Exploiting space–time–frequency diversity with MIMO–OFDM for underwater acoustic communications," *IEEE J. Ocean. Eng.*, vol. 36, no. 4, pp. 502–513, Oct. 2011.
- [13] Z. Wang, S. Zhou, J. Catipovic, and P. Willett, "Asynchronous multiuser reception for OFDM in underwater acoustic communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1050–1061, Mar. 2013.
- [14] M. C. Domingo, "Overview of channel models for underwater wireless communication networks," *Phys. Commun.*, vol. 1, no. 3, pp. 163–182, Sep. 2008.
- [15] L. M. Brekhovskikh and Y. P. Lysanov, *Fundamentals of Ocean Acoustics*, 3rd ed. New York, NY, USA: Springer-Verlag, 2003.
- [16] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.
- [17] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 371–381, Feb. 2003.
- [18] B. Zhao and M. C. Valenti, "Distributed turbo coded diversity for relay channel," *Electron. Lett.*, vol. 39, no. 2, pp. 786–787, May 2003.
- [19] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, Sep. 2005.
- [20] A. Chakrabarti, A. de Baynast, A. Sabharwal, and B. Aazhang, "Low density parity check codes for the relay channel," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 2, pp. 280–291, Feb. 2007.
- [21] J. Yuan, Z. Chen, Y. Li, and L. Chu, "Distributed space-time trellis codes for a cooperative system," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 4897–4905, Oct. 2009.
- [22] Z. Wang, J. Huang, S. Zhou, and Z. Wang, "Iterative receiver processing for OFDM modulated physical-layer network coding in underwater acoustic channels," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 541–553, Feb. 2013.
- [23] Y. Chen, Z.-H. Wang, L. Wan, H. Zhou, S. Zhou, and X. Xu, "OFDM-modulated dynamic coded cooperation in underwater acoustic channels," *IEEE J. Ocean. Eng.*, vol. 40, no. 1, pp. 159–168, Jan. 2015.
- [24] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.
- [25] Y. Li and X.-G. Xia, "A family of distributed space-time trellis codes with asynchronous cooperative diversity," *IEEE Trans. Commun.*, vol. 55, no. 4, pp. 790–800, Apr. 2007.
- [26] C. Li, G. Yue, M. A. Khojastepour, X. Wang, and M. Madihan, "LDPC-coded cooperative relay systems: Performance analysis and code design," *IEEE Trans. Commun.*, vol. 56, no. 3, pp. 485–496, Mar. 2008.
- [27] D. E. Lucani, M. Medard, and M. Stojanovic, "Underwater acoustic networks: Channel models and network coding based lower bound to transmission power for multicast," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 9, pp. 1708–1719, Dec. 2008.
- [28] M. Elfituri, W. Hamouda, and A. Ghayeb, "A convolutional-based distributed coded cooperation scheme for relay channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 655–669, Feb. 2009.
- [29] Z. Guo, B. Wang, P. Xie, W. Zeng, and J.-H. Cui, "Efficient error recovery with network coding in underwater sensor networks," *Ad Hoc Netw.*, vol. 7, no. 4, pp. 791–802, Jun. 2009.
- [30] T. E. Hunter and A. Nosratinia, "Diversity through coded cooperation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 283–289, Feb. 2006.
- [31] S. Schwandtner, A. Graell i Amat, and G. Matz, "Spatially-coupled LDPC codes for decode-and-forward relaying of two correlated sources over the BEC," *IEEE Trans. Commun.*, vol. 62, no. 4, pp. 1324–1337, Apr. 2014.
- [32] X. Bao and J. Li, "Adaptive network coded cooperation (ANCC) for wireless relay networks: Matching code-on-graph with network-on-graph," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 574–583, Feb. 2008.
- [33] Z. Ding, K. K. Leung, D. L. Goeckel, and D. Towsley, "A relay assisted cooperative transmission protocol for wireless multiple access systems," *IEEE Trans. Commun.*, vol. 58, no. 8, pp. 2425–2435, Aug. 2010.
- [34] J. L. Rebelatto, B. F. Uchoa-Filho, Y. Li, and B. Vucetic, "Multiuser cooperative diversity through network coding based on classical coding theory," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 916–926, Feb. 2012.
- [35] A. M. Jalil and A. Ghayeb, "Distributed channel coding for underwater acoustic cooperative networks," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 848–856, Mar. 2014.
- [36] D. Pompili, T. Melodia, and I. F. Akyildiz, "Three-dimensional and two-dimensional deployment analysis for underwater acoustic sensor networks," *Ad Hoc Netw.*, vol. 7, no. 4, pp. 778–790, Jun. 2009.
- [37] F. Xue and P. R. Kumar, "The number of neighbors needed for connectivity of wireless networks," *J. Wireless Netw.*, vol. 10, no. 2, pp. 169–181, Mar. 2004.
- [38] H.-W. Jeon, S.-J. Lee, and H.-N. Lee, "LDPC coded OFDM system design and performance verification on a realistic underwater acoustic channel model," in *Proc. Military Commun. Conf. (MILCOM)*, Nov. 2011, pp. 2200–2204.
- [39] R. G. Gallager, *Low Density Parity Check Codes*. Cambridge, MA, USA: MIT Press, 1963.
- [40] M. Stojanovic, "Underwater acoustic communication," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, vol. 22. New York, NY, USA: Wiley, 1999, pp. 688–698.
- [41] B. Sklar, *Digital Communications: Fundamentals and Applications*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.
- [42] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 399–431, Mar. 1999.
- [43] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

- [44] LinkQuest Inc. *LinkQuest Underwater Acoustic Modems, UWM2000 Specifications*, accessed on Sep. 27, 2013. [Online]. Available: <http://www.link-quest.com/html/uwm2000.htm>
- [45] EvoLogics GmbH. *Underwater Acoustic Modems, S2CR Acoustic Modem*, accessed on Sep. 27, 2013. [Online]. Available: http://www.evologics.de/en/products/acoustics/s2cr_18_34.html
- [46] T. Mohsenin, D. N. Truong, and B. M. Baas, "A low-complexity message-passing algorithm for reduced routing congestion in LDPC decoders," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 5, pp. 1048–1061, May 2010.
- [47] A. J. Blanksby and C. J. Howland, "A 690-mW 1-Gb/s 1024-b, rate-1/2 low-density parity-check code decoder," *IEEE J. Solid-State Circuits*, vol. 37, no. 3, pp. 404–412, Mar. 2002.
- [48] M. M. Mansour and N. R. Shanbhag, "A 640-Mb/s 2048-bit programmable LDPC decoder chip," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 684–698, Mar. 2006.
- [49] A. G. Zajic, "Statistical modeling of MIMO mobile-to-mobile underwater channels," *IEEE Trans. Veh. Technol.*, vol. 60, no. 4, pp. 1337–1351, May 2011.
- [50] M. Chitre, "A high-frequency warm shallow water acoustic communications channel model and measurements," *J. Acoust. Soc. Amer.*, vol. 122, no. 5, pp. 2580–2586, Nov. 2007.
- [51] M. Stojanovic, "Underwater acoustic communications: Design considerations on the physical layer," in *Proc. Annu. Conf. Wireless Demand Netw. Syst. Services (WONS)*, Jan. 2008, pp. 1–10.
- [52] P. Qarabaqi and M. Stojanovic, "Modeling the large scale transmission loss in underwater acoustic channels," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2011, pp. 445–452.
- [53] W.-B. Yang and T. C. Yang, "High-frequency channel characterization for M-ary frequency-shift-keying underwater acoustic communications," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2615–2626, Nov. 2006.
- [54] A. Radosevic, J. G. Proakis, and M. Stojanovic, "Statistical characterization and capacity of shallow water acoustic channels," in *Proc. IEEE EUROPE OCEANS*, May 2009, pp. 1–8.
- [55] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. Hoboken, NJ, USA: Wiley, 2000.
- [56] R. F. W. Coates, *Underwater Acoustic Systems*. New York, NY, USA: Wiley, 1989.



Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. He then moved to HRL Laboratories, LLC, Malibu, CA, USA, where he worked as a research staff member from 1999 to 2002. He was appointed as an Assistant Professor with the University of Pittsburgh, Pittsburgh, PA, USA, from 2002 to 2008. He then moved to the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, in 2009, where he is currently affiliated. He was appointed as the Dean of Research with GIST in 2015. His general areas of research include information theory, signal processing theory, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface. He has served as a member of technical program committees for several IEEE conferences, including the IEEE International Conference on Communications and the IEEE Global Communications Conference. His research efforts have been recognized with prestigious national awards, including the Top 100 National Research and Development Award from the Korean Ministry of Science, ICT, and Future Planning in 2012, the Top 50 Achievements of Fundamental Researches Award from the National Research Foundation of Korea in 2013, and the Science/Engineer of the Month (January 2014) from the National Research Foundation of Korea. He also served as the Lead Guest Editor for the European Association for *Signal Processing Journal on Wireless Communications and Networking* from 2010 to 2011. He has served as an Area Editor for the *AEU International Journal of Electronics and Communications* since 2013.



Zafar Iqbal received the bachelor's degree in computer engineering from COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2005, and the M.S. degree in information and communications from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2010. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science in GIST. He was with Shanghai R&D Center, ZTE Corporation, China, from 2005 to 2008 and worked with Vieworks Company Limited Korea, in 2011.

His research interests include wireless communication systems, digital signal processing, and design of VLSI circuits and systems. He received the Korea IT Industry Promotion Agency Scholarship for his M.S. degree, and the Korean Government Scholarship for his Ph.D. study and research.

COMPU-EYE: a high resolution computational compound eye

Woong-Bi Lee,¹ Hwanchol Jang,¹ Sangjun Park,¹ Young Min Song,² and Heung-No Lee^{1,*}

¹School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, 61005, South Korea

²Department of Electronics Engineering, Pusan National University, Busan, 46241, South Korea
heungno@gist.ac.kr

Abstract: In nature, the compound eyes of arthropods have evolved towards a wide field of view (FOV), infinite depth of field and fast motion detection. However, compound eyes have inferior resolution when compared with the camera-type eyes of vertebrates, owing to inherent structural constraints such as the optical performance and the number of ommatidia. For resolution improvements, in this paper, we propose COMPUTational compound EYE (COMPU-EYE), a new design that increases acceptance angles and uses a modern digital signal processing (DSP) technique. We demonstrate that the proposed COMPU-EYE provides at least a four-fold improvement in resolution.

©2016 Optical Society of America

OCIS codes: (110.1758) Computational imaging; (100.6640) Superresolution; (110.3010) Image reconstruction techniques; (120.4570) Optical design of instruments.

References and links

1. E. Warrant and D.-E. Nilson, *Invertebrate Vision*, (Cambridge University, 2006), Chap. 1.
2. R. Dudley, *The Biomechanics of Insect Flight: Form, Function, Evolution*, (Princeton University, 2000), Chap. 5.
3. D. Floreano, J.-C. Zufferey, M. V. Srinivasan, and C. Ellington, *Flying Insects and Robot*, (Springer, 2009), Chap. 10.
4. M. F. Land and D.-E. Nilson, *Animal Eyes* (Oxford University, 2002).
5. M. F. Land, "The optics of animal eyes," *Contemp. Phys.* **29**(5), 435–455 (1988).
6. D.-E. Nilson, "Vision optics and evolution," *Bioscience* **39**(5), 298–307 (1989).
7. A. Borst and J. Plett, "Optical devices: Seeing the world through an insect's eyes," *Nature* **497**(7447), 47–48 (2013).
8. Y. M. Song, Y. Xie, V. Malyarchuk, J. Xiao, I. Jung, K. J. Choi, Z. Liu, H. Park, C. Lu, R. H. Kim, R. Li, K. B. Crozier, Y. Huang, and J. A. Rogers, "Digital cameras with designs inspired by the arthropod eye," *Nature* **497**(7447), 95–99 (2013).
9. A. Brückner, J. Duparré, A. Bräuer, and A. Tünnermann, "Artificial compound eye applying hyperacuity," *Opt. Express* **14**(25), 12076–12084 (2006).
10. J. Duparré, F. Wippermann, P. Dannberg, and A. Bräuer, "Artificial compound eye zoom camera," *Bioinspir. Biomim.* **3**(4), 046008 (2008).
11. Y. Kitamura, R. Shogenji, K. Yamada, S. Miyatake, M. Miyamoto, T. Morimoto, Y. Masaki, N. Kondou, D. Miyazaki, J. Tanida, and Y. Ichioka, "Reconstruction of a High-Resolution Image on a Compound-Eye Image-Capturing System," *Appl. Opt.* **43**(8), 1719–1727 (2004).
12. K. H. Jeong, J. Kim, and L. P. Lee, "Biologically Inspired Artificial Compound Eyes," *Science* **312**(5773), 557–561 (2006).
13. D. P. Pulsifer, A. Lakhtakia, R. J. Martín-Palma, and C. G. Pantano, "Mass fabrication technique for polymeric replicas of arrays of insect corneas," *Bioinspir. Biomim.* **5**(3), 036001 (2010).
14. P. Qu, F. Chen, H. Liu, Q. Yang, J. Lu, J. Si, Y. Wang, and X. Hou, "A simple route to fabricate artificial compound eye structures," *Opt. Express* **20**(5), 5775–5782 (2012).
15. L. Li and A. Y. Yi, "Design and fabrication of a freeform microlens array for a compact large-field-of-view compound-eye camera," *Appl. Opt.* **51**(12), 1843–1852 (2012).
16. R. Hornsey, P. Thomas, W. Wong, S. Pepic, K. Yip, and R. Krishnasamy, "Electronic Compound-Eye Image Sensor: Construction and Calibration," *Proc. SPIE* **5301**, 13–24 (2004).
17. H. Zhang, L. Li, D. L. McCray, S. Scheiding, N. J. Naples, A. Gebhardt, S. Risse, R. Eberhardt, A. Tünnermann, and A. Y. Yi, "Development of a low cost high precision three-layer 3D artificial compound eye," *Opt. Express* **21**(19), 22232–22245 (2013).
18. D. Floreano, R. Pericet-Camara, S. Viollet, F. Ruffier, A. Brückner, R. Leitel, W. Buss, M. Menouni, F. Expert, R. Juston, M. K. Dobrzynski, G. L'Éplattenier, F. Recktenwald, H. A. Mallot, and N. Franceschini, "Miniature curved artificial compound eyes," *Proc. Natl. Acad. Sci. U.S.A.* **110**(23), 9267–9272 (2013).

19. T. Someya, T. *Stretchable Electronics* (Wiley, 2013).
20. F. Marefat, A. Partovi, and A. Mousavinia, "A Hemispherical Omni-directional Bio Inspired Optical Sensor," in *Proceedings of Iranian Conference on Electrical Engineering (ICEE)* (2012), pp. 668–672.
21. M. F. Land, "Visual Acuity in Insects," *Annu. Rev. Entomol.* **42**(1), 147–177 (1997).
22. H. B. Barlow, "The size of ommatidia in apposition eyes," *J. Exp. Biol.* **29**, 667–674 (1952).
23. F. Zettler and R. Weiler, *Neural Principles in Visions* (Springer, 1976), Chap. 2.9.
24. G. Cristóbal, L. Perrinet, and M. S. Keil, *Biologically Inspired Computer Vision: Fundamentals and Applications* (Wiley-VCH, 2015).
25. P. T. Gonzalez-Bellido, T. J. Wardill, and M. Jusuola, "Compound eyes and retinal information processing in miniature dipteran species match their specific ecological demands," *Proc. Natl. Acad. Sci. U.S.A.* **108**(10), 4224–4229 (2011).
26. H. C. Ko, M. P. Stoykovich, J. Song, V. Malyarchuk, W. M. Choi, C.-J. Yu, J. B. Geddes 3rd, J. Xiao, S. Wang, Y. Huang, and J. A. Rogers, "A hemispherical electronic eye camera based on compressible silicon optoelectronics," *Nature* **454**(7205), 748–753 (2008).
27. J. Tanida, T. Kumagai, K. Yamada, S. Miyatake, K. Ishida, T. Morimoto, N. Kondou, D. Miyazaki, and Y. Ichioka, "Thin observation module by bound optics (TOMBO): concept and experimental verification," *Appl. Opt.* **40**(11), 1806–1813 (2001).
28. C. Shi and F. Xu, "Post-digital image processing based on microlens array," *Proc. SPIE* **92701K**, 9270 (2014).
29. E. Watson, R. Muse, and F. Blommel, "Aliasing and blurring in microscanned imagery," *Proc. SPIE* **1689**, 242–250 (1992).
30. J. Oliver, W.-B. Lee, and H.-N. Lee, "Filters with random transmittance for improving resolution in filter-array-based spectrometers," *Opt. Express* **21**(4), 3969–3989 (2013).
31. H. Jang, C. Yoon, E. Chung, W. Choi, and H.-N. Lee, "Holistic random encoding for imaging through multimode fibers," *Opt. Express* **23**(5), 6705–6721 (2015).
32. J. Fang, J. Li, Y. Shen, H. Li, and S. Li, "Super-Resolution Compressed Sensing: An Iterative Reweighted Algorithm for Joint Parameter Learning and Sparse Signal Recovery," *IEEE Signal Process. Lett.* **21**(6), 761–765 (2014).
33. M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. (Springer, 2010).
34. E. J. Candès, "Compressive sampling," *Proc. Int. Congr. Mathematicians* **3**, 1433–1452 (2006).
35. E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006).
36. D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006).
37. E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Trans. Inf. Theory* **52**(12), 5406–5425 (2006).
38. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE T. Signal Process.* **54**(11), 4311–4322 (2006).
39. R. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.* **24**(4), 118–121 (2007).
40. D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory* **52**(1), 6–18 (2006).
41. E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.* **59**(8), 1207–1223 (2006).
42. A. Tavakoli and A. Pourmohammad, "Image Denoising Based on Compressed Sensing," *Int. J. Comp. Theory Eng.* **4**, 266–269 (2012).
43. J. Yang and Y. Zhang, "Alternating direction algorithms for l_1 -problems in compressive sensing," *SIAM J. Sci. Comput.* **33**(1), 250–278 (2011).
44. A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.* **2**(1), 183–202 (2009).
45. A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.* **40**(1), 120–145 (2011).
46. W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inf. Theory* **56**(6), 2969–2979 (2008).
47. E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.* **31**(1), 59–73 (2011).
48. H. Rauhut, "Compressive sensing and structured random matrices," *Theor. Found. Num. Meth. Sparse Recov.* **9**, 1–92 (2010).
49. R. Chartrand, "Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data," in *IEEE International Symposium on Biomedical Imaging (IEEE, 2009)*, pp. 262–265.
50. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.* **3**(1), 1–122 (2010).

1. Introduction

Compound eyes of arthropods such as ants, flies and bugs have attracted extensive research interest due to their unique features such as wide field-of-view (FOV), high sensitivity to motion and infinite depth of field [1–3]. An apposition type of compound eye in nature consists of integrated optical units called ommatidia, each of which includes a light diffracting

facet lens, crystalline cone, wave guiding rhabdom and photoreceptor cell [4–6]. Each ommatidium arranged along a spherical surface senses incident light within a small range of angular acceptance. Implementations of optical devices inspired by natural compound eyes exhibit great potential in various fields such as surveillance cameras on micro aerial vehicles, high-speed motion detection, endoscopic medical tools, and image guided surgery [7,8].

For years, several attempts to develop artificial compound eyes have been based on microlenses and photodetectors to imitate imaging organs of a natural ommatidium. Because most optoelectronics technologies developed earlier were intrinsically based on a planar substrate, both devices were implemented on a plane [9–11]. Planar compound eyes had low design and fabrication complexity, but they incurred a limited FOV. Later, curved microlens arrays were developed and interfaced with conventional planar sensors [12–17], but these suffered from off-axis aberrations, crosstalk between adjacent ommatidia, or restricted FOV [18]. They also required optical relay devices for beam-steering, which are complicated to fabricate [15–17]. In recent years, with the advance of flexible optoelectronics [19], curvilinear structured compound eyes, which provide larger FOVs, have been developed [8,18,20]. A hemispherical omni-directional optical sensor was implemented by a circular central board and multiple modular sensor boards [20]. Another cylindrical compound eye was introduced by bending the planar ommatidial array along a concave substrate [18]. Song *et al.* implemented a hemispherical compound eye by reformulating stretchable planar ommatidia into hemispherical ommatidia [8]. We note that the hemispherically structured compound eye developed in [8], which is most comparable to a natural compound eye is mainly considered in this paper.

It is well known that the vision of insects is far inferior to that of humans because of inherent structural constraints [21–23]. Generally, the resolution of any eye depends not only on its optical resolution but also on the number of the receptors. First, if the optics are free of defects, the resolution of any optical imaging system is determined by its diffraction limit. The resolution of a diffraction-limited imaging system is proportional to the size of its lens and inversely proportional to the wavelength of the observed light. Second, in apposition-type compound eyes, the basic sampling units are ommatidia rather than photoreceptors. In a diffraction-limited compound eye, in order to accommodate many separate ommatidia without crosstalk, the number of ommatidia is much smaller than that of photoreceptors in the retina of a human eye. In nature, the density of the photoreceptors in the human eye is about 25 times higher than the ommatidial density of the compound eye [24]. For a compound eye to achieve a resolution similar to that of a human eye, it requires a radius of about 6 m and millions of ommatidia with facet lenses as large as a pupil, which is impractical [21].

Artificial compound eyes that mimic the structure of natural compound eyes are also limited on their image resolutions. In the design of the compound eyes, the spatial resolution that the compound eye can resolve depends on the relation between the acceptance angle ($\Delta\varphi$) of the ommatidia and the interommatidial angle ($\Delta\phi$) between the optical axes of the neighboring ommatidia [21,25]. In nature, for most light-adapted diurnal animals, the acceptance angles of ommatidia approach the interommatidial angle, i.e., $\Delta\varphi \approx \Delta\phi$ [21], which achieves high spatial resolution by minimizing aliasing among neighboring ommatidia. For example, *Tenodera* has angles $\Delta\varphi = 0.7^\circ$ and $\Delta\phi = 0.6^\circ$, and *Calliphora* has angles $\Delta\varphi = 1.02^\circ$ and $\Delta\phi = 1.5^\circ$. Analogous to natural compound eyes, artificially developed compound eyes have been designed to have similar acceptance and interommatidial angles [8,18]. The acceptance and interommatidial angles have been chosen to be $\Delta\varphi = 9.7^\circ$ and $\Delta\phi = 11^\circ$, and $\Delta\varphi = 4.2^\circ$ and $\Delta\phi = \sim 4.2^\circ$ in the literature [8,18]. Compared to the human eye, the artificial compound eyes are fundamentally limited on the resolution and thus they are inappropriate for object recognition.

For improving the quality of the observed image, a scanning method was introduced by capturing the object image repeatedly with different angle of rotations in [8,26]. As a result, an image of 160×160 pixels was obtained only with 16×16 ommatidia by scanning the compound-eye camera and thus the actual resolution of the observed image was improved by

100 times [8]. However, the repeated image capturing with fine mechanical angle controls makes the scanning method inefficient. In [27], a compact imaging system called TOMBO (thin observation module by bound optics) was introduced, which consists of a multi-aperture imaging system and a post-signal processing. The TOMBO reconstructs the object image with high resolution from multiple low-resolution subimages by exploiting the relation between the object and the captured signals. Afterward, many techniques were proposed to improve the reconstruction performance of the TOMBO system [27,28]. However, the FOVs are limited because they are planar compound eyes.

In this paper, instead of enhancing the size and number of the ommatidia for improving the resolution, we propose a totally different imaging system, called COMPUTational compound EYE (COMPU-EYE), using a modern digital signal processing (DSP) technique. Conventional compound eyes are designed to have limited ommatidial acceptance angles to avoid aliasing. Thus, each ommatidium of the conventional compound eye observes an independent section of the object image. In contrast, the ommatidium of COMPU-EYE has larger acceptance angles. This increase in acceptance angle allows a single ommatidium to observe multiple pieces of information simultaneously. Because the multiple pieces of information in each observation interfere with each other, the observed image is distorted. We employ a DSP technique in COMPU-EYE to recover the object image from these observations. In the DSP, by utilizing the fact that one piece of information is observed by multiple ommatidia with different perspectives, COMPU-EYE improves the resolution of the object image.

For a classical resolution improvement technique, a microscan technique requires to capture multiple frames of a target with slightly displaced locations [29]. The sequences of frames are then integrated to form a high resolution image. In contrast, COMPU-EYE provides a high resolution image reconstruction with a single frame of the target with less number of samples. The high resolution reconstruction is achieved by solving an underdetermined linear system of equations as will be introduced in Eq. (1) in Section 2. As a fast emerging area in DSP, compressed sensing (CS) provides a sparse solution to the underdetermined system. Recently, there are other papers who studied CS with the intension of improving resolution in various areas such as spectroscopy [30], optical imaging [31], and direction of arrival estimation [32]. In this paper, we propose to design a compound eye with large ommatidial acceptance angles, which is appropriate for the framework of the CS-based super-resolution, and reconstruct the object with high resolution using the DSP technique.

This paper is organized as follows: In Section 2, the system model of the compound eye imaging system is described. In Section 3, we propose COMPU-EYE and describe how COMPU-EYE improves the resolution by comparison with the conventional compound eye, and Section 4 presents the experimental results. Section 5 concludes the paper.

2. System description

We consider the biologically inspired compound eyes of a hemispherical structure as seen in Fig. 1(a). Details of the optical design of the hemispherical compound eye is referred from [8]. Each ommatidium, a basic imaging unit can be implemented by a set of microlens, supporting posts connected to a base membrane and a photodetector. An array of microlenses and photodiodes are integrated in the planar layout and are transformed into a full hemispherical shape. Note that the ommatidium is based on a circular lattice because the microlens is hemispherical shape compared to the hexagonal lattice in compound eyes in nature [21]. Each ommatidium receives incident light within its acceptance angle defined by $\Delta\phi$ and is separated by an interommatidial angle $\Delta\phi$ from each other. We note that the optical transfer function of an ommatidium can be modeled as a Gaussian function. For simplicity, we assumed that the optical transfer function is simplified by neglecting light whose relative light intensities are smaller than a certain value. Thus, each ommatidium is modeled to collect averaged optical signal from light incident within its acceptance angle, $\Delta\phi$, as seen in Fig. 1(b). With the compound eye of the hemispherical structure, we now consider an imaging system with M ommatidia as seen in Fig. 1(c). The imaging system observes an object image

on the plane size of $U \times V$ mm, which is located L mm away from the compound eye. According to the acceptance angle and object image distance, the receptive field (i.e., visible area at the object plane) of a single ommatidium is determined. Each observation contributes to a single pixel that contains the intensity of the light collected from its corresponding receptive field. The final image is reconstructed by a set of these pixels.

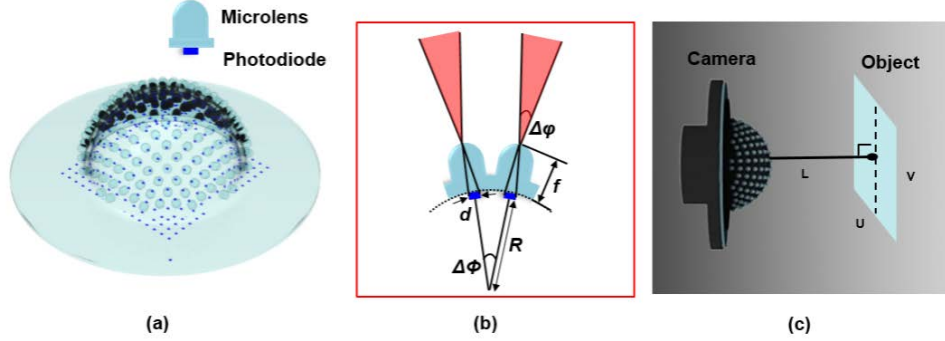


Fig. 1. (a) Illustration of the hemispherical compound eye. (b) Structure of conventional compound eye with key parameters: the acceptance angle ($\Delta\phi$) and focal length (f) for each ommatidium, the interommatidial angle ($\Delta\phi$), the diameter of a photodiode (d) and the radius of curvature of the compound eye (R) and of an individual microlens (r). (c) Compound eye imaging system

Let y_i denote an output sample at the i^{th} ommatidium for $i \in \{1, 2, \dots, M\}$. We assume that the image to be reconstructed consists of N_U by N_V pixels, each having uniform light intensity. The size of each pixel is $U/N_U \times V/N_V$ mm. The object image forms an $N \times 1$ vector $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$, where $N = N_U N_V$. On the basis of ray tracing analysis, the sample y_i can be obtained from $y_i = \mathbf{a}_i \mathbf{x}$, where \mathbf{a}_i is an $1 \times N$ vector whose elements represent the visibility of the i^{th} ommatidium at each of the N pixels. For the i^{th} ommatidium, if the j^{th} pixel for $j \in \{1, 2, \dots, N\}$ is outside the receptive field, which represents the j^{th} pixel is invisible to the i^{th} ommatidium, then the j^{th} component a_{ij} in \mathbf{a}_i becomes zero, i.e., $a_{ij} = 0$. If the j^{th} pixel is inside the receptive field, which represents the j^{th} pixel is fully observed by the i^{th} ommatidium, then $a_{ij} = 1$. Otherwise, if the j^{th} pixel is on the boundary of the receptive field, which represents the j^{th} pixel is partially observed by the i^{th} ommatidium, then $0 < a_{ij} < 1$, which is proportional to the intersection area of the receptive field and pixel. This process can be summarized as follows:

$$a_{ij} = \begin{cases} 0 & , j^{\text{th}} \text{ pixel is invisible to } i^{\text{th}} \text{ ommatidium} \\ 1 & , j^{\text{th}} \text{ pixel is fully observed by } i^{\text{th}} \text{ ommatidium} \\ 0 < a_{ij} < 1, & j^{\text{th}} \text{ pixel is partially observed by } i^{\text{th}} \text{ ommatidium} \end{cases} .$$

When collecting M samples, the ommatidial observations can be modeled as a system of linear equations as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}. \quad (1)$$

where $\mathbf{y} = [y_1, \dots, y_M]^T$ is a set of M output samples and $\mathbf{n} \in \mathbb{R}^{M \times 1}$ contains unexpected noise. Let $\mathbf{A} = [\mathbf{a}_1^T \ \mathbf{a}_2^T \ \dots \ \mathbf{a}_M^T]^T \in \mathbb{R}^{M \times N}$ denote a measurement matrix the i^{th} row of which is \mathbf{a}_i . Given

the measurement matrix \mathbf{A} and the observation \mathbf{y} , we aim to solve the system of linear equations in Eq. (1) for the object image reconstruction.

In this paper, since we are considering resolution improvements in the compound eye imaging system, the number of estimated pixels is set to be greater than the number of ommatidia. i.e., $N > M$. Thereby, we note that Eq. (1) becomes an underdetermined system of linear equations. This underdetermined system can be solved by a convex optimization if the object is represented as a sparse signal in the proper domain [33–36]. A sparse signal is often represented as a vector which has a small number of non-zero components. We note that any natural image can be sparsely represented in a certain domain such as wavelets, the discrete cosine transform (DCT), or the discrete Fourier transform [37,38].

In an underdetermined system, the solution can be found by solving the l_0 minimization problem

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon. \quad (2)$$

where $\|\mathbf{x}\|_0$ denotes the number of non-zero components in \mathbf{x} and ε is a small positive constant. However, the optimization problem in Eq. (2) is combinatorial and computationally intractable [39]. Alternatively, the l_1 norm minimization provides unique and sparse solutions for underdetermined systems by solving

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon. \quad (3)$$

We note that the l_1 norm minimization guarantees stability, which means that it can reliably reconstruct the signal without amplifying the observation errors in the process of l_1 norm minimization [40,41]. The l_1 norm minimization reconstructs a signal with explicit sparsity constraints while removing non-sparse random noise components from a corrupted signal. Due to its property of noise suppression, the l_1 norm minimization has been used as an image denoising tool [42]. Recently, many algorithms [43–45] have been proposed to solve Eq. (3). In this study, we use the alternating direction method [43], which is known to be fast and efficient for the problem in Eq. (3). If an object image of N pixels is reconstructed, where $N > M$, the resolution of COMPU-EYE is improved by a factor of N/M .

In the following section, we propose a COMPU-EYE imaging system, which is more appropriate to resolve Eq. (3) and thus to reconstruct the object image with high resolution.

3. COMPU-EYE for image acquisition and reconstruction with high resolution

In this section, we introduce COMPU-EYE. In COMPU-EYE, we propose to increase the acceptance angles of ommatidia larger than the interommatidial angle to recover the object image with computations. We first compare COMPU-EYE with the conventional compound eye imaging system in terms of resolution limit. We then explain how COMPU-EYE improves the resolution by investigating the influence of larger acceptance angles on the measurement matrix of the image capturing system.

3.1 Overview and comparison of compound eyes

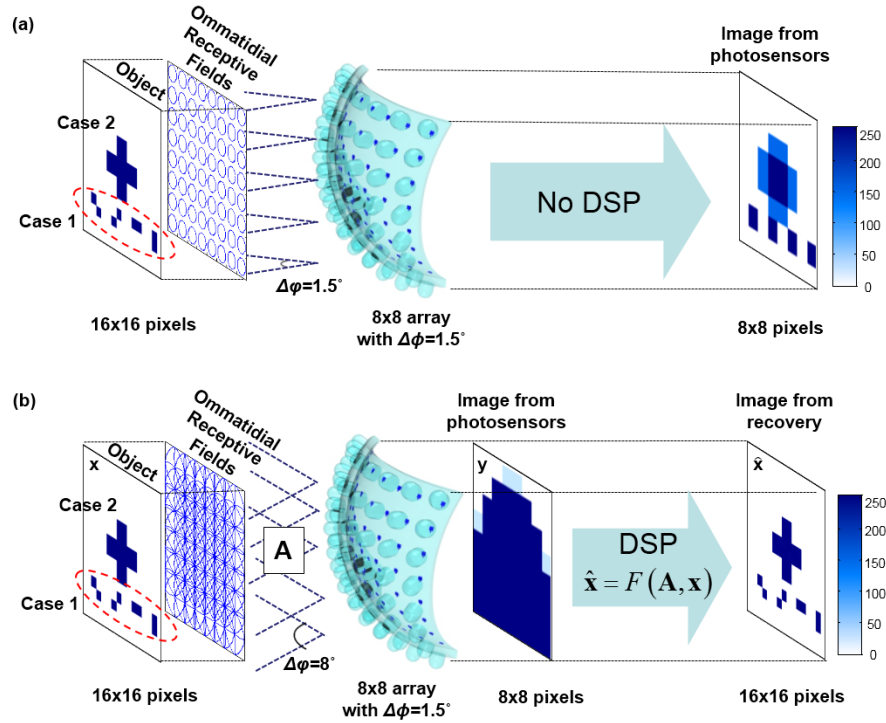


Fig. 2. Imaging systems of a conventional compound eye and the proposed COMPU-EYE (a) The conventional compound eye consists of 8×8 ommatidia with $\Delta\phi = 1.5^\circ$ and $\Delta\varphi = 1.5^\circ$. (b) COMPU-EYE consists of 8×8 ommatidia with $\Delta\phi = 1.5^\circ$ and $\Delta\varphi = 8^\circ$ as well as a DSP algorithm.

The imaging system of a conventional compound eye is depicted in Fig. 2(a). It has a hemispherical structure with a radius (R) of 6.9216 mm, and consists of 8×8 ommatidia, each of which has a height (f) of 1.35 mm. Because each ommatidium provides a single sample, the compound eye has $M = 64$ samples. An 8×8 mm object image is located at a distance of 30 mm from the compound eye. The receptive field of a single ommatidium is shown as an ellipse, and a set of these receptive fields forms the ommatidial receptive fields near the left in Fig. 2(a).

In the conventional compound eye, the acceptance angles of the ommatidia are typically designed to be similar to the interommatidial angle (i.e., $\Delta\varphi = \Delta\phi = 1.5^\circ$) in order to maximize the spatial resolution as well as to avoid overlapping ommatidial receptive fields. Accordingly, the ommatidial receptive fields are totally isolated, and each ommatidium observes an independent part of the object image. Each observation forms a single pixel in the reconstructed image. Note that no signal processing technique is needed to reconstruct the image.

To demonstrate the resolution limit of the conventional compound eye, we consider an object image comprising two parts as seen in Fig. 2(a): 1) four different patterns with the same light intensity, each of which is included in the receptive field of an ommatidium; and 2) a cross pattern that lies over several receptive fields.

Because every pattern in Case 1 is included within a receptive field, every observation appears to have a single image pixel with the same intensity of light. As a result, finer details within a receptive field cannot be resolved and the four different patterns in Case 1 cannot be distinguished by a conventional compound eye. Moreover, because its ommatidial receptive

fields are totally isolated, the fields contain *undetectable areas*, i.e., areas that are invisible to the compound eye. The undetectable areas deteriorate the image quality by decreasing the intensity of light observations as seen in right side of Fig. 2(a). This example shows that the conventional compound eye roughly recognizes object patterns, but has undetectable areas. As a result, such compound eyes suffer from limited resolution and poor image quality.

In contrast, consider the proposed COMPU-EYE imaging system in Fig. 2(b). COMPU-EYE consists of an 8×8 hemispherical array of ommatidia with acceptance angles that are larger than the interommatidial angle, i.e., $\Delta\phi = 8^\circ > \Delta\phi = 1.5^\circ$. It is also equipped with a DSP technique. Because of the increased acceptance angles, the receptive field of each ommatidium is increased to at least 28 times that of $\Delta\phi = 1.5^\circ$. Thus, the ommatidial receptive fields widely overlap, severely distorting the observations as seen in the third frame from left in Fig. 2(b). We then apply DSP to recover a high resolution object image from these highly distorted observations.

In Fig. 2(b), the proposed COMPU-EYE recovers an object image of 256 pixels from 64 observations. The resolution is improved by a factor of four. In recovered image $\hat{\mathbf{x}}$, finer details that were perceived as a single point in Fig. 2(a) can be resolved, and different patterns in Case 1 are distinguished by COMPU-EYE. Moreover, COMPU-EYE compensates for undetectable areas and hence prevents the deterioration of the recovered image quality in Case 2. As a result, COMPU-EYE provides a higher-resolution image of better quality than the conventional compound eye.

3.2 Effects of larger acceptance angles and resolution improvements

We now focus on how larger acceptance angles along with the DSP technique improve resolution with respect to measurement matrix characteristics of the conventional compound eye and COMPU-EYE.

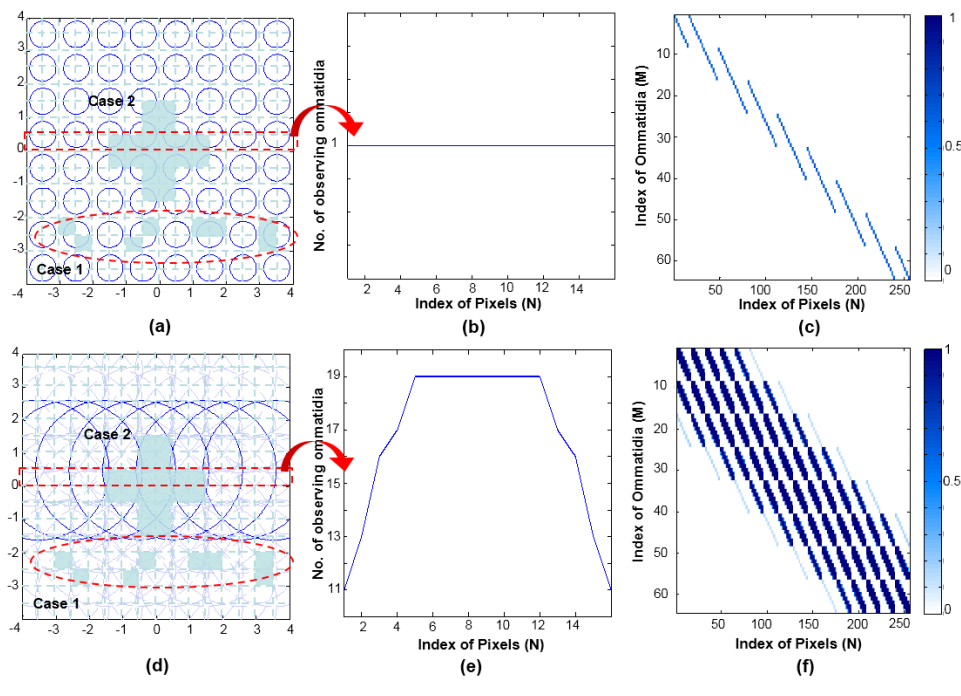


Fig. 3. Effects of acceptance angles for the conventional compound eye (top row) and COMPU-EYE (bottom row) (a)(d) Ommatidial receptive fields overlapped with the object image. (b)(e) Number of observing ommatidia corresponding to pixels in the 8th row. (c)(f) Graphical representations of the measurement matrices.

Figure 3(a) shows how the object image of a 16×16 array of pixels is projected onto the 8×8 ommatidial receptive fields of the conventional compound eye, where $\Delta\varphi = \Delta\phi = 1.5^\circ$. The measurement matrix of the conventional compound eye $\mathbf{A} \in \mathbb{R}^{M \times N}$ in Eq. (1) can be obtained from the ommatidial receptive fields and pixels of the object image in Fig. 3(a). This measurement matrix is displayed graphically in Fig. 3(c). Every element in the measurement matrix indicates the visibility of the corresponding row of an ommatidium in the corresponding column of a pixel. Because the receptive fields of the ommatidia are small and isolated, the measurement matrix has few nonzero components. In Fig. 3(a), each ommatidium separately observes four corresponding pixels, and each pixel is observed by a single ommatidium. The values of the four pixels in one receptive field are considered to be of a single light intensity. Thus, each observation and its observed pixels are in a one-to-many correspondence relation. Because the information of one pixel is contained in one ommatidium as seen in Fig. 3(b), there is no additional information regarding that pixel in other observations. Therefore, in such relationships, finer details within the receptive field cannot be resolved and the resolution of the conventional compound eye is limited by M measurements. We note that the coefficients in Fig. 3(c) are smaller than one because the pixel cannot be entirely observed by ommatidia owing to the undetectable areas.

In contrast, COMPU-EYE has a larger acceptance angle of $\Delta\varphi = 8^\circ$. Figure 3(d) shows how the object image is superimposed on the ommatidial receptive fields of COMPU-EYE. The size of each receptive field is considerably larger; a single ommatidium covers up to 76 pixels, which is considerably greater than the four pixels of the conventional compound eye. Whereas each receptive field in the conventional compound eye is small and separated, each receptive field in COMPU-EYE is large and highly overlapped. Hence, undetectable areas do not exist in the receptive fields of COMPU-EYE. As a result, the number of nonzero components increases correspondingly in the measurement matrix of COMPU-EYE in Fig. 3(f).

The measurement matrix of COMPU-EYE is appropriate for image acquisition and reconstruction rather than that of the conventional compound eye because the object elements, \mathbf{x} in Eq. (1) is more likely to be aligned with the nonzero elements of the matrix [46]. As shown in Fig. 3(e), each pixel of the object image is observed multiple times with different ommatidia. In the context of information acquisition, the total quantity of information for a pixel is increased. Each observation is not redundant to the others for it has different receptive field. Accordingly, each column in the measurement matrix has multiple nonzero elements with different coefficients in Fig. 3(f). The observation of a pixel sufficiently differs from its other observations and this provides additional information about the pixel. In the literature, it is shown that such additional information is useful for reliable signal recovery, even if the number of measurements is smaller than the dimension of the original signal [31,37,46–48]. Thus, the large acceptance angle of ommatidia with the use of DSP allows COMPU-EYE to resolve finer details of the object beyond the resolution limit of M measurements.

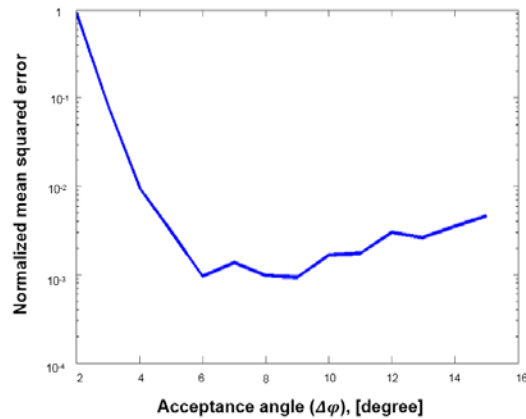


Fig. 4. NMSE against acceptance angle where $M = 8 \times 8$ ommatidia with $\Delta\phi = 1.5^\circ$ and $N = 16 \times 16$ pixels.

We now investigate the reconstruction performance of the DSP technique in accordance with the acceptance angle in the example of Fig. 2. A randomly located sparse signal with 10 nonzero components is generated with uniform distribution between 0 and 1. As a measure of the reconstruction performance evaluation, let us define the normalized mean squared error (NMSE) as $\text{NMSE} = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2$. As seen in Fig. 4, when the acceptance angle is small, the object is unable to be reconstructed with low errors. Specifically, when $\Delta\phi = 2^\circ$ which corresponds to the conventional compound eye in the example of Fig. 2(a), the 16×16 pixels cannot be recovered from 8×8 ommatidia. Associated with the measurement matrix in Fig. 3(c), each observation and its corresponding observed pixels are one-to-many correspondence. Thus, each ommatidium is unable to resolve fine details of its observation. As the acceptance angle increases, each pixel is observed multiple times by different ommatidia. The DSP technique reconstructs each pixel with low errors by solving Eq. (3). As a result, the NMSE decreases. When $\Delta\phi > 8^\circ$, it is seen that the NMSE gradually increases because each observation becomes redundant with neighboring observations. We note that the analysis on the optimal acceptance angle is remained as our future works.

We note that the acceptance angle can be easily increased in many possible ways in an artificial compound eye. The acceptance angle within an ommatidium can be represented as $\Delta\phi_o = \sqrt{(d/f)^2 + (\lambda/D)^2}$, where D is the lens diameter, λ is the light wavelength, d is the photosensor diameter, and f is the focal length of the ommatidial optics [21]. According to Snell's law, the acceptance angle $\Delta\phi$ outside the ommatidium can be obtained by $\Delta\phi = 2 \sin^{-1}((n_o/n_i) \sin(\Delta\phi_o/2))$, where the refractive indices of the lens material and air are defined as n_o and n_i , respectively. Thus, the acceptance angle $\Delta\phi$ can be increased by using a material of higher refractive index for the ommatidia, decreasing the focal length f , or increasing the diameter d of the photodetector. Note that increasing the diameter of the photodetector may lead to increase the size of the ommatidia and the size of the compound eye. On the other hand, decreasing the radius of the curvature of the microlens for reducing the focal length can increase the acceptance angle without increasing the size of the ommatidia.

4. Results

To evaluate the performance of our design, we consider a hemispherical compound eye with a radius of $R = 6.9216$ mm, where each ommatidium has a height of $f = 1.35$ mm in Fig. 1(b). The compound eye consists of a varying number M of ommatidia of uniform spacing with the

interommatidial angle $\Delta\phi = 180/\sqrt{M}^\circ$. The object image to be reconstructed is composed of $N = 160 \times 160$ pixels. As a sparsity measure of the image, we use the Sparsity Ratio (SR) defined as a ratio of the number of nonzero elements to the total length of the signal.

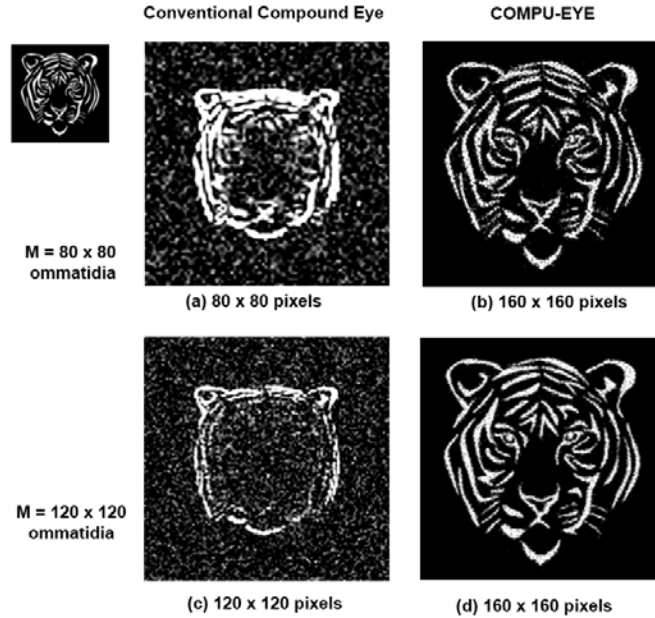


Fig. 5. For $M = 80 \times 80$ and $\Delta\phi = 2.25^\circ$, (a) Output image of the conventional compound eye with $\Delta\phi = 2.25^\circ$ (b) Image recovered by COMPU-EYE with $\Delta\phi = 60^\circ$. For $M = 120 \times 120$ and $\Delta\phi = 1.5^\circ$, (c) Output image of the conventional compound eye with $\Delta\phi = 1.5^\circ$ (d) Image recovered by COMPU-EYE with $\Delta\phi = 60^\circ$.

We demonstrate the performance of COMPU-EYE with an image in the presence of noise. The object image is a line-art illustration of a tiger, which consists of 160×160 pixels each of which contains an 8-bit quantized light intensity. The sparsity ratio of the tiger image is $SR = 0.2335$. The conventional compound eye consists of $M = 80 \times 80$ ommatidia with $\Delta\phi = 2.25^\circ$. On the other hand, COMPU-EYE consists of $M = 80 \times 80$ ommatidia of much larger acceptance angles, $\Delta\phi = 60^\circ$ than $\Delta\phi = 2.25^\circ$. The object image size of 60×60 mm is at a distance of 10mm from the compound eyes. An additive observation noise in Eq. (1) is assumed to be Gaussian with zero mean and covariance matrix $\sigma^2 \mathbf{I}_M$ where $\sigma^2 = 0.1$. Figure 5(a) shows the output image of the conventional compound eye. The output image is corrupted by noise. Because of the resolution limit determined by M and undetectable areas in the ommatidial receptive fields, the observed image of the conventional compound eye is poor quality. Figure 5(b) shows the image recovered by COMPU-EYE equipped with the DSP technique. Compared to the Fig. 5(a), COMPU-EYE provides a higher resolution imaging as well as denoising effects. Due to the stability of the l_1 norm minimization, the unexpected noise is efficiently removed in the reconstructed image without any denoising algorithm. When the number of ommatidia is increased to $M = 120 \times 120$ with $\Delta\phi = 1.5^\circ$, the output image of the conventional compound eye and the recovered image of COMPU-EYE are shown in Figs. 5(c) and 5(d), respectively. As we increase the number of ommatidia, the object image is more clearly seen. For a measure of the resolution improvement, we define a *pixel resolution* as the total number of pixels to be reconstructed with $NMSE < \delta$, where $\delta > 0$ is a user-defined positive number. Since the number of pixels to be recovered is increased from 80^2 to 160^2 in Fig. 5(b) and from 120^2 to 160^2 in Fig. 5(d), the gain in the pixel

resolution is 4 and 1.78, respectively. We note that the size of the observed image in a conventional compound eye is smaller than that of the recovered image of the proposed eye; this is because the ommatidia on the edge of a conventional compound eye are unable to detect the object image owing to their small range of acceptance angle.

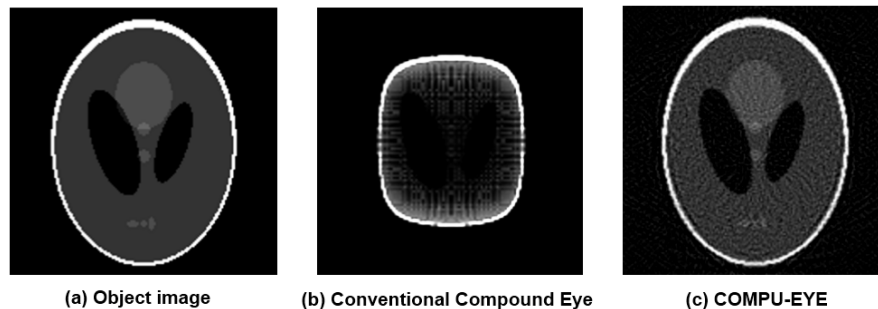


Fig. 6. For the compound eyes, $M = 120 \times 120$ and $\Delta\phi = 1.5^\circ$. (a) Object image of 8-bit grayscale image with 160×160 pixels (b) Output image of the conventional compound eye with $\Delta\phi = 1.5^\circ$ and. (c) Image recovered by COMPU-EYE with $\Delta\phi = 60^\circ$.

We now investigate the performance of COMPU-EYE with a non-sparse phantom image which is used in image processing [49]. The phantom image in Fig. 6(a) consists of 160×160 pixels, each of which contains an 8-bit intensity of light. SR of the phantom image is 0.4928. The number of ommatidia is set to be 120×120 with $\Delta\phi = 1.5^\circ$ and $\Delta\phi = 1.5^\circ$ for the conventional compound eye and $\Delta\phi = 60^\circ$ and $\Delta\phi = 1.5^\circ$ for COMPU-EYE. The object image size of 60×60 mm is at a distance of 10 mm from the compound eyes. In the reconstruction of the image, DCT is used for a sparsifying basis. As seen in Fig. 6(b), the direct observation of the conventional compound eye provides poor resolution and the object is distorted. Figure 6(c) shows the reconstructed image by COMPU-EYE. The resolution is improved by a factor of $N/M = 1.78$. We note that the distortion comes from a discrepancy in receptive fields of ommatidia, i.e., as an ommatidium is closely located to the edge of the compound eye, its corresponding receptive field becomes larger. In contrast, the reconstructed image of COMPU-EYE in Fig. 6(c) is not distorted because COMPU-EYE recovers the designated pixel values \mathbf{x} in the object. As a result, COMPU-EYE can also reconstruct the non-sparse object image with a high resolution.

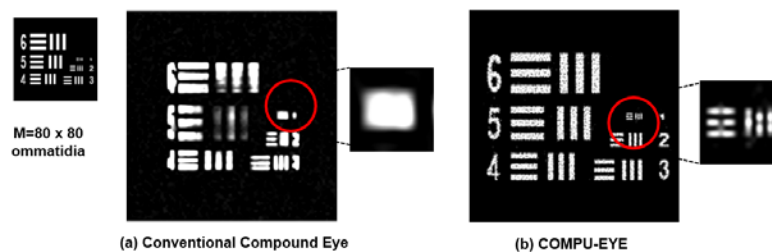


Fig. 7. Resolution test: (a) Conventional compound eye consisting of 80×80 ommatidia with $\Delta\phi = \Delta\phi = 2.25^\circ$. (b) COMPU-EYE consisting of 80×80 ommatidia with $\Delta\phi = 60^\circ$ and $\Delta\phi = 2.25^\circ$.

Figure 7 illustrates optical resolution tests of the conventional compound eye and COMPU-EYE. The 60×60 mm object image at a distance of 10 mm is composed of 160×160 pixels. The object image is a target image similar to the US Air Force (USAF) test, where the minimum spacing of gratings is a single pixel, i.e., 0.375 mm. The lines of the row labeled

“1” have single pixel spacing, those of the row labeled “2” have two-pixel spacing, and so on. Both compound eyes are composed of 80×80 ommatidia with $\Delta\phi = 2.25^\circ$ and $\Delta\varphi = 2.25^\circ$ for the conventional compound eye and with $\Delta\phi = 2.25^\circ$ and $\Delta\varphi = 60^\circ$ for COMPU-EYE. Because the achievable optical resolution of the conventional 80×80 compound eye with $\Delta\varphi = \Delta\phi = 2.25^\circ$ is 0.7179×0.7179 mm, which is obtained from the distance of resolvable gratings in the object plane, it cannot distinguish the smallest grating as shown in Fig. 7(a). However, COMPU-EYE can sharply resolve the smallest grating because the resolvable resolution is the unit of a single pixel. Thus, the achievable minimum optical resolution of COMPU-EYE is 0.375×0.375 mm, an improvement in resolution of about 3.66 times. We note that the observation at the center of the conventional compound eye in Fig. 7(a) suffers from lack of incoming light due to the relatively small sized receptive fields and its resulting undetectable area.

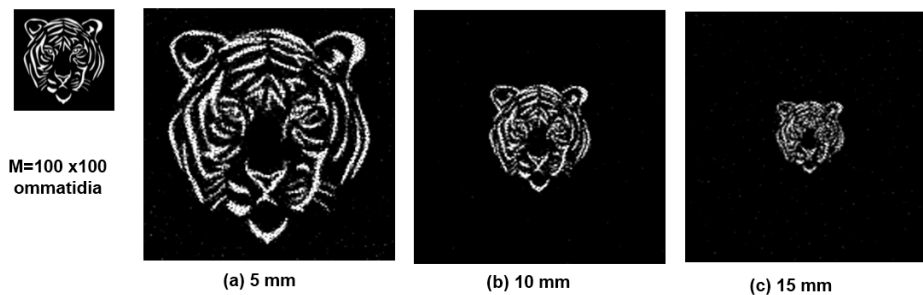


Fig. 8. Depth test: Image recovered by COMPU-EYE consisting of 100×100 ommatidia with $\Delta\varphi = 60^\circ$ and $\Delta\phi = 1.8^\circ$, where the dimension of the final object image is (a) 30×30 mm at 5 mm, (b) 60×60 mm at 10 mm, (c) 90×90 mm at 15 mm. The actual tiger picture is 30×30 mm.

Figure 8 shows the image recovered by the proposed COMPU-EYE at various object image distances. The size of the visible area of the compound eye is proportional to the distance of the object image, and the measurement matrices are generated according to the distances of the object image. Given the measurement matrices at distances of 5, 10, and 15 mm, the image can be reconstructed from \mathbf{y} . As seen in Fig. 8, the recovered images are still clear and focus is maintained as the object image moves away from the compound eye.

5. Summary

In this study, we proposed the COMPU-EYE imaging system to improve the resolution of compound eyes. COMPU-EYE uses ommatidia with acceptance angles that are larger than the interommatidial angle as well as a DSP technique. By increasing the acceptance angles, each ommatidium covers wider areas, and each observation is different from the others because of its receptive field. Finer details can be resolved by the DSP technique. As a result, the proposed COMPU-EYE provides at least a four-fold improvement in resolution.

Natural compound eyes have the ability to detect high-speed motion owing to the simple ON/OFF detection structure of the ommatidium. In contrast, COMPU-EYE views the object only through computation and it necessarily requires certain computation time and cost for imaging. The computation requires solving a convex optimization problem; this problem can be solved in polynomial time by many state-of-the-art algorithms including YALL1 [43], FISTA [44], and CP [45]. Thus, the additional computation time required for the compound eyes is practically feasible with modern DSP devices. For example, when we measure the computation time using MATLAB with a 3.6-GHz Intel i7 processor, it takes 47 ms to recover $N = 256$ pixels from $M = 64$, as shown in Fig. 2(b). We note that the computation time can be reduced by using a multicore processor or graphic processing unit because the

algorithms [43–45] conduct matrix multiplications and additions, and these operations can be computed in parallel [50].

Generally, the acceptance angles are proportional to the light sensitivity of ommatidia. But, the large acceptance angles cause overlapping among neighboring ommatidia and necessarily result in low spatial resolution. By resolving the aliasing caused by the overlapping using a DSP technique, COMPU-EYE is expected to have high sensitivity with high resolution. Moreover, the technique for resolution improvements used in COMPU-EYE can be applied to other designs of artificial compound eyes. It would be interesting to compare resolution of Curvace design in [18] consisting of more ommatidia and the hemispherical compound eye in [8] consisting of less ommatidia but equipped with the DSP technique. In this paper, we have focused on the apposition compound eye. But, we note that the concept of COMPU-EYE can also be applied to other types of compound eyes, i.e., superposition compound eyes. For example, in the neural superposition compound eyes which are specialized for light sensitivity, each object point is imaged by multiple photoreceptors from different ommatidia and the related signals are combined to form an image with high sensitivity and high resolution [21]. By applying the design concept of larger acceptance angles and the DSP technique, the neural superposition compound eyes can improve the resolution and sensitivity. In the real implementation of compound eye devices, COMPU-EYE is more efficient in terms of multiple observations. If some ommatidia are disjointed or damaged, the conventional compound eye could lose vision in the corresponding area. However, in COMPU-EYE, each area is observed by multiple ommatidia. Thus, even though some ommatidia are lost, they do not have a significant influence on the overall observation.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the South Korean government (NRF-2015R1A2A1A05001826).



Simple adaptive sparse representation based classification schemes for EEG based brain–computer interface applications



Younghak Shin^a, Seungchan Lee^a, Minkyu Ahn^b, Hohyun Cho^a, Sung Chan Jun^a, Heung-No Lee^{a,*}

^a School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

^b Department of Neuroscience, Brown University, RI, USA

ARTICLE INFO

Article history:

Received 27 May 2015

Accepted 24 August 2015

Keywords:

Electroencephalogram (EEG)
Brain–computer interface (BCI)
Sparse representation based classification (SRC)
Common spatial pattern (CSP)
L1 minimization
Non-stationarity

ABSTRACT

One of the main problems related to electroencephalogram (EEG) based brain–computer interface (BCI) systems is the non-stationarity of the underlying EEG signals. This results in the deterioration of the classification performance during experimental sessions. Therefore, adaptive classification techniques are required for EEG based BCI applications. In this paper, we propose simple adaptive sparse representation based classification (SRC) schemes. Supervised and unsupervised dictionary update techniques for new test data and a dictionary modification method by using the incoherence measure of the training data are investigated. The proposed methods are very simple and additional computation for the re-training of the classifier is not needed. The proposed adaptive SRC schemes are evaluated using two BCI experimental datasets. The proposed methods are assessed by comparing classification results with the conventional SRC and other adaptive classification methods. On the basis of the results, we find that the proposed adaptive schemes show relatively improved classification accuracy as compared to conventional methods without requiring additional computation.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Brain–computer interface (BCI) systems provide a new communication and control channel between human brain and an external device without any muscle movements [1]. Due to the convenient usability and high temporal resolution compared to other brain imaging equipment such as functional magnetic resonance imaging (fMRI) and magnetoencephalogram (MEG), research of noninvasive electroencephalogram (EEG) based brain–computer interface (BCI) systems is continuously progressed [1–3].

In the beginning of BCI research, BCI systems have been developed mostly to provide alternative communication means to people who have severe motor disabilities [2,4,5]. Recently, much research effort focused on development of portable BCI systems for normal person by using headset shaped scalp electrodes [6,7] and also dry electrodes which do not need conductive gel for preparation of EEG recording [8,9]. In addition, with the progress of portable BCI systems and EEG sensor technologies, many BCI applications are developed for general public [9,10]. However, for the BCI systems going beyond laboratory researches, the most important issue is stable classification performance.

Normally, EEG based BCI experiment can be categorized as a training (calibration) stage and a real time testing (feedback) stage. In the training stage, translation algorithm such as classification is designed using collected training signals. Then, an application device such as neural prosthesis is controlled by using the classification algorithm in real time testing stage. However, EEG signals have inherent non-stationary characteristics and there exist significant day-to-day and even session-to-session variability [12,27,29]. Thus, features of experimental EEG signals are changed from the offline training sessions to online testing sessions [11]. Due to this, classification performance is unavoidably deteriorated in BCI experiment with time. In addition, the training session (15–35 min) is conventionally carried out every time before using the BCI systems even for experienced subjects [12]. These are major obstacles of real-time online BCI applications.

To overcome the performance decrease caused by the non-stationarity of EEG signals, many adaptive signal processing methods are proposed. In [27–29], adaptive feature extraction methods are proposed for the motor imagery based BCI systems. For the adaptive classification scheme, in [13], mean and covariance matrix of a statistical classifier are iteratively updated using each class data. The study [11] proposes a bias adaptation scheme of linear discriminant analysis (LDA) classification using class labels of several test trials. They have shown that simple bias

* Corresponding author. Tel.: +82 62 715 2237; fax: +82 62 715 2204.
E-mail address: heungno@gist.ac.kr (H.-N. Lee).

adaptation is effective for online test data. In [14], they propose an expectation-maximization (EM) algorithm based unsupervised adaptive classification method. Using EM algorithm, common spatial pattern (CSP) features are re-extracted and parameters of Bayes classifier are updated in each iteration step. Similarly, [15] suggest unsupervised bias adaptation of LDA without using class label information. Previous studies for adaptive classification method need classifier re-adjustment (training) such as parameters and bias adaptation for new test trials. However, for this re-training, additional computation is needed in each update (adjustment) step.

Recently, with much progress of L1 minimization technique in compressive sensing field [21,22], sparse representation has received a lot of attention in signal processing and pattern recognition fields. Especially, sparse representation based classification (SRC) has shown an increased interest [16,23,24]. SRC framework is first introduced by Huang et al [16]. A test data from one class is predominantly represented by the same class training data from dictionary. The dictionary is composed by all class training data and usually underdetermined. Sparse representation of the test data using the dictionary can effectively be solved by the L1 minimization tool, and the classification is performed by comparing the representation error for each class.

SRC have been also studied for EEG signal classification [17,18,25]. In [18] and [25], SRC scheme is applied to vigilance detection and epileptic seizure detection problem respectively. In addition, SRC scheme is first introduced for motor imagery based BCI application in [17]. They have shown that the SRC exhibits better classification performance than the conventional LDA method using two experimental datasets. Another study [31] also revealed that the SRC shows better classification accuracy and noise robustness than the well-known SVM method. However, no research has been studied for adaptive SRC scheme for online BCI applications.

Compared to other fixed decision rule based classification method such as linear discriminant analysis (LDA) and support vector machine (SVM), in the SRC, the sparse representation is adaptively performed for each test data by utilizing all training data in the dictionary. Along with this inherent adaptive characteristic of the SRC, in this study, we propose simple adaptive SRC schemes for real-time BCI applications. We suggest a dictionary update rule and an incoherence based dictionary modification (IDM) method. For the dictionary update rule, supervised and unsupervised adaptive schemes and also accumulated and fixed update rules are considered. Proposed dictionary update methods are very simple and additional computation for adaptation is not needed. In the part of IDM method, our aim is to create a maximally incoherent dictionary via an incoherence measure of training data. This method is applied to the training data before performing the sparse representation. Using online motor imagery based BCI experimental datasets, we evaluate classification performance of the proposed adaptive method by comparing with the conventional SRC and other adaptive classification methods.

This paper is organized as follows. In Section 2, our experiment and dataset are explained. In Section 3, technical methods such as feature extraction, sparse representation based classification (SRC) method and proposed adaptive SRC schemes are introduced. We explain experimental evaluation strategy and results in Section 4. In Section 5, we discuss some experimental results. Finally, we conclude the paper in Section 6.

2. Experiment

For evaluation of adaptive classification scheme, we performed online motor imagery based BCI experiment. The experiment was

approved by the Institutional Review Board of Gwangju Institute of Science and Technology. Ten subjects who signed a written informed consent letter participated in our online experiment. The experiment was performed on multiple days (two or three days). In each day, just one session experiment was executed. The number of sessions for each subject was determined by classification results and condition of each subject. Right hand (R), left hand (L) and foot (F) motor imagery were performed for each subject. For this experiment, we used Active Two EEG measurement system made by Biosemi, Inc. The sampling rate of these datasets was 512 samples per second and the number of EEG channels was 64. The channel positions were selected from international 10/20 standard.

The detailed experimental paradigm was illustrated in Fig. 1. The same paradigm was used for both training (calibration) and online testing (feedback) phases. In the training phase, one session consisted of three runs and one run consisted of 20 trials for each class. Thus, we collected a total of 60 training trials for each class. All participants were naïve subjects for this motor imagery experiment. Therefore, it was difficult to achieve satisfactory classification performance without sufficient training time. In addition, each subject had a different discrimination potential for a different pair of motor imagery signals. In this study, to find the most discriminative motor imagery pair for each subject, we performed the initial classification for all pairs of (R), (L), and (F) by using the dataset of the first run in the training phase. The best pair of motor imagery was selected using the CSP feature with the LDA classifier and used for a further experiment in the training and testing session. As shown in Fig. 1, in each trial, the target bar was represented on 0 s at left, right or down side of monitor screen corresponding to the left, right or foot motor imagery. On 2 s after cue onset, subject was instructed to perform the motor imagery task. Then, subject imagined their left, right hand or foot movement such as grasping and releasing hand. In this period, subject was also instructed to stare a red dot during motor imagery to avoid eye movement artifacts. In the training session, to design a classifier that would be used in the testing session, we just collected the training trials for each motor imagery signal. At that time, the classifier had not been designed. Therefore, the yellow ball (feedback) was set to move into the target direction automatically.

In the online testing (feedback) phase, same experimental paradigm was used. However, the online feedback was provided in each trial. Thus, the yellow ball was controlled by the classified result which was analyzed from intention of each subject using the EEG data collected from 2 to 4 s. We recorded 75 test trials for each class. One run consisted of 25 trials and we performed total three runs. Thus, in the one session experiment, total 60 offline and 75 online trials per class were collected for each subject. Both data were segmented from 2 to 4 s after cue onset for further signal processing.

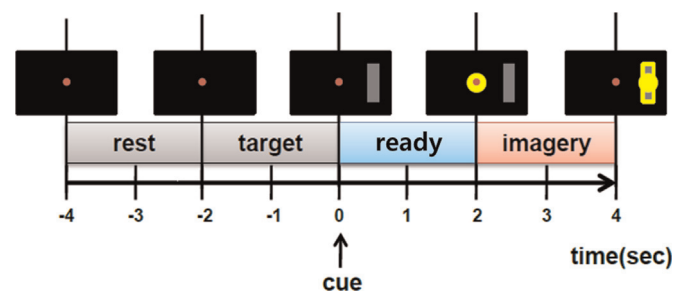


Fig. 1. One trial experimental paradigm for motor imagery experiment.

3. Methods

3.1. Preprocessing and feature extraction

For preprocessing of experimental EEG dataset, we apply same procedures to all datasets and classification methods. First, we perform band pass filtering to eliminate the frequencies which are not related to motor imagery signals. In this study, we use fourth order Butterworth filter with 5 and 30 of cut off frequencies.

EEG signals are very noisy and have poor spatial resolution. Thus, an electrode placed on the scalp measures the EEG signals generated not only from the motor cortex area but also from other cortical regions. Therefore, it is important to find maximally discriminative information from the original high-dimensional data. For this purpose, we perform common spatial pattern (CSP) filtering. The CSP filtering is a well-known feature extraction method for two-classes motor imagery dataset [12,17,19]. The CSP filtering algorithm finds the filters $\mathbf{W} \in \mathbb{R}^{C \times C} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C]$ which transforms the EEG data $\mathbf{X} \in \mathbb{R}^{C \times S}$ (C and S denote the number of EEG channels and time samples) into a spatially filtered space: $\mathbf{X}_{CSP} = \mathbf{W}^T \cdot \mathbf{X}$. Generally, \mathbf{W} is computed by simultaneous diagonalization of the covariance matrices, Σ_1 and Σ_2 , of the two classes data. This is equivalent to solving the generalized eigenvalue problem, i.e., $\Sigma_1 \mathbf{w} = \lambda \Sigma_2 \mathbf{w}$, where λ is eigenvalue. In practice, first and last k columns of the \mathbf{W} corresponding to the k largest and k smallest eigenvalues are used for CSP filtering. For fair comparison, we set the k equal to five for all our datasets in this study. The obtained CSP filters maximize the variance of the spatially filtered signal for one class data while minimizing it for the other class data. Detailed information about the CSP filtering algorithm can be found in [17,19]. After CSP filtering, we compute the band power (BP) of sensorimotor rhythm (8–15 Hz). BP is the power of the signal within specific frequency bands. Because of the physiological background of the motor imagery signals, ERD based band power (BP) of the sensorimotor rhythm is a well-known feature form in many EEG based BCI studies [12,17,20].

3.2. Sparse representation based classification

In this paper, based on the sparse representation classification (SRC) scheme we propose adaptive SRC methods. Therefore, in this section, we simply introduce conventional SRC framework. We also use the SRC method to provide a baseline classification result for this study to compare results of the proposed adaptive SRC methods.

In [17], we propose a SRC scheme for motor imagery based BCI applications. In the SRC framework, if training samples in a dictionary is sufficiently large, a test sample can be sparsely represented with same class training samples over the dictionary. The SRC method can be categorized as sparse coding step and identification step. The sparse coding step is formulated as $\mathbf{y} = \mathbf{A}\mathbf{x}$. Where, \mathbf{y} and \mathbf{A} indicate a test feature vector and a collection of training feature vectors. Also, \mathbf{x} is an unknown coefficient vector. \mathbf{A} is called a dictionary formed by class-dictionary $\mathbf{A}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \dots, \mathbf{a}_{i,N_i}]$, where $i = 1, 2, \dots, C$ represents class information and N_i denotes the number of training trials for class i . In this study, C is equal to 2. $\mathbf{a}_{ij} \in \mathbb{R}^{m \times 1}$ is the j -th training feature vector of dimension $m=2k$ from the class i . In this study, each element of \mathbf{a} is the band power feature of the CSP filtered data. The dictionary \mathbf{A} is formed by $\mathbf{A} = [\mathbf{A}_1; \mathbf{A}_2] \in \mathbb{R}^{m \times N}$, where N denotes the total number of training trials. Thus, in this study, $N = 2N_i$ for two class problems.

In the SRC algorithm, first, the columns of dictionary \mathbf{A} are normalized to have a unit L2 norm. Then, in the sparse coding step, unknown coefficient vector \mathbf{x} can be recovered by solving following optimization problem via L1 norm minimization tool:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad (1)$$

Note that equation (1) is an under-determined system. The literature of compressive sensing (CS) shows that the L1 norm minimization algorithm can solve this optimization problem effectively in polynomial time [21,22]. Using the recovered coefficient vector \mathbf{x} by L1 minimization, class identification is performed as follows:

$$\text{class}(\mathbf{y}) = \min_i r_i(\mathbf{y}), \quad (2)$$

where $r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}_i \mathbf{x}_i\|_2$ is representation residual corresponding to the class i . Thus, we identify the class of the test sample \mathbf{y} as i when residual $r_i(\mathbf{y})$ is minimal.

3.3. Adaptive SRC schemes

To overcome inherent non-stationarity of EEG signals, we propose simple adaptive classification schemes based on the SRC method. In this study, we suggest two schemes, dictionary update method and incoherence based dictionary modification (IDM) method. Each scheme works with the conventional SRC method independently. In addition, both schemes can be incorporated as one combined adaptive SRC method. In the following subsections, we introduce each adaptive scheme.

3.3.1. Incoherence based dictionary modification method

Previous SRC studies for motor imagery based EEG classification [17] have revealed that when a dictionary is incoherent, a test signal from one particular class can be predominantly represented by the columns of the same class in the dictionary. The uncertainty principle (UP) [30] in the sparse representation theory dictates that a signal cannot be sparsely represented in both classes simultaneously. This phenomenon intensifies as the degree of incoherence of the dictionary increases. An incoherent dictionary can be explained from the definition of mutual coherence of class-dictionary. The coherence measures the correlation between the two class-dictionaries defined as following:

$$C(\mathbf{A}_L, \mathbf{A}_R) \triangleq \max \left\{ \left| \langle \mathbf{a}_{L,j}, \mathbf{a}_{R,k} \rangle \right| : j, k = 1, 2, \dots, N_i \right\}, \quad (3)$$

The vector $\mathbf{a}_{L,j}$ and $\mathbf{a}_{R,k}$ are the j -th column of \mathbf{A}_L and the k -th column of \mathbf{A}_R respectively. The notation $\langle \mathbf{a}_{L,j}, \mathbf{a}_{R,k} \rangle$ denotes the inner product of the two vectors. We call C the measure of mutual coherence of two class-dictionaries. In the SRC algorithm, we normalize the columns of dictionary \mathbf{A} . Therefore, C measures the smallest angle between any pair of columns of two classes. When the value of C obtained from the two class-dictionaries is small, i.e., the cosine angle between two columns is large, we consider the dictionary incoherent. Due to the characteristics of the CSP filtering, i.e., CSP filters maximize the variance of the spatially filtered signal for one class data while minimizing it for the other class data, the CSP features can be used for constructing incoherent dictionary [17]. After applying CSP filtering, in the proposed IDM method, we aim to eliminate some training trials that have a high average cross coherence value with training trials of a different class. Thus, the eliminated training trials have features similar to those of many training trials of a different class. Therefore, we expect to further increase the incoherence of the dictionary by using the IDM method; this might lead to a high discrimination capability for training trials of two different classes.

In the IDM method, coherence value of the dictionary \mathbf{A} can be simply estimated by each element of $\mathbf{G} = \mathbf{A}^T \mathbf{A}$. Thus, $\mathbf{G}(i, j)$ indicates the coherence value between i and j -th column of the dictionary. Therefore, $\mathbf{G}(i, j)$ is equal to $\mathbf{G}(j, i)$. For example, if the number of training trials of each class-dictionary is five, then the

	1	2	3	4	5	Avg.
6	9	1	8	1	1	4
7	1	8	2	2	1	2.8
8	3	2	9	9	9	6.4
9	2	1	2	2	2	1.8
10	2	9	8	1	1	4.2
Avg.	3.4	4.2	5.8	3	2.8	

Eliminate 3rd trial of A_1

Eliminate 8th trial of A_2

Fig. 2. Example of incoherence based dictionary modification (IDM) method.

dimension of \mathbf{G} is 10×10 . From the \mathbf{G} , we focus on the cross coherence part between the two classes. Thus, we extract columns from 1-th to 5-th and rows from 6-th to 10-th of the \mathbf{G} which are corresponding to the class 1 and class 2 respectively. Therefore, the dimension of cross coherence part is 5×5 in this example. We represent this cross coherence part as \mathbf{G}_{cc} . Using the \mathbf{G}_{cc} , we can easily check which trials of class 1 dictionary have large coherence values with trials from class 2 dictionary and vice versa.

Fig. 2 shows example values of cross coherence $\mathbf{G}_{cc} \in \mathbb{R}^{5 \times 5}$ and concept of the IDM method. In this figure, each number means the coherence value ranged from 1 to 9. Red colored elements represent high coherence values which are set up to be the values greater than or equal to 8. The values of last row and column represent the averaged value of five columns and rows respectively. In this example, we set the number of elimination trials n equal to one. Thus, we aim to eliminate the highest average value for each column and row respectively.

From the averaged value of cross coherence, the third row and third column shows highest averaged value of 6.4 and 5.8. This means that 8-th row (8-th trial from class 2 dictionary) and third column (third trial from class 1 dictionary) shows high coherence value with many trials, i.e., many red colored elements, from the other class-dictionary. Therefore, we can eliminate the one trial in the each class-dictionary.

We summarize the incoherence based dictionary modification (IDM) algorithm as follows:

1. Set n the number of elimination trials.
2. Compute the average value of each column of \mathbf{G}_{cc} .
3. Collect the indices of column numbers which have n highest average coherence values.
4. Eliminate n indices from original class-dictionary.
5. Repeat 2–4 steps for row of \mathbf{G}_{cc} .

For each subject dataset, we apply the IDM algorithm to the dictionary. After then, we perform the SRC steps with the modified dictionary.

3.3.2. Dictionary update methods

Normally, in motor imagery based BCI systems, a translation algorithm such as a classifier is designed using the collected training data. Then, an application device or program is controlled by using the classification algorithm in each test trial. However, because of the inherent non-stationarity of EEG, the classification performance deteriorates from the training to the test session in a BCI experiment. To overcome this drawback, many adaptive classification schemes are proposed. The main concept of the adaptive

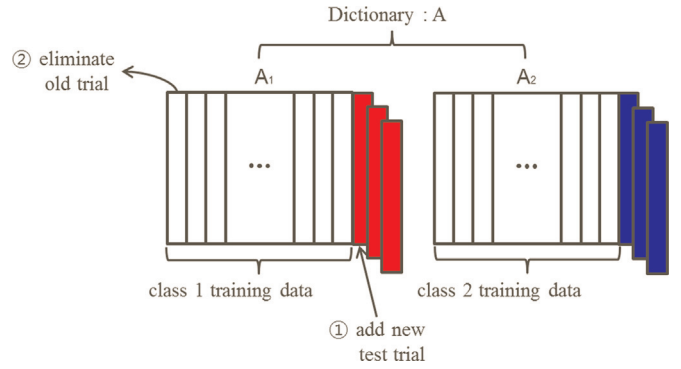


Fig. 3. Concept of the proposed dictionary update rule.

classification is re-adjustment (re-training) of the classifier for the new test data. On the other hand, in the SRC scheme, one important characteristic is that training (or parameter decision) of a classifier is not needed unlike in other decision hyper-plane based classification methods such as LDA and SVM [31]. Thus, in the SRC scheme, a dictionary is simply formed by collecting the training feature vectors as columns of the dictionary. Then, using the dictionary sparse coding step is performed for each test data. Due to this unique classification mechanism, a simple intuitive method for adaptive SRC is dictionary update.

As we mentioned in Section 3.2, the dictionary \mathbf{A} is formed by class-dictionary $\mathbf{A}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \dots, \mathbf{a}_{i,N_i}]$ in the SRC method. Each column vector \mathbf{a}_{ij} is a j -th training feature vector of class i . Therefore, for each test trial in the online testing phase, a feature vector of a new test trial \mathbf{y} can be easily updated as a new column of the dictionary. Then, characteristics of the test feature can be applied into the dictionary while the online testing experiment is performed. And therefore, we can expect the classification performance of the online testing phase is not deteriorated.

In this study, we consider four types of dictionary update rule, supervised accumulated update (SAU), supervised fixed update (SFU), unsupervised accumulated update (UAU) and unsupervised fixed update (UFU) rule. In our online experimental paradigm, as shown in Fig. 1, a target class label is first provided as the position of the target bar. Then, subjects perform motor imagery corresponding to the class label information for each trial. In the supervised update rule, the target class label of test trials is used for updating the online test trials. Thus, a new test trial which has same class label of training trials in the class-dictionary is updated into the corresponding class-dictionary. However, this strategy is not practical for a general online scenario. Therefore, we also consider the unsupervised update rule. In the unsupervised update rule, class label information of the test trial is not used. Thus, each test trial is updated into the corresponding class-dictionary based on the estimated result of the current classifier, which is represented by the direction of the yellow ball movement shown in Fig. 1.

For the case of accumulated update method, as shown in ① of Fig. 3, all updated test trials are just stacked at the end (last column) of the class-dictionary based on the class label and classified result for SAU and UAU respectively. However, for the case of fixed update rule, SFU and UFU, the oldest training trial, i.e., the first training trial of the class-dictionary is eliminated as shown in ② of Fig. 3 when each new test trial is updated. Note that if available training data in the dictionary is large enough and online testing phase is long, i.e., the number of test trials is large; the dictionary will be a fat matrix in the case of accumulated update rule. In this case, computation time for sparse representation is also increased. Therefore, in this study, we consider fixed update rule which has a same size dictionary, i.e., number of columns in the dictionary,

with the original training dictionary. We compare computation time between accumulated and fixed update rule in Section 6.2.

4. Results

4.1. Evaluation strategy

Using the online experimental dataset, we aim to evaluate proposed adaptive SRC schemes, i.e., four dictionary update methods (supervised accumulated update (SAU), supervised fixed update (SFU), unsupervised accumulated update (UAU) and unsupervised fixed update (UFU) rule) and an incoherence based dictionary modification (IDM) method. From the multi session datasets of 10 subjects, 12 session datasets are selected for evaluation of proposed methods. In this selection, for a reliable assessment of classification methods, we choose datasets over 60% classification accuracy in the online experiment (in the binary classification, theoretical random chance level is 50%). Each session dataset consists of 60 training trials and 75 test trials for each class.

In this study, for the two class classification problems of the conventional SRC method, the dimension of the dictionary \mathbf{A} is 10×120 , i.e., $m=10$ CSP features and $N=120$ training trials. For each subject, 150 test trials where each has the same 10 dimension features are evaluated with dictionary \mathbf{A} . For the proposed adaptive methods, we perform the incoherence based dictionary modification (IDM) method using the original dictionary \mathbf{A} . After then, for each new test trial, we perform the each proposed dictionary update method for adaptation of test data.

Due to the inherent non-stationarity of EEG signals, online test data have different feature characteristics compared to training data [11,26,27]. And therefore, even though classifier is well trained for training data, satisfactory classification performance is not guaranteed for online data. We expect that in the SRC method the proposed incoherence based dictionary modification (IDM) method is effective for proper dictionary design by maximizing incoherence between two classes. In addition, to overcome the non-stationarity of EEG, new test features will be applied into the original dictionary using updated new test trials from the proposed dictionary update method. Using online experimental dataset, we evaluate classification accuracy of the conventional SRC, each dictionary update method and IDM based adaptive SRC method. In addition, we also compare the classification results of

the proposed methods with other adaptive classification methods such as adaptive LDA and SVM method.

4.2. Experimental results

To evaluate classification performance of the proposed adaptive SRC schemes, we compare classification accuracy (%) of proposed methods with that of conventional SRC method using the online experimental dataset of 12 motor imagery sessions. Table 1 shows the classification accuracy of the SRC and the proposed dictionary update based SRC methods with and without IDM method. For fair comparison, we set the same value of n (the number of elimination trials of IDM) of 10 for all subjects and all IDM based adaptive SRC methods.

From the results of Table 1, all five methods with IDM show better mean classification accuracy than the without IDM method. Thus, the proposed IDM method is effective for the SRC framework. Furthermore, the proposed simple dictionary update methods with and without IDM show improved mean classification accuracy than the conventional SRC method. Supervised update methods, i.e., SAU and SFU, show more improved results than the unsupervised methods, UAU and UFU. However, mean difference between SAU/ SFU with IDM and UAU/ UFU with IDM is not much.

For further analysis, in Fig. 4, we investigate the comparison of the classification accuracy of 12 datasets using scatter plots. Each point indicates the classification accuracy of each dataset which is used for computing mean classification accuracy in Table 1. Fig. 4 left shows the comparison results between the SRC and the two supervised dictionary update methods with IDM. Classification accuracies of the SRC and supervised methods are represented in X and Y-axis respectively. For the supervised methods (Y-axis), blue square points indicate the SAU with IDM method and red circle points indicate the SFU with IDM method. Similarly, Fig. 4 right shows the comparison results between the SRC and the two unsupervised dictionary update methods.

From the results of Fig. 4 left, both SAU and SFU with IDM show higher classification accuracies than the SRC method for eleven datasets. Thus, the 11 data points positioned over the black linear-line which indicates the same classification accuracy between SRC and proposed methods. On the right figure, we also observe that the both UAU and UFU IDM show higher classification accuracies than the SRC for 10 datasets. In addition, p -values obtained from

Table 1

Classification accuracy of conventional SRC and proposed adaptive SRC schemes (SRC_SAU, SRC_SFU, and SRC_USU) for 12 session datasets. We present the classification accuracy (%) of each method with and without IDM. The highest classification accuracy for each dataset is highlighted in bold.

Dataset	SRC		SRC_SAU		SRC_SFU		SRC_UAU		SRC_UFU	
	w/o IDM	w/ IDM	w/o IDM	w/ IDM	w/o IDM	w/ IDM	w/o IDM	w/ IDM	w/o IDM	w/ IDM
1	66	66.7	67.3	70.7	66.0	64.7	66.0	67.3	66.0	67.3
2	86	86.7	88.0	88.0	88.0	88.0	87.3	89.3	82.7	90.7
3	88.7	90.7	90.0	90.0	89.3	90.7	90.0	90.7	90.7	88.7
4	96.4	96.4	96.4	96.4	97.1	97.1	96.4	96.4	96.4	96.4
5	83.3	89.3	93.3	96.0	96.0	96.7	93.3	95.3	94.7	97.3
6	82.7	78.7	86.7	86.7	84.0	84.0	80.0	84.0	80.7	83.3
7	77.3	75.3	78.0	80.0	78.7	79.3	76.7	77.3	79.3	78.0
8	73.3	88.0	88.7	88.7	89.3	91.3	78.0	89.3	84.7	90.7
9	70.0	75.3	74.0	74.7	73.3	74.0	70.0	72.0	70.0	71.3
10	62.0	64.0	66.0	68.7	67.3	71.3	62.0	63.3	68.0	66.7
11	84.0	87.3	88.7	89.3	88.7	89.3	86.7	88.0	88.0	88.7
12	96.7	96.0	97.3	98.0	97.3	98.0	96.7	98.0	96.7	98.0
Mean	80.5	82.9	84.5	85.6	84.6	85.4	81.9	84.3	83.1	84.8
Std.	11.13	10.74	10.69	9.94	10.99	10.89	11.73	11.64	10.84	11.40

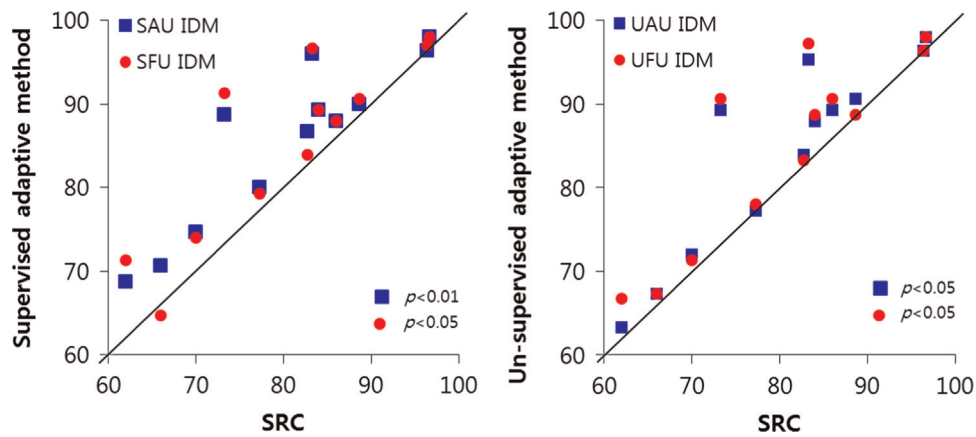


Fig. 4. Comparison of classification accuracy of all 12 datasets. (Left): Scatter plot of classification accuracies between conventional SRC (X-axis) and the both supervised update methods SAU and SFU with IDM (Y-axis). (Right): Scatter plot of classification accuracies between conventional SRC (X-axis) and the both unsupervised update methods UAU and UFU with IDM (Y-axis).

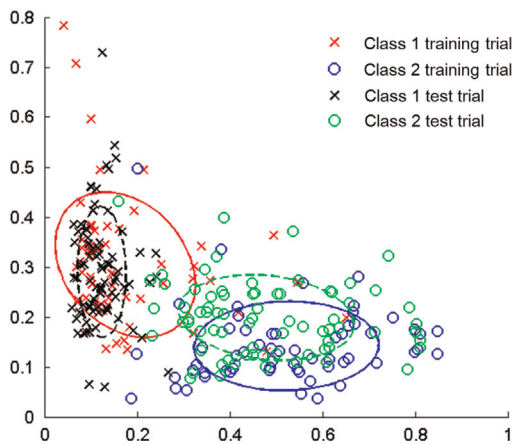


Fig. 5. Scatter plot of training and test features for two different classes in two dimensional feature spaces using an example dataset 5. All training and test samples are scattered and fitted by Gaussian distribution. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

paired *t*-test are smaller than 0.05 for all comparisons between the SRC and proposed methods in Fig. 4.

To evaluate the effect of the proposed methods, we analyze one dataset in the feature space. Fig. 5 shows scatter plots of training and test features of dataset 5 used in Table 1. For ease of visualization, we use two-dimensional feature spaces which are corresponding to the first and the last CSP filters. In Fig. 5, the red and black x marks indicate the 60 training and 75 test features for one class, respectively. On the other hand, the blue and green circles indicate the 60 training and 75 test features for another class, respectively. Each class training and test data element is fitted by a Gaussian distribution. Therefore, we can easily check the distribution change from the training to the test data during the experimental sessions. When the distribution of the test data is changed from that of the training data, the previously designed dictionary based on the training data is not optimal for the classification of new test data.

Fig. 6 shows one classification instance of a test trial, which is represented by a filled green point (class 2) in the left figure. In this test, the test feature is not correctly classified, i.e., classified as class 1, by the conventional SRC without IDM method. All training features in the dictionary of classes 1 and 2 shown in Fig. 5 are utilized for the classification of the test feature without the use of any adaptation techniques. Fig. 6 right shows the coefficients

recovered by the conventional SRC for the test feature represented in the left figure. The X-axis represents the training trial number (column number) of the dictionary, and the red dotted line denotes the boundary of two different classes. In the right figure, the numbering ①, ② and ③ represent the coefficients corresponding to the training trials of black x marks ①, ② and ③ in the left figure. Because the three training points of class 1 are used for the sparse representation of the test trial and have large coefficient values, the test feature is classified as class 1 by using the minimum residual rule in Eq. (2).

On the other hand, Fig. 7 shows the classification results of SRC_UAU IDM for the same test trial used in Fig. 6. In Fig. 7 left, we can see that some training features which are originally positioned at the area of different class features including the black x marks ①, ② and ③ in Fig. 6 left are effectively eliminated by the IDM method. In addition, new test trials represented by the black x marks and the green and black circles are also updated before the classification of the current test trial, which is represented by the filled green circles. From the result of Fig. 7 right, we conclude that the test trial is correctly classified as class 2 and the three updated test trials represented by black circles ①, ② and ③ in the left figure have large coefficients. Therefore, for the classification of new test trials, IDM and the dictionary update method in SRC are very effective, and we can see that the proposed methods with IDM show relatively improved classification accuracy compared to the conventional SRC from the results of dataset 5, presented in Table 1.

In Table 2, we compare the classification accuracy of the conventional SRC and the proposed adaptive SRC methods with the non-adaptive and adaptive LDA and SVM classification methods using our experimental dataset. The LDA and SVM are widely used classification methods in many EEG based BCI researches [26]. For the adaptive LDA and SVM methods, first, linear decision hyper-plane is chosen from training data. Then in the testing session, the decision hyper-plane is re-trained for new test sample as shown in [11]. We only consider supervised adaptation for the LDA and SVM methods.

From the results presented in Table 2, we can first see that the conventional SRC exhibits better mean classification accuracy than the non-adaptive LDA and SVM methods. These results are consistent with those of the previous studies [17,31] mentioned in Section 1. Second, the proposed adaptive SRC methods show better mean classification accuracy than the other adaptive LDA and SVM methods. Note that even though the accuracy difference between the unsupervised adaptive SRC methods and adaptive SVM method is not much, in the conventional adaptive methods,

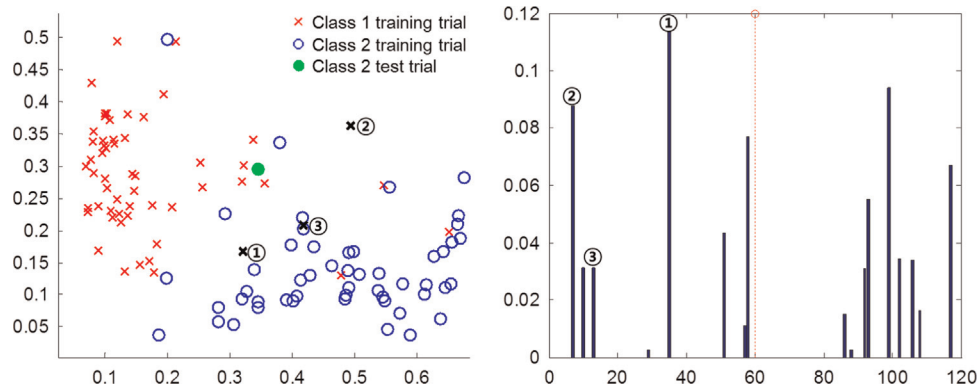


Fig. 6. Classification results of conventional SRC for one test sample of dataset 5. (Left): Scatter plot of training features for two classes and one test feature of class 2. (Right): Sparse representation results of one test feature shown in left figure from the conventional SRC. X-axis represents the training trial number in dictionary and Red dotted line means the boundary of two different classes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

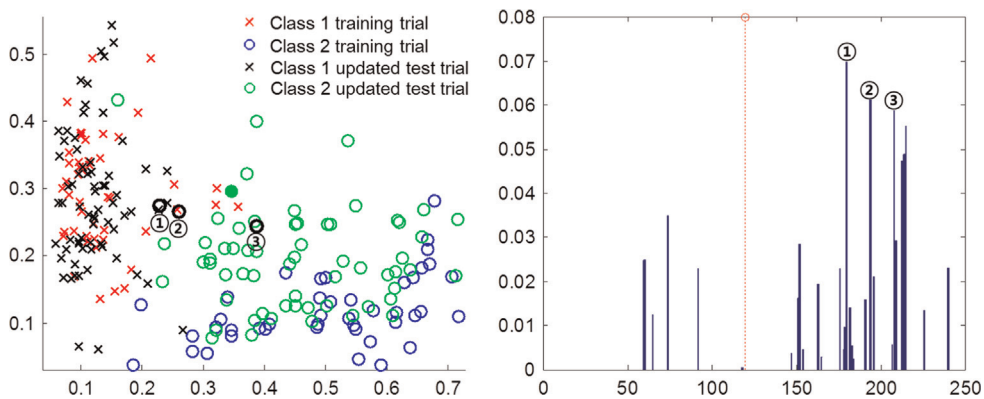


Fig. 7. Classification results of SRC_UAU IDM for the same test sample in Fig. 6. (Left): Scatter plot of training features for two classes and one test feature of class 2. (Right): Sparse representation results of one test feature shown in left figure from the SRC_UAU IDM. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

re-training (re-adjustment) of the decision hyper-plane for new test data is time consuming process. However, in the proposed methods, dictionary update for adaptation of each test sample is very simple process and re-training of classifier is not needed.

5. Discussions

5.1. Results for public dataset

For the evaluation of the proposed methods, we use a public dataset obtained from Dataset Ivc of BCI Competition III [32]. In this dataset, the test data were separately recorded for more than 3 h after the acquisition of the training data. Therefore, the distribution of some EEG features could be effected by non-stationarities. This dataset was recorded from a healthy subject. He sat in a comfortable chair with his arms resting on the armrests. The training dataset consists of the data of the first three (non-feedback) sessions. In all, 210 training trials (105 for each class) were obtained. The visual cues (letter presentation) indicated for 3.5 s which of the following two motor imagery that the subject had to perform: (L) left hand and (F) right foot. The target cues were presented at intervals of random length ranging from 1.75 to 2.25 s, in which the subject could relax. In the test sessions, total 280 test trials (140 for each class) were recorded. The experimental setup was similar to the setup of the training sessions, but the motor imagery had to be performed for 1 s only, compared to 3.5 s in the training sessions. The recording was made using BrainAmp amplifiers and a 128-channel Ag/AgCl electrode cap

from ECI. A total of 118 EEG channels were measured at the positions of the extended international 10/20 system. Signals were band-pass filtered between 0.05 and 200 Hz, and then digitized at 1000 Hz.

Table 3 shows the classification accuracy of the public dataset for conventional SRC and the four proposed adaptive SRC schemes when the number of elimination trials n is varied from 0 (no IDM) to 30. For this dataset, six CSP filters are used for feature extraction, and thus, the dimension of dictionary \mathbf{A} is 6×210 for the original SRC. In all, 280 test trials are classified by each classification method. From the results presented in Table 3, we find that all proposed adaptive SRC methods exhibit improved classification accuracy compared to the conventional SRC method irrespective of the value n of IDM. Supervised dictionary update methods (SAU and SFU IDM) show better classification accuracy than the unsupervised methods (UAU and UFU IDM); however, the difference is very small (within 1%). Further, the difference between the accumulated (SAU and UAU IDM) and the fixed dictionary update methods (SFU and UFU IDM) is more small and negligible for this dataset.

5.2. Comparison between proposed adaptive schemes

In this section, first, we compare the accumulated and fixed dictionary update rule for each supervised and unsupervised dictionary update method. From the results of Table 1, the mean difference between SRC_SAU and SRC_SFU with IDM is just 0.2%. For the unsupervised case, SRC_UAU and SRC_UFU with IDM exhibit a mean difference of 0.5%. To analyze the statistical

Table 2
Comparison of classification accuracy (%) between conventional non-adaptive classification methods (LDA, SVM, and SRC) and adaptive classification methods (including the proposed adaptive SRC schemes). The highest classification accuracy for each dataset is highlighted in bold.

Dataset	LDA	Adaptive LDA	SVM	Adaptive SVM	SRC	SRC_SAU IDM	SRC_SFU IDM	SRC_UAU IDM	SRC_UFU IDM
1	56.0	62.7	68.7	69.3	66.0	70.7	64.7	67.3	67.3
2	88.0	87.3	88.0	88.0	86.0	88.0	88.0	89.3	90.7
3	87.3	86.7	86.0	86.0	88.7	90.0	90.7	90.7	88.7
4	94.3	94.3	95.7	95.0	96.4	96.4	97.1	96.4	96.4
5	78.0	84.0	80.0	89.3	83.3	96.0	96.7	95.3	97.3
6	79.3	82.0	84.7	90.7	82.7	86.7	84.0	84.0	83.3
7	68.7	74.0	71.3	80.0	77.3	80.0	79.3	77.3	78.0
8	84.7	89.3	70.7	89.3	73.3	88.7	91.3	89.3	90.7
9	70.7	74.0	69.3	73.3	70.0	74.7	74.0	72.0	71.3
10	53.3	63.3	58.0	62.7	62.0	68.7	71.3	63.3	66.7
11	79.3	82.7	70.0	87.3	84.0	89.3	89.3	88.0	88.7
12	87.3	91.3	94.0	95.3	96.7	98.0	98.0	98.0	98.0
Mean	77.2	81	78	83.9	80.5	85.6	85.4	84.3	84.8
Std.	12.84	10.36	11.70	10.36	11.13	9.94	10.89	11.64	11.40

Table 3
Classification accuracy (%) of conventional SRC and the proposed adaptive SRC methods for the BCI competition dataset.

<i>n</i> of IDM	SRC	SRC_SAU IDM	SRC_SFU IDM	SRC_UAU IDM	SRC_UFU IDM
0	92.5	95.36	95.36	93.93	94.64
5	92.86	96.07	95.71	94.64	94.64
10	90	95.36	95.71	93.93	93.93
15	92.86	95.36	95.36	94.64	94.64
20	91.43	95.36	95.71	95.36	94.64
30	91.79	95	95	94.64	94.64
Mean	91.91	95.42	95.48	94.52	94.52

significance of the mean differences, we perform the paired *t*-test for the accuracy of each subject. The obtained *p*-values of the *t*-test are larger than 0.05 for the comparisons of the accumulated and the fixed update rule, which means that the differences are not statistically significant. As we mentioned in Section 3.3.2, when the number of original training trials in the dictionary and that of the updated new test trials are large, the computation time of the accumulated dictionary update based SRC method might be increased to solve the sparse coding step, i.e., Eq. (1), by using L1 minimization as compared to the fixed dictionary update based SRC method. Thus, in the fixed update rule, the dictionary size is fixed for all test trials and the computation time for sparse coding is not increased. However, in the accumulated update rule, the dictionary size is increased in every test trial, and therefore, the computation time for the sparse coding step is also increased. We compare the running time (computation time) of the accumulated and fixed dictionary update methods. Because of the number of training trials and that of the test trials of the competition dataset, which is used in Section 6.1 (210 and 280), are larger than our dataset (120 and 150), we use the competition dataset to evaluate the running time. The *tic* and *toc* MATLAB commands are used for measuring the running time of the sparse coding step in the SRC algorithm. We repeat 100 times and measure the average running time for each method. For a single test trial, the average running time of the sparse coding step in SRC_SAU and SFU are 5.47 ms and 4.29 ms respectively. Further, the SRC_UAU and UFU show the average running time of 5.45 ms and 4.26 ms for the sparse coding step, respectively. Therefore, for a single test trial, the differences in the running time between the accumulated and the fixed update rule are very small and negligible for online BCI applications.

Second, we investigate supervised and unsupervised dictionary update methods. From the results presented in Table 1, we find that the mean difference between SRC_SAU and SRC_UAU with

IDM is 1.3%. For this comparison, we obtained a *p*-value of 0.04 from the paired *t*-test. For the unsupervised case, the mean difference between SRC_SFU and SRC_UFU with IDM is 0.6% and the obtained *p*-value is larger than 0.05. Even though the mean differences are not much, all supervised methods consistently show better mean classification accuracy than the unsupervised methods for our dataset and the public dataset presented in Tables 1 and 3, respectively. In the unsupervised dictionary update method, the class labels of the test trials are determined by the results of the current classifier. Unfortunately, the classifier usually does not provide perfect classification results for all test trials because of the non-stationarity of EEG. Few incorrectly classified test trials are also updated in a different class-dictionary with the original target class. These trials affect the sparse coding step in the SRC algorithm. Therefore, this might be the reason that the unsupervised methods exhibit lower mean classification accuracy than the supervised methods. However, from the results for our dataset and the public dataset, we find that the unsupervised methods still show improved classification results compared to the original SRC.

5.3. Analysis of IDM method

As shown in the results of Table 3, the classification accuracy of IDM based SRC methods may vary on the basis of the value *n* of IDM. The value *n* can be heuristically chosen to optimize the classification accuracy. In this section, we analyze the effect of the number of elimination trials *n* of IDM by using our experimental dataset. In the results presented in Table 1, for a fair comparison, we set the same value of *n* of 10 for all 12 datasets. For the same datasets, in Fig. 8, we compute average classification accuracy over all datasets when the number of elimination trials of SAU, SFU, UAU and UFU IDM is varied from 0 to 30. From the results of Fig. 8, the optimal number *n* is different for each method. This means that there is a place to improve classification performance of IDM based adaptive SRC method by finding optimal *n* for each method and also each subject dataset. In Fig. 8, compared to the results of supervised update methods average accuracy is decreased with the large value of *n* in the case of unsupervised update methods. This might be because if the number of elimination trials *n* is large, number of training trials is decreased in the dictionary. Thus, the role for classification task of updated new test trials is increased. However, in the case of unsupervised method, class label of new test trials is not always correctly updated. Therefore, for the unsupervised update methods with IDM, the value *n* is needed to choose more carefully.

Next, we analyze the effect of the incoherence based dictionary modification (IDM) method. As we mentioned in Section 3.1.1, we

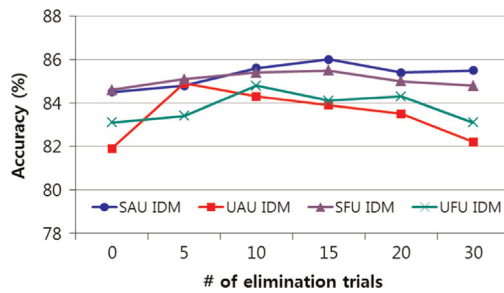


Fig. 8. Average classification accuracy of SAU IDM and UAU IDM when the number of elimination trials n is varied.

propose an IDM method to make more incoherent dictionary after applying the CSP filtering. Incoherence of dictionary can be measured by coherence value C introduced in Eq. (3). To evaluate the change in the coherence value, we measure the C value of SRC without IDM and with IDM method. From the average results over twelve datasets, The SRC without IDM shows 0.983 value of C . On the other hand, the SRC with IDM shows 0.934 value of C . This means that after applying the IDM method, we can make more incoherent dictionary than the without IDM method.

6. Conclusion

Because of the inherent non-stationarity of EEG signals, performance degradation is an inevitable phenomenon in EEG based BCI systems. In particular, an already designed classifier by the training data does not guarantee satisfactory classification accuracy for new test data in the online feedback stage. In this paper, we propose dictionary update methods with incoherence based dictionary modification (IDM) as adaptive SRC schemes to compensate for the non-stationary effects. We consider supervised/unsupervised and accumulated/fixed dictionary update rules with IDM. With the unique classification mechanism of the SRC, i.e., a fixed decision rule is not required for the classification, in the proposed dictionary update methods, the test data are easily updated and utilized for the classification of other new test data without requiring any additional computation. In addition, in the IDM algorithm, we try to create a maximally incoherent dictionary for SRC by using a simple incoherence measure of the training data. By using two online motor imagery based BCI experimental datasets, we evaluate the classification performance of the proposed adaptive schemes. From the results, we find that the proposed IDM based adaptive SRC schemes show improved classification results compared to the conventional SRC. Further, unsupervised adaptive SRC schemes that are more practically applicable in BCI exhibit competitive classification accuracy than other adaptive LDA and SVM methods. An analysis of a stable dictionary to overcome the inter-subject variation in BCI systems and a fully adaptive classification method developed by combining adaptive CSP filtering with adaptive SRC will be interesting future works.

Conflict of interest statement

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgments

This work was supported by the National Research Foundation (NRF) of Korea Grant funded by the Korean government (NRF-2015R1A2A1A05001826).

References

- [1] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, Brain-computer interfaces for communication and control, *Clin. Neurophysiol.* 113 (6) (2002) 767–791.
- [2] J.R. Wolpaw, D.J. McFarland, G.W. Neat, C.A. Forneris, An EEG-based brain-computer interface for cursor control, *Electroencephalogr. Clin. Neurophysiol.* 78 (1991) 252–259.
- [3] G. Dornhege, R. del, J. Millán, T. Hinterberger, D.J. McFarland, K.-R. Müller, *Toward Brain-Computer Interfacing*, The MIT Press, Cambridge, Massachusetts (2007), p. 213–215.
- [4] G. Pfurtscheller, D. Flotzinger, J. Kalcher, Brain-computer interface – a new communication device for handicapped persons, *J. Microcomput. Appl.* 16 (1993) 293–299.
- [5] E. Sellers, E. Donchin, A P300-based brain-computer interface: Initial tests by ALS patient, *Clin. Neurophysiol.* 117 (2006) 538–548.
- [6] (<http://emotiv.com/store/epoc-detail/>).
- [7] (http://www.quasarusa.com/products_dsi.htm).
- [8] Y.M. Chi, Y.T. Wang, Y. Wang, C. Maier, T.P. Jung, G. Cauwenberghs, Dry and noncontact EEG sensors for mobile brain-computer interfaces, *IEEE Trans. Neural Syst. Rehabil. Eng.* 20 (2) (2012) 228–235.
- [9] L.D. Liao, C.Y. Chen, I.J. Wang, S.F. Chen, S.Y. Li, B.W. Chen, J.Y. Chang, C.T. Lin, Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors, *J. Neuroeng. Rehabil.* 9 (2012) 5, <http://dx.doi.org/10.1186/1743-0003-9-5>.
- [10] C.T. Lin, B.S. Lin, F.C. Lin, C.J. Chang, Brain computer interface-based smart living environmental auto-adjustment control system in UPnP home networking, *IEEE Syst. J.* 8 (2014) 363–370.
- [11] P. Shenoy, M. Krauledat, B. Blankertz, R.P.N. Rao, K.-R. Müller, Towards adaptive classification for BCI, *J. Neural Eng.* 3 (2006) R13–R23.
- [12] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K.R. Müller, Optimizing spatial filters for robust EEG single-trial analysis, *IEEE Signal Process. Mag.* 25 (1) (2008) 41–56.
- [13] J.D.R. Millán, On the need for on-line learning in brain-computer interfaces, in: *Proceedings of the International Joint Conference on Neural Networks*, 2004, pp. 2877–2882.
- [14] Y. Li, C. Guan, An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces, *Neural Comput.* 18 (11) (2006) 2730–2761.
- [15] C. Vidaurre, M. Kawanabe, P. von Büna, B. Blankertz, K.-R. Müller, Toward unsupervised adaptation of LDA for brain-computer interfaces, *IEEE Trans. Biomed. Eng.* 58 (2011) 587–597.
- [16] K. Huang, S. Aiyente, Sparse representation for signal classification, *Adv. Neural Inf. Process. Syst.* 19 (2006) 609–616.
- [17] S. Younghak, L. Seungchan, L. Junho, L. Heung-No, Sparse representation-based classification scheme for motor imagery-based brain-computer interface systems, *J. Neural Eng.* 9 (2012) 056002.
- [18] H. Yu, H. Lu, T. Ouyang, H. Liu, and B.L. Lu, Vigilance detection based on sparse representation of EEG, in: *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 2439–2442.
- [19] H. Ramoser, J. Müller-Gerking, G. Pfurtscheller, Optimal spatial filtering of single trial EEG during imagined hand movement, *IEEE Trans. Rehabil. Eng.* 8 (4) (2000) 441–447.
- [20] G. Pfurtscheller, C. Neuper, Motor imagery and direct brain-computer communication, *Proc. IEEE* 89 (2001) 1123–1134.
- [21] D.L. Donoho, M. Elad, Optimally sparse representation in general (non-orthogonal) dictionaries via L1 minimization, *Proc. Natl. Acad. Sci.* 100 (2003) 2197–2202.
- [22] D. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory* 52 (2006) 1289–1306.
- [23] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2) (2009) 210–227.
- [24] J.F. Gemmeke, T. Virtanen, A. Hurmalainen, Exemplar-based sparse representations for noise robust automatic speech recognition, *IEEE Trans. Audio, Speech, Lang. Process.* 19 (7) (2011) 2067–2080.
- [25] Q. Yuan, W. Zhou, S. Yuan, X. Li, J. Wang, G. Jia, Epileptic EEG classification based on kernel sparse representation, *Int. J. Neural Syst.* 24 (4) (2014) 1450015.
- [26] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, B. Arnaldi, A review of classification algorithms for EEG-based brain-computer interfaces, *J. Neural Eng.* 4 (2) (2007) R1–R13.
- [27] W. Samek, C. Vidaurre, K.R. Müller, M. Kawanabe, Stationary common spatial patterns for brain-computer interfacing, *J. Neural Eng.* 9 (2) (2012) 026013.
- [28] M. Arvaneh, C. Guan, K.K. Ang, C. Quek, Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-

- computer interface, *IEEE Trans. Neural Netw. Learn. Syst.* 24 (4) (2013) 610–619.
- [29] M. Kawanabe, W. Samek, K.R. Müller, C. Vidaurre, Robust common spatial filters with a maxmin approach, *Neural Comput.* 26 (2) (2014) 1–28.
- [30] D.L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. Inf. Theory* 47 (2001) 2845–2862.
- [31] S. Younghak, L. Seungchan, A. Minkyu, C. Hohyun, J. Sung Chan, L. Heung-No, Noise robustness analysis of sparse representation based classification method for non-stationary EEG signal classification, *Biomed. Signal Process. Control* 21 (2015) 8–18.
- [32] B. Blankertz, Berlin Brain–Computer Interface. (http://www.bbci.de/competition/iii/desc_IVc.html).



Magnetic field induced one-dimensional nano/micro structures growth on the surface of iron oxide thin film



Pawan Kumar^a, Rajesh Kumar^{b,*}, Heung-No Lee^b

^a Jaypee University of Information Technology, Waknaghat, Solan, 173234 Himachal Pradesh, India

^a Gwangju Institute of Science and Technology, Gwangju 500712, Republic of Korea

ARTICLE INFO

Article history:

Received 17 March 2015

Received in revised form 21 July 2015

Accepted 30 August 2015

Available online 2 September 2015

Keywords:

Crystal growth

Magnetic materials

Thin films

Nanoparticles

Microstructure

Magnetic field

Oriented growth

ABSTRACT

The influence of external magnetic field on the morphology of α -Fe₂O₃ thin film formed at liquid–vapor interface has been investigated. Application of magnetic field during the growth of film resulted in the magnetic moment ordering of constituent nanoparticles. Thus formed α -Fe₂O₃ thin film was transferred to a glass substrate, which upon annealing converted into one dimensional (1D) nanostructured thin film due to the oriented attachment of magnetically ordered nanoparticles. The effect of dopants viz. Ni²⁺ and Co²⁺ on the directional growth, and magnetic properties of nanostructures has also been investigated. The Ni²⁺ and Co²⁺ doped α -Fe₂O₃ 1D nanostructured thin films show superparamagnetic and ferromagnetic behavior, respectively, whereas undoped α -Fe₂O₃ film exhibits superparamagnetism. From the room temperature magnetization measurements of films, it is found that the magnetization depends upon the morphology and magneto-crystalline anisotropy attributes of the film nanostructures.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Owing to outstanding electrical, magnetic and optical properties the nano/micro sized structures of iron oxide have attracted great attention as compared with the bulk counterparts [1–2]. Among iron oxides, hematite (α -Fe₂O₃) is typically nontoxic and environment friendly iron oxide with band gap $E_g = 2.1$ eV [3]. In the case of film, the morphology and size of constituent α -Fe₂O₃ nanoparticles have great impact on its physical and chemical properties. The nano/micron sized structures of α -Fe₂O₃ have applications in diverse fields including catalysis [4], sensors [5], lithium-ion batteries [6], and environment protection, etc. [7].

Since the nanomaterials exhibit shape and/or size dependent properties [8], therefore, various efforts have been made to synthesize one-dimensional iron oxide structures for specific applications [9]. Here, we quote some of the methods reported for the growth of one-dimensional (1D) iron oxide structures, such as solution method [10–11], thermal oxidation [12–14], forced hydrolysis [15–16], hydrothermal [17–18] chemical precipitation [19], and solvothermal method [20]. For practical applications, such as integrated devices, these one-dimensional nanostructures (nanowires and nanorods) should be grown on substrate to form vertically aligned arrays. Still now, despite of tremendous efforts, it is challenging to develop a simple and versatile way to form α -Fe₂O₃ thin film composing 1D structure. However, for the synthesis of

structured iron oxide film, the magnetic field may be considered as one of the synthesis parameters alike to the temperature and pressure. The applied magnetic field is not sensitive to the surface charges and solution pH, therefore, it does not influence the reaction mechanism as the other parameters do (electric field or current).

There are few reports where the magnetic field has been employed for the synthesis and assemblies of 1D and two-dimensional (2D) aggregates. During synthesis, the applied magnetic field enhances the dipole–dipole interaction by decreasing the surface energy, which results in the directional growth along the easy axis of magnetization. The effect of magnetic field is more in the case of materials possessing higher magnetic susceptibility due to their easy formation in the system, which is due to the magnetic field effect on Gibbs free energy leading to tremendous impact on structures and properties of materials [21]. The spin state of ions in the crystal structure can be changed by applying magnetic field during the synthesis process. Applied magnetic field generates novel magnetic domains in sample. In literature also, the application of magnetic field is reported an elegant way to orient and assemble disordered structures into highly ordered structures [22–30]. Nowadays, magnetic fields have been widely employed in the nanomaterials research area [31–34]. The response of magnetic field is different for ferromagnetic, paramagnetic and diamagnetic materials. In the case of ferromagnetic/ferrimagnetic materials, the growth of nanostructures in the presence of weak magnetic fields can induce anisotropy leading to the formation of 1D growth of nanostructures in the easy magnetization direction. The field strength and orientation can be varied or kept

* Corresponding author.

E-mail address: rajesh.kumar@juit.ac.in (R. Kumar).

constant, for each of these magnetic nanomaterials in space and time or in both. When the magnetic field is applied, the Brownian motion on the surface of the solution gets diminished due to magnetic field lines, and the applied magnetic field forces the nucleated nanoparticle to align along their easy axes parallel to magnetic field.

Here, we report the formation of nano/micro structures on the surface of α - Fe_2O_3 thin films by applying external magnetic field during the film formation process. Along with undoped α - Fe_2O_3 thin films, films doped with Ni^{2+} and Co^{2+} were also formed in the presence of external magnetic field. The effect of Ni^{2+} and Co^{2+} doping on the morphology and magnetic properties of the formed iron oxide structures is studied. The as prepared nanostructured thin films were studied for structural, morphological and magnetic properties. The present study gives a new method of directional growth of one dimensional nanostructures, opening up a new way for controlled synthesis of nanostructured thin films with various dimensionalities and morphologies.

2. Experimental

Initially, a precursor solution containing 24 mM FeCl_2 , 22 mM $\text{FeCl}_3 \cdot 6\text{H}_2\text{O}$ and 64 μM of polyvinyl alcohol (PVA) was formed. The measured pH value of solution was 2.8. The solution was placed in

an ice-chamber to reduce the thermal fluctuations [35]. After cooling the solution, an out of plan magnetic field (~ 0.8 T) was applied on the surface of solution by using an electromagnet (with poles diameter 2 in.). A gap of 2.5 mm was kept between the solution surface and electromagnetic pole. Then to form a thin film on the surface of solution, NH_3 vapor (6 volume %) was introduced into the chamber. The NH_3 vapor interacts with the Fe^{3+} ions in precursor solution residing on the surface, and forms an iron oxide-poly vinyl alcohol (PVA) composite thin film (as shown in schematic of Fig. 1). The obtained film was transferred to the glass substrate, and then annealed at 500°C in a horizontal tube furnace.

The thin film formation method, described above, was also applied to obtain doped (Ni^{2+} and Co^{2+} doping) iron oxide nanostructured thin films. The salts of NiCl_2 and CoCl_2 were taken in 15 molar percent, and added to the precursor solution, in two separate experiments. Thus obtained undoped and doped iron oxide thin films were also formed for horizontal magnetic field (in plane). Finally, these nanostructured thin films were characterized for structural properties by using X-ray Diffraction (XRD, PANalytical X'pert-PRO) employing $\text{Cu K}\alpha$ ($\lambda = 1.5406 \text{ \AA}$, $2\theta = 20$ to 60°) radiation, and for morphological study using Scanning Electron Microscopy (SEM, Hitachi, S-4700). The elemental composition and magnetic properties of the prepared samples were analyzed by Energy Dispersive X-ray spectroscopy (EDX,

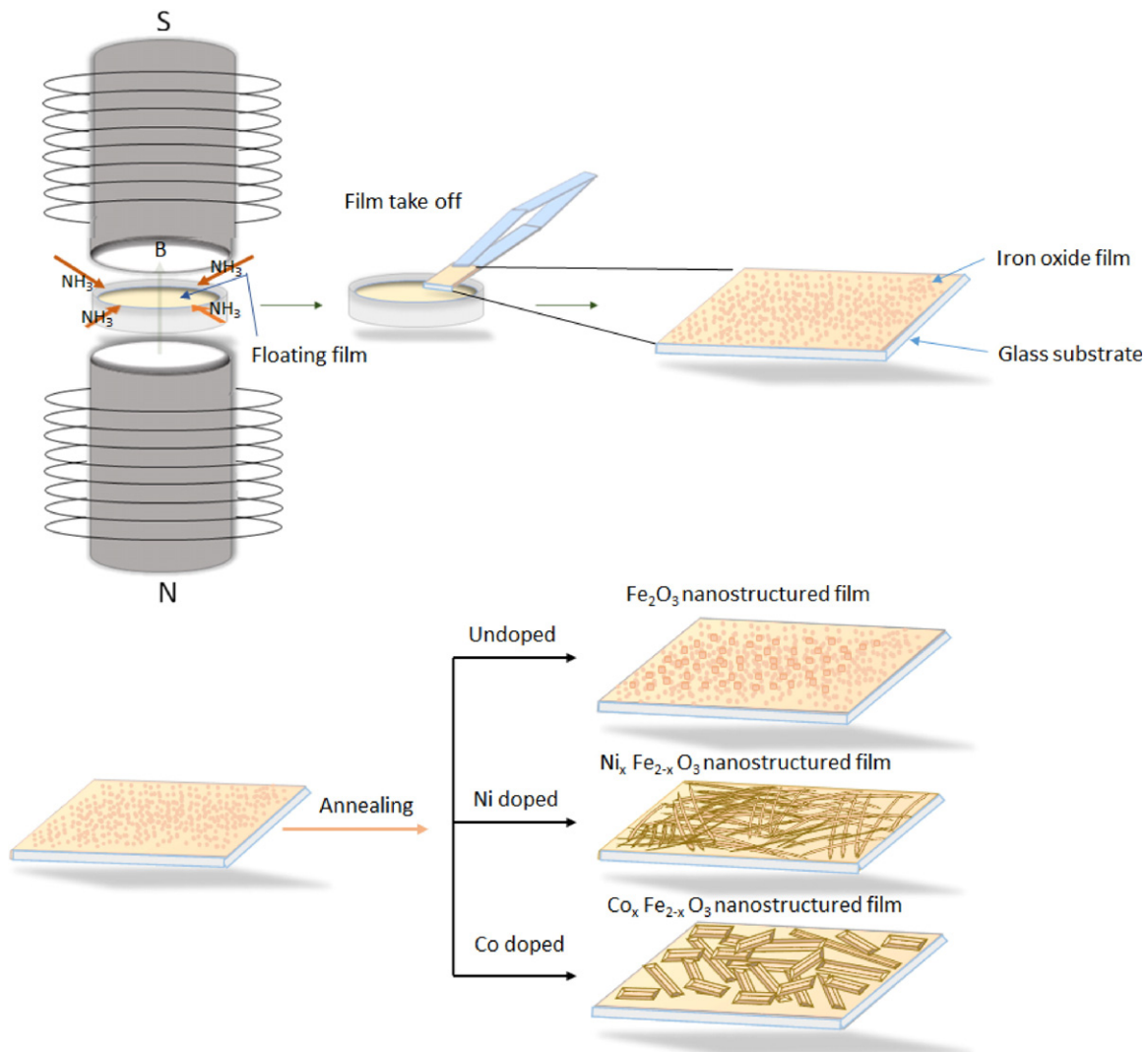


Fig. 1. Schematic presentation of the thin film formation process at the surface of solution in the presence of magnetic field (B) and ammonia (NH_3) vapor.

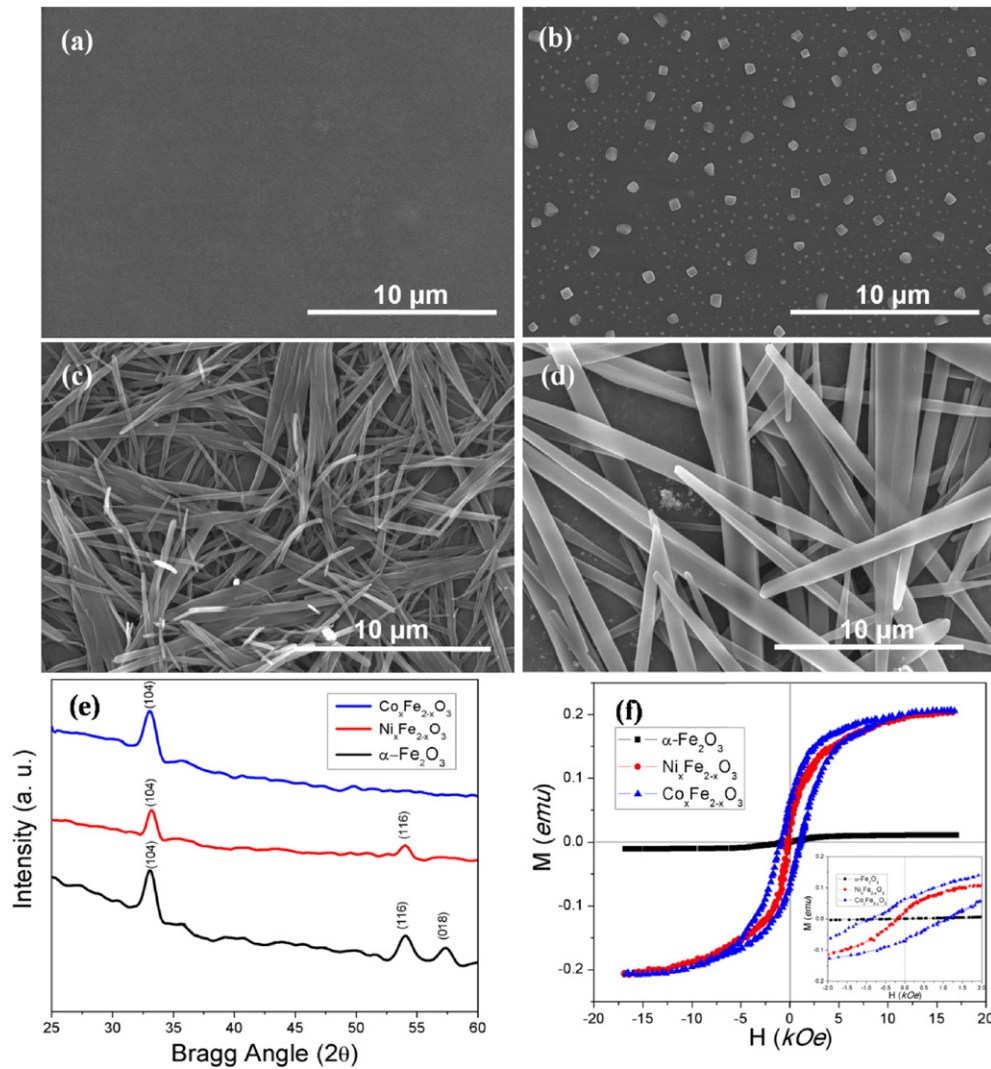


Fig. 2. SEM images of iron oxide thin films formed (a) without magnetic field, (b) with magnetic field (out of plane), (c) 15% Ni^{2+} doping with magnetic field applied, (d) 15% Co^{2+} doping with magnetic field applied and (e) XRD, and (f) VSM of the corresponding films (inset shows the magnetic behavior at the low magnetic field). All the films were annealed at 500 °C.

Oxford instruments, INCA PENTA FETX3) and Vibrating Sample Magnetometer (VSM) PAR 155.

3. Results and discussion

To estimate the effect of applied magnetic field, iron oxide thin films were formed both in the absence, and presence of external magnetic field. Fig. 2(a) and (b) shows the iron oxide thin film formed in the absence and presence of external magnetic field, respectively. The magnetic field was applied out of plane on the liquid–vapor interface. The films formed in the presence of external magnetic field possessing nanostructures (Fig. 2(b)) indicate that the external magnetic field has an effect on the surface morphology of the film. The iron oxide thin film is composed of nano and micrometer size particles. However, we observed that before annealing, both of the films that are formed in the absence and in the presence of magnetic field have similar surface morphology. But when annealed, the surface of the latter covered with nanostructured, whereas the former remained the same. Since the change in surface morphology appeared only after annealing, therefore, it may be inferred that the applied magnetic field has an effect on the magnetic moments of nanoparticles during the formation of film, which upon annealing resulted in nanostructured surface. The magnetic moment of iron oxide film can be enhanced by adding Ni^{2+} and Co^{2+} in the film [36–37]. In the present study, we also included Ni^{2+} and Co^{2+}

ions in the precursor solution, and investigated their magnetic moment's effect on the morphology of iron oxide films. Fig. 2(c) shows the SEM image of Ni^{2+} doped iron oxide thin film. In this case, one dimensional nanostructures can be observed on the top of the film, whereas the film formed without any doping has nanoparticles on its surface. The Ni^{2+} in this case enhanced magnetic moment of nanostructures, and led to one dimensional form of nanostructures. Similar results were observed in the case of Co^{2+} doping. However, in the case of Co^{2+} doping, the surface was covered with microstructures as shown in Fig. 2(d).

Fig. 2(e) is XRD intensity patterns corresponding to undoped, Ni^{2+} and Co^{2+} doped iron oxide thin films. From the XRD pattern, it is observed that all the films are well crystalline, and match their diffraction peaks with those of $\alpha\text{-Fe}_2\text{O}_3$ (JCPDS no. 89–8104). Also, there is no other secondary phase due to Ni^{2+} and Co^{2+} doping. The crystalline size calculated using Scherrer's formula, $D_{\text{hkl}} = 0.9/\beta \cos\theta$ is 6.7 nm, 7 nm and 6.3 nm for $\alpha\text{-Fe}_2\text{O}_3$, Ni^{2+} and Co^{2+} films, respectively. To study the magnetic behavior of the fabricated films, the M–H measurements were performed at room temperature. The M–H curve of undoped and doped iron oxide thin films is shown in Fig. 2(f). Both the undoped and Ni^{2+} doped samples show superparamagnetic behavior. In the magnetic curves, the undoped sample saturates at 7.95×10^5 A/m (or 10,000 Oe, as shown in the graph), whereas Ni^{2+} doped sample saturates above than 13.5×10^5 A/m (or 17,000 Oe). Also, the

magnetization value in the case of Ni^{2+} doped sample is higher than undoped sample. In the case of Co^{2+} doped sample, a ferromagnetic behavior with relatively larger coercivity value of 7.95×10^4 A/m (or 1000 Oe), and larger remanence is observed. The observed higher coercivity and remanence in Co^{2+} doped sample are attributed to enhanced shape of structures and related magneto-crystalline anisotropy [38]. We know that the magnetic iron oxide film doped with Co^{2+} ions has a stronger spin-order interaction than Fe^{2+} ions [39]. The doping of Co^{2+} ions decompensates the antiferromagnetic order of the lattice, which leads to an enhanced effective magnetic field seen by the Fe^{3+} nucleus [40]. Due to higher value of magneto-crystalline anisotropy of Co^{2+} ion, the post synthesis annealing resulted in large directional growth of nanostructure which prevented them from magnetizing in the directions other than that along their easy magnetic axes, leading to a higher directional growth and coercivity. The presence of Ni^{2+} and Co^{2+} was confirmed by EDX analysis. Fig. 3 shows the EDX of Ni^{2+} and Co^{2+} doped structures, these nanostructures have Ni^{2+} and Co^{2+} with the atomic percent of 15 and 14, respectively.

In literature, different magnetic behaviors of $\alpha\text{-Fe}_2\text{O}_3$ nanostructure are reported. There are few studies [41,42], which indicate $\alpha\text{-Fe}_2\text{O}_3$ nanostructures synthesized via sol-gel and hydrothermal methods to be superparamagnetic. However, the other studies report that $\alpha\text{-Fe}_2\text{O}_3$ nanostructures are ferromagnetic [43,44]. In this study, we have obtained undoped and Ni^{2+} doped $\alpha\text{-Fe}_2\text{O}_3$ structures which show superparamagnetic behavior, and doped with Co shows ferromagnetic behavior. In our case, the undoped thin film has small size of $\alpha\text{-Fe}_2\text{O}_3$ nanocrystals, which should have uncompensated surface spin at their boundaries. The uncompensated spins lead the undoped $\alpha\text{-Fe}_2\text{O}_3$ thin film to be superparamagnetic.

In the case of Co doping, due to smaller ionic radii of cobalt (72 pm), as compare with iron (74 pm), it may occupy the interstitial positions or sit on the grain boundaries. The XRD data indicates polycrystalline nature of the sample, possessing large number of grain boundaries. Here, the Co atoms will destroy the crystalline structure, which results into a decreased crystalline size, and therefore disappearance of the (116) and (018) peaks from the data. The Co with electronic configuration $[\text{Ar}] 3d^7 4s^2$ has one electron in excess than Fe $[\text{Ar}] 3d^6 4s^2$ which has less energy of d state. When Co^{2+} with spin down electron substitutes Fe^{3+} ion, the spin down d band gets completely filled with remaining one d electron in the spin up band, which results in a net magnetization of $1 \mu_B$ [45]. The increase in the magnetization value of Co-doped Fe_2O_3 takes place due to the canting of spin structure. The canting of spin structure is created by the imbalance resulted from the incorporation of Co^{2+} in Fe_2O_3 lattice [46]. A similar increased magnetization value behavior has been observed by Wieslaw A. Kaczmarek et al. (1996) [47]. The canting of spin produces an uncompensated magnetic moment of Fe^{3+} cation, resulting in a ferromagnetic behavior of the sample.

Similarly, in the case of Ni doping, the d bands of Ni ($3d^8 4s^2$) have lower energy than those of Fe. Here, the five d states in the down spin channel are occupied, and the remaining two d electrons are on the t_{2g} states of the Ni^{2+} site which are 2/3 filled. The local moment on the

Ni^{2+} is $3 \mu_B$, and is polarized in the same direction as that of substituted Fe^{3+} , which gives a net magnetic moment of $2 \mu_B$ in the direction opposite to the magnetic moment of the substituted Fe^{3+} [45]. The increase in the saturation magnetization of Ni^{2+} substituting at Fe^{3+} sites is due to the higher surface spins of electron. This occurs due to the increase in surface spin that causes an enhancement of the magnetization of anti-ferromagnetic nanoparticles. The over occupancy of Ni^{2+} ions in the tetrahedral sites of $\alpha\text{-Fe}_2\text{O}_3$ creates more dense structure of pinning centers and discourages irreversible domain wall movement, and decreases the coercivity of Ni^{2+} doped $\alpha\text{-Fe}_2\text{O}_3$ [45] resulting in a superparamagnetic thin film.

To estimate the effective direction of applied magnetic field, which give rise to the structured surface morphology of the film, we also investigated the effect of magnetic field which was applied parallel (in plane) to the liquid-vapor interface. Fig. 4(a) shows the corresponding SEM images of $\alpha\text{-Fe}_2\text{O}_3$ thin film formed in the presence of magnetic field applied parallel to the liquid-vapor interface; the corresponding film obtained after annealing is shown in Fig. 4(b).

The formation of worm like nanostructures of $\alpha\text{-Fe}_2\text{O}_3$ on the film surface took place after annealing the Ni^{2+} and Co^{2+} doped samples (Fig. 4(c) and (d)). The worm like structures are formed due to the crack formation on the film surface during the synthesis process in the presence of external (in plane) magnetic field, and size of the nanostructures changed due to the change in magnetic moment by doping.

4. Mechanism of the nano/micro structure formation

To ensure the formation of structures on the film surface due to annealing, we investigated thin film samples at different annealing temperatures. For this study, Co^{2+} doped iron oxide thin film was selected, and annealed at 100, 300 and 500 °C temperature. Fig. 5 shows SEM images of films formed after annealing at different temperatures. From the SEM images, it can be seen that without annealing, no nanostructure protudes on the film surface (Fig. 5(a)) but for the film annealed at 100 °C, small grains started to agglomerate on the film surface as shown in Fig. 5(b). For 300 °C of annealing temperature, one dimensional structures emerge out of the film surface (Fig. 5(c)), which enhanced to a length of micrometers at 500 °C as shown in Fig. 5(d). These results show that the growth of nanostructure takes place during the annealing process, and the applied magnetic field induces a directional magnetic moment inside the oxide nanoparticles during the formation of film.

The effect of external magnetic field on the magnetic moment of nucleated nanoparticles can be understood in the following way. We know that the magnetic force $F(z)$ on metal ions at a position z is expressed by [40];

$$F(z) = \chi n H(z) \frac{\partial H(z)}{\partial z}. \quad (1)$$

Where n is number mole of Fe ions, χ is magnetic susceptibility, and $H(z)$ is applied magnetic field. When magnetic field is employed on

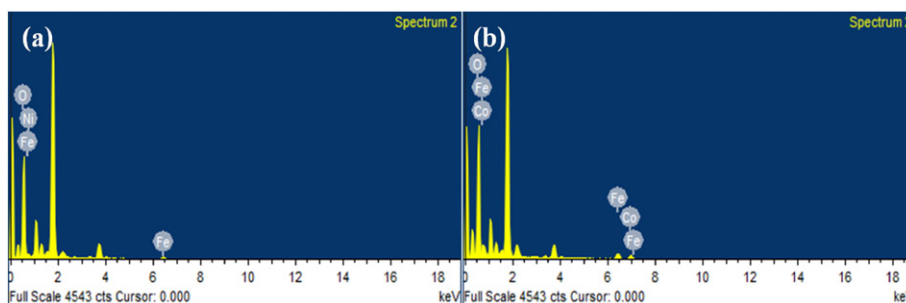


Fig. 3. The EDX of (a) Ni^{2+} doped and (b) Co^{2+} doped $\alpha\text{-Fe}_2\text{O}_3$ structure formed on the surface of thin film.

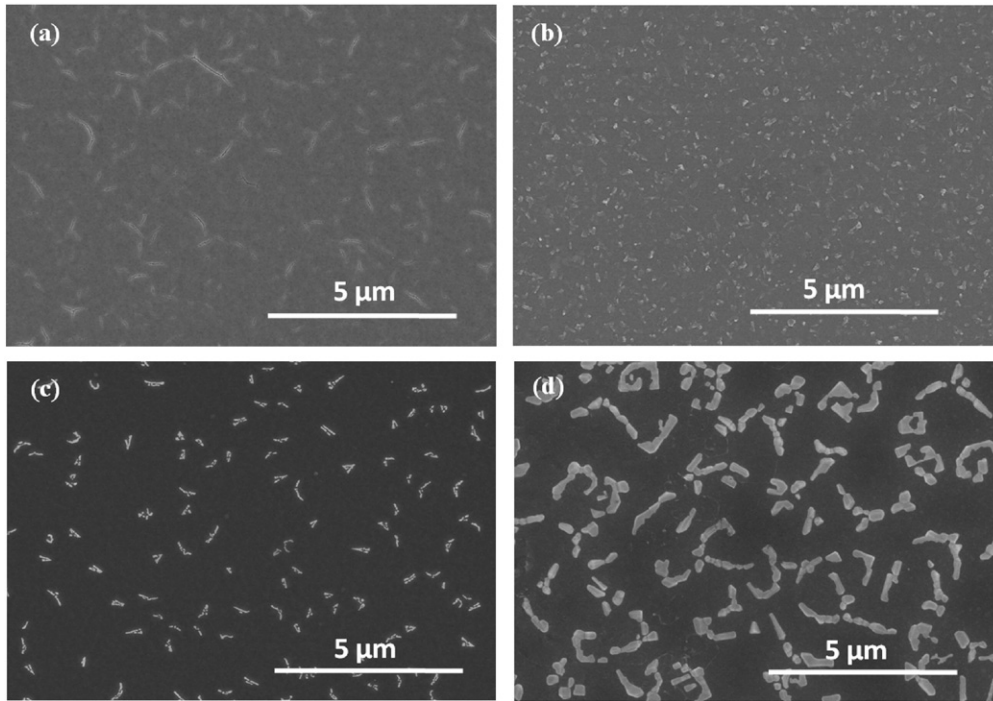


Fig. 4. The SEM images of α -Fe₂O₃ thin films (a) un-annealed and (b) annealed at 500 °C (c) 15% Ni²⁺ doped α -Fe₂O₃, and (d) 15% Co²⁺ doped α -Fe₂O₃ formed with the external magnetic field applied parallel (in plane) to the liquid–vapor interface.

liquid–vapor interface, it creates a change in the transport of Fe³⁺ ion and changes the Gibbs free energy of the reaction.

When a magnetic field is applied in a solution phase, the Fe³⁺ ions preferentially migrate and start to agglomerate along the magnetic line of force due to magnetic attraction, and the reaction occurs along the magnetic line of force. Thus the grain orientation of materials with magnetic anisotropy can be enhanced by applying a

magnetic field as the material in the presence of magnetic field will produce magnetic energy [40]. The difference in magnetization directions produces different magnetic energy. The energy difference could be described as [40]:

$$\Delta E = -\frac{1}{2\mu_0} \Delta\chi VB^2. \tag{2}$$

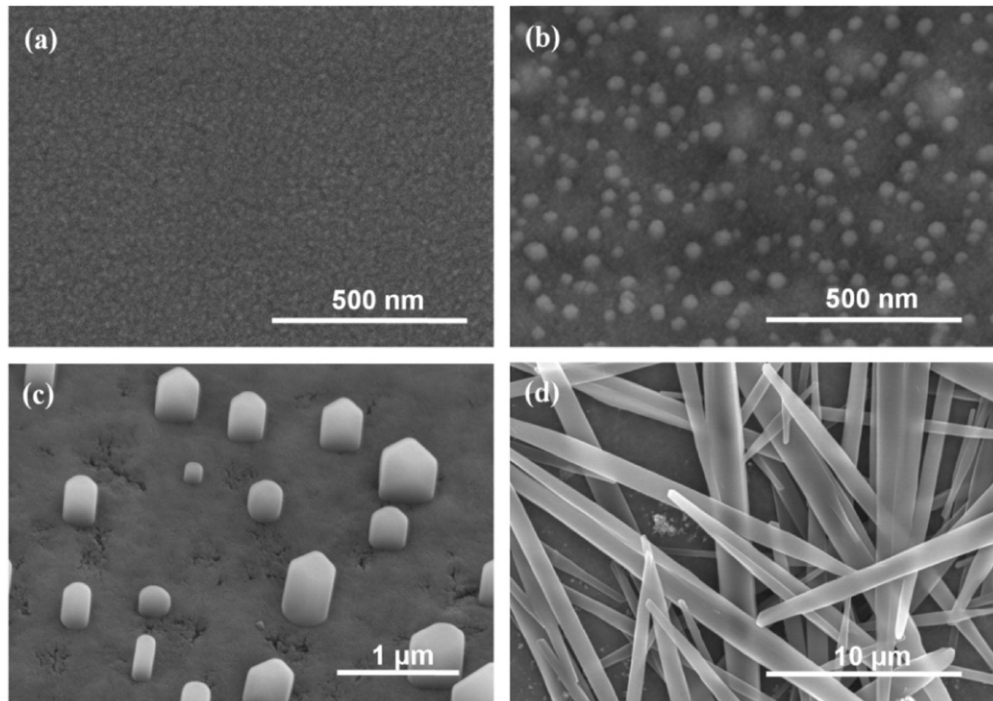


Fig. 5. SEM images of Co²⁺ doped iron oxide thin films (a) un-annealed, and annealed at (b) 100 °C, (c) 300 °C and (d) 500 °C temperature.

This orientation effect of a magnetic field is applicable to all (ferromagnetic/ferrimagnetic material, paramagnetic and diamagnetic) materials.

If the magnetic anisotropy is greater than the thermal energy, nucleated units will orientate with easy axes parallel to the applied field. For ferromagnetic and paramagnetic materials ($\chi > 0$), the largest magnetic susceptibility direction is parallel to the magnetic field direction and opposite for diamagnetic ($\chi < 0$) materials. Obviously, the orientation effect is associated with magnetic anisotropy and magnetic field intensity which influence the free energy (ΔG_M) of a chemical reaction, as given by [40]:

$$\Delta G_M = -\frac{1}{2}\mu_0(\Delta\chi_M)H^2. \quad (3)$$

Here, $\Delta\chi_M$ is the change in the susceptibility during reaction. The applied magnetic field determines the direction of any chemical change by controlling the ΔG_M . The generation of magnetic field effect is also due to the Zeeman interaction of the unpaired electron spins in Fe^{3+} ions with an external magnetic field. The increases length (L) in the presence of the applied magnetic field is given by the equation [48]:

$$L = L_0 + \delta L (1 - e^{-\alpha H_{app}}). \quad (4)$$

This equation shows dependence of L on the Boltzmann distribution factor ' $e^{-\alpha H_{app}}$ ', i.e. the ratio of Zeeman energy over thermal energy (αH_{app}). Zeeman energy of Fe^{3+} ions being in competition with the thermal activation energy in the presence of magnetic field [40] results to the nucleation of nanoparticles in the direction of easy axis that can minimize the energy of magnetization vector of material. Therefore, applied magnetic field might induce nucleation of $\alpha-Fe_2O_3$ grains along the easy magnetic axis, which upon annealing results in the formation of 1D nanostructure due to orientation arrangement. The 1D nanostructure results due to the oriented growth of materials determined by the surface energy of the material and experimental conditions.

The overall growth mechanism of 1D structure formation can be understood schematically by Fig. 6. Initially, when no magnetic field is applied, the iron oxide grains have magnetic moments oriented in the random direction, which after annealing do not show a directional growth (Fig. 6(a)). But when an external magnetic field is applied during the film formation, the nucleated grains might be having their magnetic moments aligned in the direction of magnetic field as shown in Fig. 6(b). These films upon annealing give directional growth due to oriented attachment of nanoparticles [49] as shown in Fig. 6(d). The high temperature annealing, evaporates PVA content from the film, and the magnetized grains arrange themselves to reduce their magneto-crystalline anisotropy energy, and results in a directional growth of the nanostructures.

5. Conclusions

Nano/micro structures are produced on the surface of thin film due to the application of magnetic field. The applied magnetic field produces an effect on the magnetic moment of nucleated iron oxide nanoparticles inside the film. The induced magnetic moment of nanoparticles align them along the direction of applied magnetic field, and upon annealing an oriented attachment nanoparticles form one dimensional structure on the film surface. Thus formed $\alpha-Fe_2O_3$ and Ni^{2+} doped $\alpha-Fe_2O_3$ films are superparamagnetic, whereas Co^{2+} doped film is ferromagnetic. The magnetic moment of $\alpha-Fe_2O_3$ film is successfully enhanced with the doping of Ni^{2+} and Co^{2+} ions. A larger value of magneto-crystalline anisotropy in Co^{2+} doped samples as compared with undoped and Ni^{2+} doped iron oxide films results in enlargement of 1D structures on the film surface.

Acknowledgments

This work was supported by nanotechnology research grant of Jaypee University of Information Technology and the National Research Foundation of Korea (NRF) grant funded by the Korean government (NRF-2015R1A2A1A05001826).

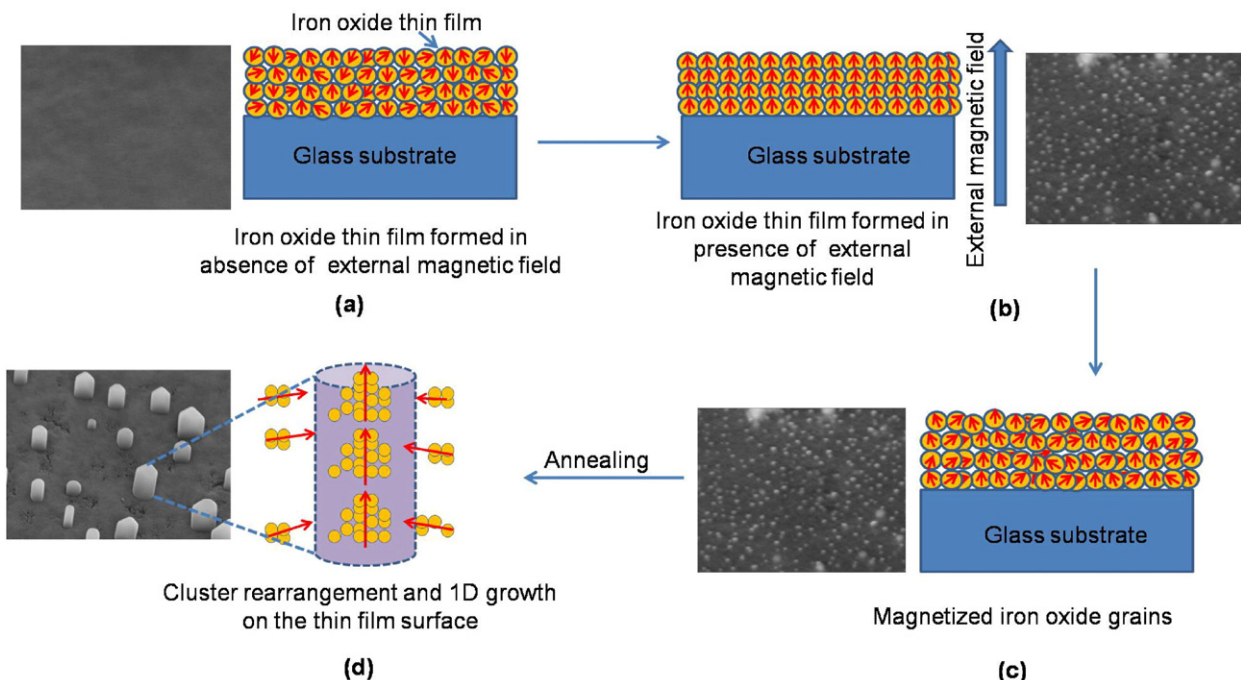


Fig. 6. Schematic of nanostructures formation mechanism on the surface of iron oxide thin film in the presence of external magnetic field.

References

- [1] A.M. Xavier, F.F. Ferreira, F.L. Souza, Morphological and structural evolution from akaganeite to hematite of nanorods monitored by ex situ synchrotron X-ray powder diffraction, *RSC Adv.* 4 (2014) 17753.
- [2] Y.M. Zhao, Y.H. Li, R.Z. Ma, M.J. Roe, D.G. McCartney, Y.Q. Zhu, Growth and characterization of iron oxide nanorods/nanobelts prepared by a simple iron–water reaction, *Small* 2 (2006) 422–427.
- [3] X. Wen, S. Wang, Y. Ding, Z.L. Wang, S. Yang, Controlled growth of large-area, uniform, vertically aligned arrays of α -Fe₂O₃ nanobelts and nanowires, *J. Phys. Chem. B* 109 (2005) 215–220.
- [4] J.Y. Kim, G. Magesh, D.H. Youn, J.W. Jang, J. Kubota, K. Domen, J.S. Lee, Single-crystalline, wormlike hematite photoanodes for efficient solar water splitting, *Sci. Rep.* 3 (2013) 2681.
- [5] V.V. Jadhava, S.A. Patil, D.V. Shindeb, S.D. Waghmare, M.K. Zatea, R.S. Manea, S.-H. Hanb, Hematite nanostructures: morphology-mediated liquefied petroleum gas sensors, *Sensors Actuators B* 188 (2013) 669–674.
- [6] J. Zhu, K.Y. Simon Ng, D. Deng, Hollow cocoon-like hematite mesoparticles of nanoparticle aggregates: structural evolution and superior performances in lithium ion batteries, *ACS Appl. Mater. Interfaces* 6 (2014) 2996–3001.
- [7] H. Liang, X. Xu, W. Chen, B. Xuab, Z. Wang, Facile synthesis of hematite nanostructures with controlled hollowness and porosity and their comparative photocatalytic activities, *Cryst. Eng. Commun.* 16 (2014) 959–963.
- [8] J. Velev, A. Bandyopadhyay, W.H. Butler, S. Sarker, Electronic and magnetic structure of transition-metal-doped α -hematite, *Phys. Rev. B* 71 (2005) 205208.
- [9] J.S. Chen, T. Zhu, C.M. Li, X.W. Lou, Building hematite nanostructures via oriented attachment, *Angew. Chem. Int. Ed.* 50 (2011) 650.
- [10] K. Woo, H.J. Lee, J.P. Ahn, Y.S. Park, Sol–gel mediated synthesis of Fe₂O₃ nanorods, *Adv. Mater.* 15 (2003) 1761.
- [11] M.G. Sung, K. Sassa, T. Tagawa, T. Miyata, H. Ogawa, M. Doyama, S. Yamada, S. Asai, Application of a high magnetic field in the carbonization process to increase the strength of carbon fibers, *Carbon* 40 (2002) 2013–2020.
- [12] H. Zhou, S.S. Wong, A facile and mild synthesis of 1-D ZnO, CuO, and α -Fe₂O₃ nanostructures and nanostructured arrays, *ACS Nano* 2 (2008) 944.
- [13] L. Yuan, R. Cai, J.I. Jang, W. Zhu, C. Wang, Y. Wang, G. Zhou, Morphological transformation of hematite nanostructures during oxidation of iron, *Nanoscale* 5 (2013) 7581.
- [14] W. Merchan-Merchan, A.V. Saveliev, A.M. Taylor, High rate flame synthesis of highly crystalline iron oxide nanorods, *Nanotechnology* 19 (2008) 125605.
- [15] S. Musić, S. Krehula, S. Popović, Ž. Skoko, Some factors influencing forced hydrolysis of FeCl₃ solutions, *Mater. Lett.* 57 (2003) 1096.
- [16] Z. Pu, M. Cao, J. Yang, K. Huang, C. Hu, Controlled synthesis and growth mechanism of hematite nanorhombhedra, nanorods and nanocubes, *Nanotechnology* 17 (2006) 799–804.
- [17] T.K. Van, H.G. Cha, C.K. Nguyen, S.W. Kim, M.H. Jung, Y.S. Kang, Nanocrystals of hematite with unconventional shape-truncated hexagonal bipyramid and its optical and magnetic properties, *Cryst. Growth Des.* 12 (2012) 862–868.
- [18] Y.N. Li, P. Zhang, Z. Guo, H. Liu, Shape evolution of α -Fe₂O₃ and its size-dependent electrochemical properties for lithium-ion batteries, *J. Electrochem. Soc.* 155 (2008) A196.
- [19] W.F. Tan, Y.T. Yu, M.X. Wang, F. Liu, L.K. Koopal, Shape evolution synthesis of mono-disperse spherical, ellipsoidal, and elongated hematite (α -Fe₂O₃) nanoparticles using ascorbic acid, *Cryst. Growth Des.* 14 (2014) 157–164.
- [20] G. Du, P. Liu, W. Guo, Y. Han, J. Zhang, Z. Jang, Z. Ma, J. Han, Z. Liu, K. Yao, The influence of high magnetic field on electric-dipole emission spectra of Eu³⁺ in different single crystals, *J. Mater. Chem. C* 1 (2013) 7608–7613.
- [21] J.H. Wang, Y.W. Ma, K. Watanabe, Magnetic-field-induced synthesis of magnetic γ -Fe₂O₃ nanotubes, *Chem. Mater.* 20 (2008) 20–22.
- [22] R.S.M. Rikken, R.J.M. Nolte, J.C. Maan, J.C.M. Hest, D.A. Wilsonb, P.C.M. Christianen, Manipulation of micro- and nanostructure motion with magnetic fields, *Soft Matter* 10 (2014) 1295.
- [23] D. Fragouli, R. Buonsanti, G. Bertoni, C. Sangregorio, C. Innocenti, A. Falqui, D. Gatteschi, P.D. Cozzoli, A. Athanassiou, R. Cingolani, Dynamical formation of spatially localized arrays of aligned nanowires in plastic films with magnetic anisotropy, *ACS Nano* 4 (2010) 1873–1878.
- [24] Y. Lu, Y.D. Yin, Y.N. Xia, Three-dimensional photonic crystals with non-spherical colloids as building blocks, *Adv. Mater.* 13 (2001) 415–420.
- [25] T. Ding, K. Song, K. Clays, C.H. Tung, Fabrication of 3D photonic crystals of ellipsoids: convective self-assembly in magnetic field, *Adv. Mater.* 21 (2009) 1936–1940.
- [26] K. Cheng, Q.W. Chen, Z.D. Wu, M.S. Wang, H. Wang, Colloids of superparamagnetic shell: synthesis and self-assembly into 3D colloidal crystals with anomalous optical properties, *Cryst. Eng. Commun.* 13 (2011) 5394–5400.
- [27] H. Wang, Q.W. Chen, Y.F. Yu, K. Cheng, Assembly of superparamagnetic colloidal nanoparticles into field-responsive purple Bragg reflectors, *Dalton Trans.* 40 (2011) 4810 H. Wang, Y.F. Yu, Q.W. Chen, K. Cheng, *Dalton Trans.* 40 (2011) 559–563.
- [28] H. Wang, Y.F. Yu, Q.W. Chen, K. Cheng, Carboxyl-functionalized nanoparticles with magnetic core and mesopore carbon shell as adsorbents for the removal of heavy metal ions from aqueous solution, *Dalton Trans.* 40 (2011) 559–563.
- [29] H. Wang, Q.W. Chen, Y.F. Yu, K. Cheng, Y.B. Sun, Size and solvent-dependent magnetically responsive optical diffraction of carbon-encapsulated superparamagnetic colloidal photonic crystals, *J. Phys. Chem. C* 115 (2011) 11427.
- [30] H. Hu, C. Chen, Q. Chen, Magnetically controllable colloidal photonic crystals: unique features and intriguing applications, *J. Mater. Chem. C* 1 (2013) 6013.
- [31] W.L. Zhou, A. Kumbhar, J. Wiemann, J.Y. Fang, E.E. Carpenter, C.J. ÓConnor, Gold-coated iron (Fe@Au) nanoparticles: synthesis, characterization, and magnetic field-induced self-assembly, *J. Solid State Chem.* 159 (2001) 26–31.
- [32] V. Raman, A. Bose, B.D. Olsen, T.A. Hatton, Long-range ordering of symmetric block copolymer domains by chaining of superparamagnetic nanoparticles in external magnetic fields, *Macromolecules* 45 (2012) 9373.
- [33] A. Sinha, S. Nayar, B.K. Nath, D. Das, P.K. Mukhopadhyay, Magnetic field induced synthesis and self-assembly of super paramagnetic particles in a protein matrix, *Colloids Surf. B* 43 (2005) 7–11.
- [34] H. Singh, P.E. Laibinis, T.A. Hatton, Rigid, superparamagnetic chains of permanently linked beads coated with magnetic nanoparticles. Synthesis and rotational dynamics under applied magnetic fields, *Langmuir* 21 (2005) 11500–11509.
- [35] P. Kumar, R.K. Singh, N. Rawat, P.B. Barman, S.C. Katyal, H. Jang, H.N. Lee, R. Kumar, A novel method for controlled synthesis of nanosized hematite (α -Fe₂O₃) thin film on liquid–vapor interface, *J. Nanoparticle Res.* 15 (2013) 1532.
- [36] J. Velev, A. Bandyopadhyay, W.H. Butler, S. Sarker, Electronic and magnetic structure of transition-metal-doped α -hematite, *Phys. Rev. B* 71 (2005) 205208.
- [37] D. Tripathy, A.O. Adeyeye, C.B. Boothroyd, S.N. Piramanayagam, Magnetic and transport properties of Co-doped Fe₂O₃ films, *J. Appl. Phys.* 101 (2007) 013904.
- [38] G.A. Petrakovskii, I. Pankrats, V.M. Sosnin, V.N. Vasil'ev, Effect of doping with Co²⁺ ions on the resonant and static magnetic properties of hematite, *Sov. Phys. - JETP* 58 (1983) 403.
- [39] C. Saragovi, J. Arpe, E. Sileo, R. Zysler, L.C. Sanchez, C.A. Barrero, Changes in the structural and magnetic properties of Ni-substituted hematite prepared from metal oxinates, *Phys. Chem. Miner.* 31 (2004) 625.
- [40] L. Hu, R. Zhang, Q. Chen, Synthesis and assembly of nanomaterials under magnetic fields, *Nanoscale* 6 (2014) 14064–14105.
- [41] M. Tadić, D. Markovic, V. Spasojević, V. Kusigerski, M. Remškar, J. Pirnat, Z. Jagličić, Synthesis and magnetic properties of concentrated α -Fe₂O₃ nanoparticles in silica matrix, *J. Alloys Compd.* 441 (2007) 291–296.
- [42] M. Tadić, M. Panjan, V. Damjanovic, I. Milosevic, Magnetic properties of hematite (α -Fe₂O₃) nanoparticles prepared by hydrothermal synthesis method, *Appl. Surf. Sci.* 320 (2014) 183–187.
- [43] M. Tadić, N. Čitaković, M. Panjan, Z. Stanojević, D. Markovic, D. Jovanovic, V. Spasojević, Synthesis, morphology and microstructure and magnetic properties of hematite submicron particles, *J. Alloys Compd.* 509 (2011) 7639–7644.
- [44] M. Tadić, N. Čitaković, M. Panjan, B. Stanojević, D. Markovic, D. Jovanovic, V. Spasojević, Synthesis, morphology and microstructure of pomegranate-like hematite (α -Fe₂O₃) superstructure with high coercivity, *J. Alloys Compd.* 543 (2012) 118–124.
- [45] J. Velev, A. Bandyopadhyay, W.H. Butler, S. Sarker, Electronic and magnetic structure of transition-metal-doped α -hematite, *Phys. Rev. B* 71 (2005) 205208 (2005).
- [46] A. Akbar, S. Riaz, R. Ashraf, S. Naseem, Magnetic and magnetization properties of Co-doped Fe₂O₃ thin films, *IEEE Trans. Magn.* 50 (8) (2014) 2201204.
- [47] W.A. Kaczmarek, Structural and magnetic properties of cobalt-doped iron oxide particles prepared by novel mechanochemical method, *J. Magn. Magn. Mater.* 157 (158) (1996) 264–265.
- [48] R. Wang, P. Li, C. Chen, Template-free synthesis and self-assembly of aligned nickel nanochains under magnetic fields, *J. Nanosci. Nanotechnol.* 11 (2011) 1–5.
- [49] B. Jia, L. Gao, Growth of well-defined cubic hematite single crystals: oriented aggregation and Ostwald ripening, *Cryst. Growth Des.* 8 (2008) 1372.

Noise Robustness Analysis of Sparse Representation based Classification Method for Non-stationary EEG Signal Classification

Younghak Shin¹, Seungchan Lee¹, Minkyu Ahn², Hohyun Cho¹, Sung Chan Jun¹, and Heung-No Lee^{1*}

¹*School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea*

²*Department of Neuroscience, Brown University, Rhode Island, USA*

E-mail: heungno@gist.ac.kr

ABSTRACT

In the electroencephalogram (EEG)-based brain-computer interface (BCI) systems, classification is an important signal processing step to control external devices using brain activity. However, scalp-recorded EEG signals have inherent non-stationary characteristics; thus, the classification performance is deteriorated by changing the background activity of the EEG during the BCI experiment. Recently, the sparse representation-based classification (SRC) method has shown a robust classification performance in many pattern recognition fields including BCI. In this study, we aim to analyze noise robustness of the SRC method to evaluate the capability of the SRC for non-stationary EEG signal classification. For this purpose, we generate noisy test signals by adding a noise source such as random Gaussian and scalp-recorded background noise into the original motor imagery based EEG signals. Using the noisy test signals and real online-experimental dataset, we compare the classification performance of the SRC and support vector machine (SVM). Furthermore, we analyze the unique classification mechanism of the SRC. We observed that the SRC method provided better classification accuracy and noise robustness compared with the SVM method. In addition, the SRC has an inherent adaptive classification mechanism that makes it suitable for time-varying EEG signal classification for online BCI systems.

Keywords: Brain-computer interface (BCI), Electroencephalogram (EEG), Sparse representation based classification (SRC), Common spatial pattern (CSP), Non-stationarity.

1. Introduction

Brain-computer interface (BCI) systems provide a new communication and control channel between people and external devices [1]. In these systems, users can control an external device using their intention or imagination without making any real muscle movement. Therefore, these systems are very helpful for people who are suffering from severe motor diseases. The electroencephalogram (EEG) is widely used for measuring brain signals in BCI systems because of its low cost, no space restriction, and high temporal resolution compared with other equipment such as functional magnetic resonance imaging (fMRI) and magneto encephalogram (MEG) [2,3]. However, scalp-recorded EEG signals are very sensitive to noise. In particular, in the case of motor imagery based BCI, which uses induced EEG signals while the subject imagines limb movements [2,3], the instability of imagery task, non-stationarity of signals, and lack of concentration are among main obstacles to effectively process the EEG signals. In addition, it is difficult to collect a large set of training samples because of the subject's fatigue. The raw EEG signals are associated with high dimension owing to the large number of EEG channels; hence, it is difficult to collect volume of data samples that are large enough for good training. Therefore, EEG signal processing is very important and many research efforts have been focused on this issue [5–7].

The signal processing steps in BCI can be categorized as preprocessing, feature extraction, and classification. In the preprocessing step, the artifact detection and rejection are conducted. The purpose of feature extraction is to make a meaningful low-dimensional data, i.e., a feature vector, from the original high-dimensional data. This feature vector should be distinguishable for different classes. Typically, the feature extraction is performed using a dimensionality reduction method. The principal component analysis (PCA), independent component analysis (ICA), and common spatial pattern (CSP) are

popular methods for dimensionality reduction in the motor imagery based BCI systems [7,20].

Another important signal processing step is classification. In the BCI systems, the purpose of classification is to translate the extracted feature of a user's intention into a computer command, which can then be used to control external devices. Typically, this translation is done using the classification algorithms, which are adopted from pattern recognition area. Frequently used classification methods in the EEG based BCI systems are linear classifiers such as linear discriminant analysis (LDA) and support vector machine (SVM) [6]. In many BCI studies, the SVM has been recognized as a robust classification method with generalization ability and has shown to provide the best classification results [6,14,15].

Recently, in the field of pattern recognition, the concept of sparse representation based classification, namely SRC, has been introduced [8]. The basic idea of SRC is to parsimoniously represent a test signal \mathbf{y} via the so-called sparsification step, i.e., $\mathbf{y} = \mathbf{A}\mathbf{x}$, where \mathbf{A} is a dictionary whose columns are a collection of training signals. This sparsification step leads to the representation of the test signal \mathbf{y} with the training signals from the same class predominantly. The L1 minimization algorithm is employed to perform the sparse representation of the test signal with a given set of training signals.

The robust classification performance of the SRC framework has been shown in various applications such as face recognition [9,12,13,24], digit classification [8], and speech recognition [10]. Particularly, in [9], Yang *et al.* presented that SRC obtains robust face recognition performance for occlusion and corruption on facial images. In addition, SRC has been successfully applied to the EEG based BCI application [11] and EEG based vigilance detection [28]. However, in the EEG signal classification, SRC is rarely studied. The previous SRC study for the motor imagery

based EEG signal classification focused on algorithm construction and evaluated the classification performance compared with a conventional classifier such as LDA in [11]. To the best of our knowledge, there has been no literature to systematically evaluate the noise robustness and classification characteristics of SRC for the scalp recorded EEG signals.

It is well known that EEG signals are non-stationary. The non-stationarity can be observed during the change in alertness and wakefulness, eye blinking, and in the event-related potential (ERP) and evoked potential (EP) such as motor imagery signals [32]. Because of the non-stationarity of the EEG, we can observe that the test feature positions vary from the original training feature positions in the feature space [6,16]. This is one of the major obstacles in EEG signal classification. Thus, a classifier that is optimized for a particular training data may not work for online BCI with a new test data.

Recently, extensive research efforts have been devoted to overcome the non-stationary issue in the motor imagery based EEG classification. In [38–40], robust feature extraction methods were proposed for common spatial pattern (CSP), which is the most widely used technique for feature extraction in the motor imagery BCI. In the classification stage, supervised and unsupervised adaptive classification schemes were studied for the conventional LDA and SVM methods [16,27,41].

In this study, our aim is to evaluate the robustness of SRC for non-stationary EEG signal classification. First, we compare the classification performance, i.e., classification accuracy and computation time, of the SRC with SVM, which has been known as the state of the art classifier in many studies. Second, we evaluate the noise robustness of the SRC and SVM methods. For this purpose, we generate noisy test signals which have different feature distribution with original test signals. The noisy test signals are generated with the addition of random Gaussian noise and scalp recorded background EEG signal into the original test signal. Then, we assess the noise robustness of both SRC and SVM methods. Third, in addition to the simple performance comparison, we examine working mechanism of SRC by analyzing advantages and disadvantages as the role of classifier compared with the conventional SVM. Moreover, we discuss why SRC outperforms SVM for the noisy test signal. Finally, we evaluate the SRC method using an online experimental dataset where non-stationarity occurs from training to testing sessions. Our work is intended to provide evaluation and analysis of SRC to researchers who want to apply the SRC framework to non-stationary EEG signal classification.

This paper is organized as follows: In section 2, the experiment and EEG signal processing methods such as feature extraction and classification are described. In addition, noise robustness analysis method is explained in this section. Section 3 shows the experimental results. In section 4, discussions and analysis are provided. Finally, we conclude this paper in section 5.

2. Methods

2.1. Experiment

In this study, to evaluate and analyze the SRC method, we perform two-class EEG based motor imagery experiment. Twenty healthy subjects (11 male and 9 female subjects whose average age is 24.05 ± 3.76) participated in this experiment. Therefore, we collected 20 motor imagery datasets. Each dataset contains EEG signals generated from the left and right hand motor imagery experiment. Experiment included five runs. One

run consisted of 20 trials for each class. Thus, the total number of trials was 100 for each instruction (class).

Fig. 1 shows a single trial experimental paradigm of our motor imagery experiment. Cue line indicates the starting point of motor imagery. One trial consisted of 4–6 sec of resting time period and 3 sec of imagery time period. In the resting period, a blank screen appeared on the monitor. The resting time was randomly selected in the range of 4 to 6 sec. In the imagery period, one of the motor imagery instructions was represented at the center of the screen, then subjects imagined their left or right hand movements for tasks such as grasping and releasing hand. In each trial, instruction was randomly selected from the left and right hand class.

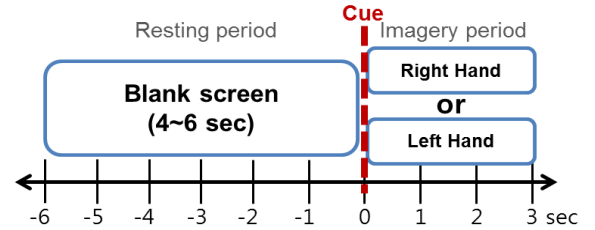


Fig. 1. Single trial time procedure of motor imagery experiment.

In addition, we recorded resting state EEG signals for each subject to estimate the subject-specific background noise. In this recording, subjects were instructed to open their eyes for 60 sec without any experimental task.

These experimental datasets were recorded by an active electrode cap. We used Active Two EEG measurement system made by Biosemi, Inc. The sampling rate for these datasets was 512 samples per second, and the number of EEG channels was 64. The channel positions were selected from the international 10/20 standard.

2.2. Preprocessing and Feature Extraction

Preprocessing and feature extraction steps are common to both SRC and SVM classification algorithms. Using the motor imagery dataset obtained from each subject, we perform the data preprocessing. Before preprocessing, raw EEG signals are segmented. After an instruction (left or right hand) appears on the screen, the time samples from 1 to 2 sec were collected for all trial data. We apply the band pass filter to the trial data to eliminate the frequencies that are not related to motor imagery signals. In this study, sensorimotor rhythm, 8 to 15 Hz, is used for band pass filtering [11]. For fair comparison of the classification performance, we fixed the time and frequency range for all subjects. Then, we reduce the dimension of EEG signal using the common spatial pattern (CSP) filtering, which is a widely used feature selection method for motor imagery based BCI systems [5,11,20]. CSP filters maximize the variance of the spatially filtered signal under one class condition while minimizing it for the other class condition. The CSP filtering algorithm finds the filters, $\mathbf{W} \in \mathbb{R}^{C \times C} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C]$ which transforms the EEG data $\mathbf{X} \in \mathbb{R}^{C \times S}$ (C and S denote the number of EEG channels and time samples) into a spatially filtered space: $\mathbf{X}_{CSP} = \mathbf{W}^T \cdot \mathbf{X}$. Generally, \mathbf{W} is computed by simultaneous diagonalization of the covariance matrices, Σ_1 and Σ_2 , of the two classes of data. This is equivalent to solving the generalized eigenvalue problem, i.e., $\Sigma_1 \mathbf{w} = \lambda \Sigma_2 \mathbf{w}$, where λ is the eigenvalue. In practice, the first and last n columns of the \mathbf{W} correspond to the n largest and n smallest eigenvalues that are used for CSP filtering. However, the optimal number of CSP filters, $m = 2n$, which shows the maximum classification accuracy varies and has

to be chosen empirically [20]. After CSP filtering, for each CSP filtered trial, we compute the frequency power of sensorimotor rhythm (8–15 Hz) which is the widely used band power (BP) feature in motor imagery based BCI classification [6,11]. Various feature types including BP, AR (autoregressive) [6], and functional connectivity [42] can be used for motor imagery classification. However, in this study, we focus on the evaluation of classification methods using a common feature type.

2.3. Classification Methods

2.3.1. Sparse Representation based EEG Signal Classification

The SRC framework was introduced to the EEG based motor imagery BCI application in [11]. There, the SRC method showed a better classification accuracy over the conventional LDA method.

In the SRC method, dictionary is first formed using the processed training feature. Let $\mathbf{A}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \dots, \mathbf{a}_{i,N_i}]$ be the class-dictionary for classes $i = L$ and R where L and R represent class information of left hand and right hand motor imagery respectively, and N_i is the total number of training trials, i.e., $N_i = 99$ for each class in this study. Then, the final dictionary \mathbf{A} is formed by $\mathbf{A} := [\mathbf{A}_L; \mathbf{A}_R]$. Each column vector $\mathbf{a} \in \mathbb{R}^{m \times 1}$ where m is the number of applied CSP filters. In this study, we used 64 EEG channels; thus, m is varied from 2 to 64. Each entry of \mathbf{a} is obtained by computing the frequency power of sensorimotor rhythm after the CSP filtering. Let \mathbf{y} denote a testing feature with the same dimension as \mathbf{a} .

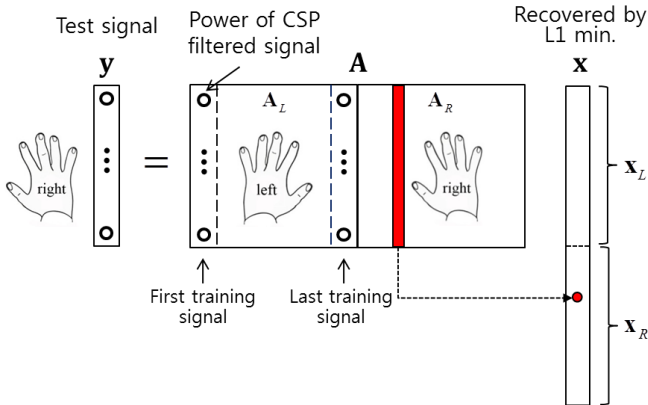


Fig. 2. Dictionary design and linear sparse representation model for SRC.

Fig. 2 shows the formed dictionary \mathbf{A} and model of sparse representation for motor imagery based EEG signals. In this example, a certain test feature \mathbf{y} of the right hand class can be sparsely represented with a linear combination of training feature of the right hand class. This is represented by the nonzero scalar coefficients \mathbf{x} in the position of corresponding class.

The SRC method can be summarized in the following two steps. The first step is to sparsely represent \mathbf{y} using \mathbf{A} via L1 norm minimization. This step is the sparsification step:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x} \quad (1)$$

where \mathbf{x} is a scalar coefficient vector and $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the dictionary.

Note that the linear system in Eq. (1) is under-determined. The literature of compressive sensing (CS) shows that the L1 minimization algorithm can solve this optimization problem in polynomial time [17,18,21].

The second step is to classify the test signal via minimum residual. This step is the identification step:

$$\text{class}(\mathbf{y}) = \min_i r_i(\mathbf{y}) \quad (2)$$

where $r_i(\mathbf{y}) := \|\mathbf{y} - \mathbf{A}_i \mathbf{x}_i\|_2$, \mathbf{x}_i is the scalar coefficient vector corresponding to the class i .

2.3.2. Support Vector Machine

SVM is a well-known classification method in the area of pattern recognition and machine learning. In the BCI field, the SVM has shown a robust classification performance in many experimental studies [6,14,15]. SVM is recognized for its excellent generalization performance, i.e., small error rate for test data. This property is achieved through the idea of margin maximization. As shown in Fig. 3, the margin d is twice the distance between the support vector (the black and white circles that are on the dashed line) and the decision hyperplane. The hyperplane can be described by a weight vector \mathbf{w} and a bias b . The SVM finds the decision hyperplane by solving the following optimization problem [19]:

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_n \xi_n, \\ & \text{subject to} \quad t_n (\mathbf{w}^T \Phi(\mathbf{y}_n) + b) \geq 1 - \xi_n \\ & \quad \quad \quad \xi_n \geq 0, n = 1, \dots, N \end{aligned} \quad (3)$$

where \mathbf{y}_n is the training feature vector, $t_n \in \{+1, -1\}$ is the class information and n indicates the training trial number. To consider the training error, a slack variable ξ and a regularization parameter C are included [19]. Using ξ , we can consider the training error which is positioned inside the support vectors. C is a user defined regularization parameter to control the importance between the maximum margin and the training error.

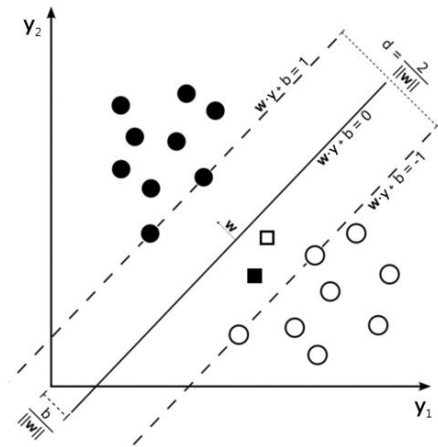


Fig. 3. The main idea of SVM. The SVM algorithm tries to find the decision hyperplane, which has the maximum margin d .

In the SVM optimization problem, mapping function $\Phi(\cdot)$ can be used to map an inseparable feature vector onto a higher-dimensional space using a kernel function $K(\mathbf{x}, \mathbf{y})$. In BCI research, the Radial Basis Function (RBF) kernel (4) is widely used and has shown robust classification performance [6,15]:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \quad (4)$$

Therefore, in this study, we consider a linear SVM and an RBF kernel based SVM for comparison of the classification performance with the SRC method. For both SVM algorithms, we use the MATLAB Bioinformatics Toolbox (SVMtrain) [23].

In the SVM algorithm, selection of parameters is important to obtain the robust performance. We optimize the regularization parameter C in (3) for linear SVM and kernel parameter σ in (4) with combination of C for RBF SVM. We adopt a coarse grid search method using cross-validation to find optimal parameters that provide the best classification accuracy [25]. In the exhaustive coarse grid search, we first find a better region on the loose grid, then fine grid search on that region is conducted. For two parameters C and σ , we set the same grid sequence as follows: C and $\sigma = [10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3]$. Then, for the best region, we optimize the parameters using a fine tuning.

2.4. Noise Robustness Analysis Method

In this study, we aim to evaluate the noise robustness of the SRC and SVM classification methods when our test data is contaminated by an additive random Gaussian noise and scalp recorded background noise. The ultimate goal of this evaluation is to assess the classification performance of both methods for non-stationary EEG signal. As it is known, EEG signals have inherent non-stationary characteristics. Therefore, BCI features vary from training sessions to test sessions during a BCI experiment [6,16,38]. There are many reasons to change EEG signals in the motor imagery task such as physical and mental drifts, misalignment of sensors, and task-irrelevant background activity [33,38]. During the imagery period in the motor imagery experiment, when we assume subjects exclusively perform motor imagery task, the task-irrelevant background activity can be the main reason for a change in EEG signals [16,38]. In [36–37], authors also considered the resting state signal as task-irrelevant noise in the motor imagery task. In addition, in [16], it was showed that EEG signals were changed from training to online testing sessions in feature space by changing the background activity. Therefore, in this context, we aim to model the modified noisy test signals by adding background activity estimated by the resting state recording into the original test signal.

For robustness analysis, we generate the modified test data by introducing two different noise sources such as white Gaussian and background noise into the original test data. Each noise source signal is separately applied to the EEG test data. Thus, we evaluate the classification performance of both classifiers for two types of noise corrupted test data. In result section, we show the position shift in the noisy test feature that is generated by the background noise (see Fig 10).

Fig. 4 shows the generation concept of the polluted noisy test data using one noise source. In the online BCI experiment, the power of EEG test data varies. Therefore, to evaluate the noise robustness of the classifiers systematically, we generate five different noisy test data with various SNR levels. Thus, we control the noise power of each noise source in five levels.

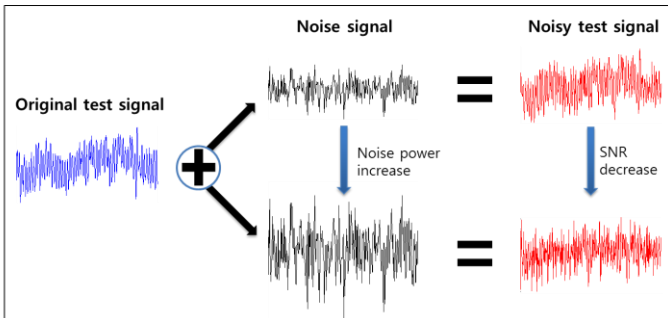


Fig. 4. Noisy test signal generation using different power of noise signal.

For the Gaussian noise, we control the noise power by varying the standard deviation of Gaussian distribution. For the

background noise, we use a scale factor α to control the noise power as follows:

$$\text{polluted test signal} = \text{test signal} + \alpha(\text{resting noise}) \quad (5)$$

For each subject's dataset, the classification performance of the SRC and SVM methods is evaluated using both types of noisy test data.

Random Gaussian noise is artificially generated by m -dimensional Gaussian distribution, i.e., $N_m(\mu, \sigma^2)$ where μ and σ^2 are the mean and variance. We use a MATLAB built-in function to generate the zero mean Gaussian distribution with different standard deviation σ . To make polluted EEG test data by Gaussian noise, we generate the same dimension of Gaussian noise to the segmented EEG signal, i.e., noise dimension is 64 by 512. We also apply the band pass filter to the generated Gaussian noise with 8–15 Hz cutoff frequency, which is used in the preprocessing of EEG signal.

Subject-specific background noise is measured by the EEG recording of the resting state. In this recording, subject is instructed to just open their eyes without any task for one minute. We apply the band pass filter to the recorded resting state signal. To make polluted EEG test data by background noise, we collect one-second time samples (512 samples) from the resting state signal. In this study, both classifiers are evaluated using 100 test trials. Therefore, we generate 100 noise signals using the moving window from the total resting state signal. The size of the moving window is 256 samples (0.5 second).

To evaluate and compare the classification accuracy of the SRC and SVM methods, we use the leave-one-out (LOO) cross-validation, which is useful for increasing the number of independent classification tests with a given set of limited data trials [22]. Thus, one trial out of 100 training trials is selected as the test trial, and the remaining trials are used as the training trials. This test is repeated for 100 times with different combination of training and test trials. To obtain noisy test trials, we apply 100 different noise signals for each noise source into the 100 test trials acquired from LOO cross-validation. Therefore, we have 100 noisy test trials for each Gaussian and background noise. In this study, we calculate the classification accuracy as follows:

$$\text{Accuracy(\%)} = \frac{\text{the number of correct test trials}}{\text{the number of total test trials}} \times 100 \quad (6)$$

3. Results

3.1. Comparison of Classification Results

First, we evaluate the classification accuracy of the SRC and SVM methods for the original experimental datasets that are not contaminated by noise sources. Fig. 5 shows the comparison result of the classification accuracy for the SRC, linear SVM, and RBF SVM. For each subject, we computed the classification accuracy (in %) using the LOO cross-validation. We used 18 CSP filters for both classification methods, which are determined heuristically (see Fig. 6).

In Fig. 5, we observe that SRC achieves competitive classification accuracy over both linear and RBF kernel-based SVM. The classification accuracy of SRC was found to be better than linear SVM for 15 subjects and RBF SVM for 14 subjects over 20 subjects. In addition, the mean difference of the classification accuracy between the SRC and both SVM methods was statistically significant using the paired t-test ($p < 0.01$).

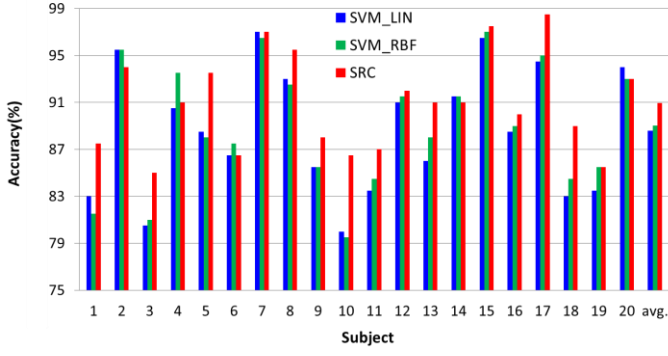


Fig. 5. Comparison of classification accuracy for the linear SVM, RBF kernel SVM, and SRC method using 20 non-noisy experimental datasets.

Moreover, we investigated the impact of varying the feature dimension on the non-noisy classification performance in each method (see Fig. 6). In this study, we used the CSP filtering as a feature selection method. The number of CSP filters (feature dimension) was varied from 2 to 64. Usually, the optimal number of CSP filters, which showed the maximum classification accuracy was chosen empirically. However, the optimal number of CSP filters was different depending on the classification method and dataset. Therefore, we evaluated the classification performance of each classification method when the feature dimension was varied. Fig. 6 shows the average classification accuracy over all subjects when the number of feature dimensions m was varied from 2 to 64. We found that the classification accuracy of SRC method consistently outperformed the linear and RBF kernel based SVM methods regardless of their feature dimension. There was not much difference in the classification accuracy between the SVM methods. However, the RBF SVM showed a better classification accuracy when the number of CSP filters was over 18.

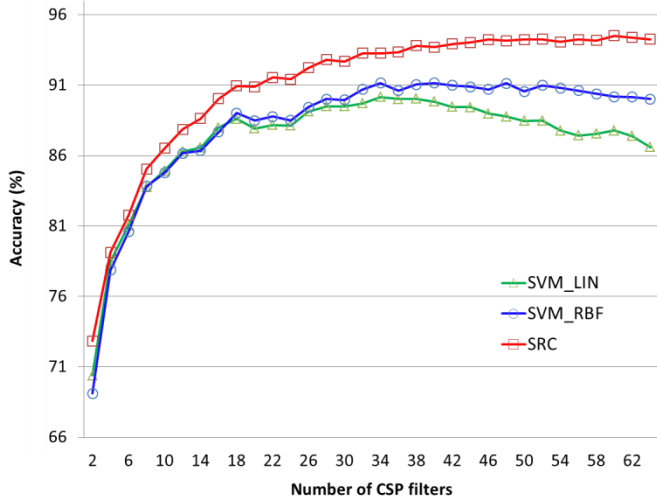


Fig. 6. Average classification accuracy over 20 non-noisy datasets when the number of CSP filters (feature dimension) is varied from 2 to 64.

We used the fixed 18 CSP filters for all classification methods that are shown in Fig. 5. However, the results in Fig. 6 shows that this number was not optimal for certain classification methods. When we used more CSP filters, the difference in the classification accuracy between the SRC and SVM methods was increased.

3.2. Classification Results for Noise Robustness

In this section, we evaluate noise robustness of the RBF kernel based SVM and SRC methods. For the noise robustness analysis, we used polluted test signals that were generated by adding two noise sources, i.e., white Gaussian noise and background noise, into the original test signal as mentioned in section 2.4.

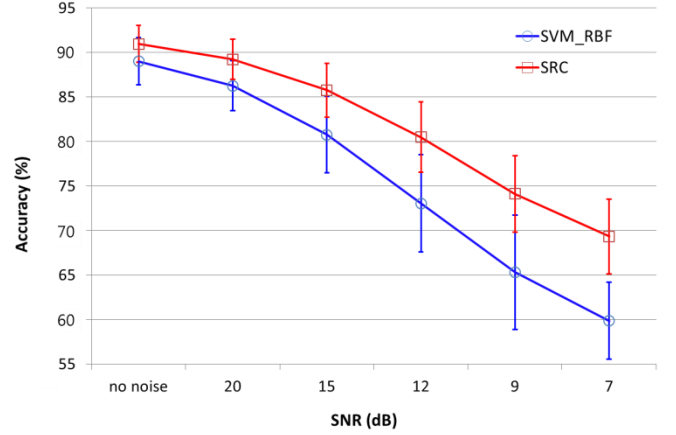


Fig. 7. Comparison of the average classification accuracy over 20 subjects. Average classification accuracy for Gaussian noise is represented as a function of SNR. Vertical line indicates the standard deviation of the accuracy for each SNR.

Fig. 7 shows the noise robustness results of the SRC and RBF kernel based SVM methods for the Gaussian noise. The average classification accuracy over all subjects was assessed when the noise power was varied. For the Gaussian noise, we controlled the noise power by changing the standard deviation, and the SNR was computed for different noise powers. In this study, SNR computation was defined as follows:

$$\text{SNR(dB)} = 10 \log_{10} \left(\frac{P_S}{P_N} \right) \quad (7)$$

where P_S and P_N indicate the signal and noise power, respectively. For the SNR computation, we investigated the average SNR over all the channels and subjects. As shown in Fig. 7, we found that the classification accuracy of SRC was higher than that of the RBF SVM for all SNR cases. The difference in the classification accuracy between the SRC and RBF SVM was increased with the SNR increase.

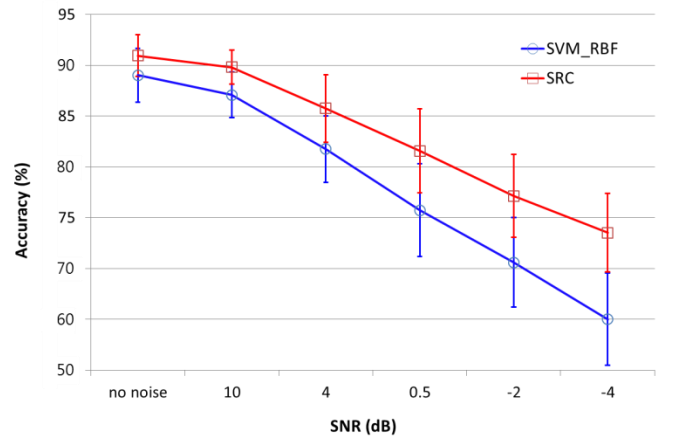


Fig. 8. Comparison of the average classification accuracy over 20 subjects. Average classification accuracy for background noise is represented as a function of SNR.

Similarly, Fig. 8 shows the noise robustness results of the SRC and RBF kernel based SVM methods for the background

noise, which was measured by the recorded resting state. For the background noise, the noise power was controlled by scale factor α (see Eq. (5)). It was found that the classification accuracy of SRC was higher than that of the RBF SVM for all SNR cases. In addition, when the noise power increased, the accuracy difference between the SRC and SVM increased. For example, in the noiseless case, the average accuracy difference between the SVM and SRC was 1.9%. However, in the case of 0.5 and -4dB SNR, the difference was 5.8% and 8.5%. This means that the SRC method was more robust than the SVM for the polluted test signal in the background noise case.

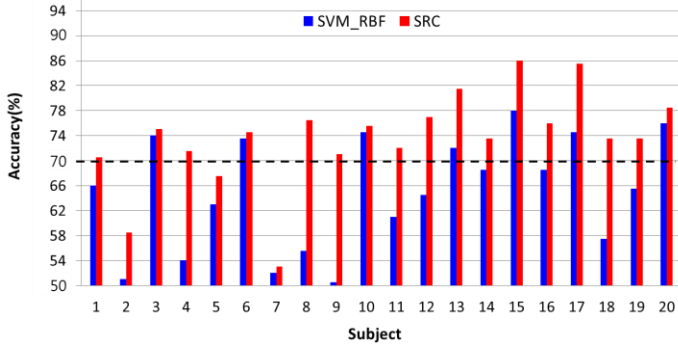


Fig. 9. Classification accuracy of RBF based SVM and SRC method for polluted test data by background noise (-4dB).

In two-class classification problems, the theoretical chance level is 50%. However, in many EEG based BCI studies [26,34,35], at least 70% classification accuracy is considered as a threshold for an acceptable communication and device control. In Fig. 9, we examine the classification performance for the polluted test data. Fig. 9 shows the classification accuracy of all subjects for the -4dB SNR for background noise cases shown in Fig. 8. The threshold of 70% classification accuracy is represented by black dotted line. For this threshold, the SVM has seven datasets that are over the threshold and the SRC has seventeen datasets. This means that for the noisy test data, 10 more subjects can use a reliable BCI system with the SRC compared to the SVM method.

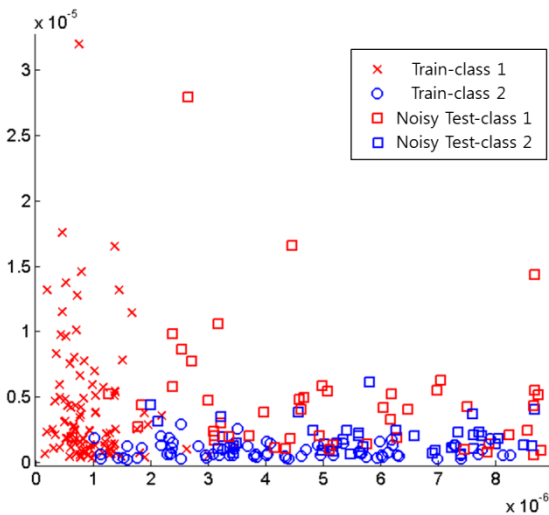


Fig. 10. Scatter plot of training data and noisy test data in two-dimensional feature space (2 CSP filters) for one subject dataset. Noisy test data are generated using background noise with 4 dB SNR.

Fig. 10 shows an example of training and polluted test features for one subject dataset. In this example, the background noise

with 4 dB SNR (shown in Fig. 8) was used for the polluted test data. The positions of noisy test features (red and blue squares in Fig. 10) in two-dimensional feature space were relocated from the positions of the original training features (red x-marks and blue circles) to places with a particular direction. This represents a typical situation that occurs in real-time BCI scenario where the online test data has different background noise compared to the training data [16]. In this study, the positions of the noisy test features were varied according to the SNR of the test data.

4. Discussions

4.1. Comparison of Classification Mechanism

In this section, we examine the algorithmic difference between the SRC and SVM methods as the role of signal classification. Fig. 11 shows the classification algorithms for both methods. Feature vectors for the training data were used as an input for both classification algorithms.

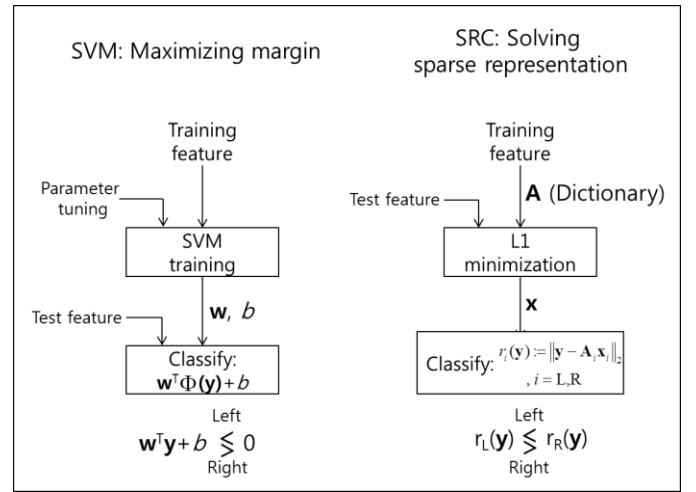


Fig. 11. Comparison of the SVM and SRC classification algorithm.

In the SVM algorithm, the input feature data and model parameters were used and the training was performed to find the parameters \mathbf{w} and b for decision boundary as shown in Eq. (3). Based on the boundary, the test feature was classified. Thus, the y class information was determined by the decision boundary.

In the SRC algorithm, the dictionary was simply formed by collecting the input training feature vectors as the columns of the dictionary. Then, using the dictionary, sparse representation was performed for each test data. Thus, scalar coefficient vector \mathbf{x} was obtained by solving L1 minimization as shown in Eq. (1). Using \mathbf{x} , class information was determined by computing the residual $r(\mathbf{y})$ in Eq. (2).

Our aim was to highlight the important difference of the classification mechanism of the SRC and SVM methods as follows:

- In SVM, a fixed decision rule (decision boundary) was obtained for the entire set of training signals. Then, for each test signal, this fixed decision rule was used for signal classification.
- In SRC, the sparse representation was adaptively performed for each test signal by utilizing all training signals in the dictionary.

4.2. Robustness Analysis of SRC

The experimental results presented in section 3 shows that

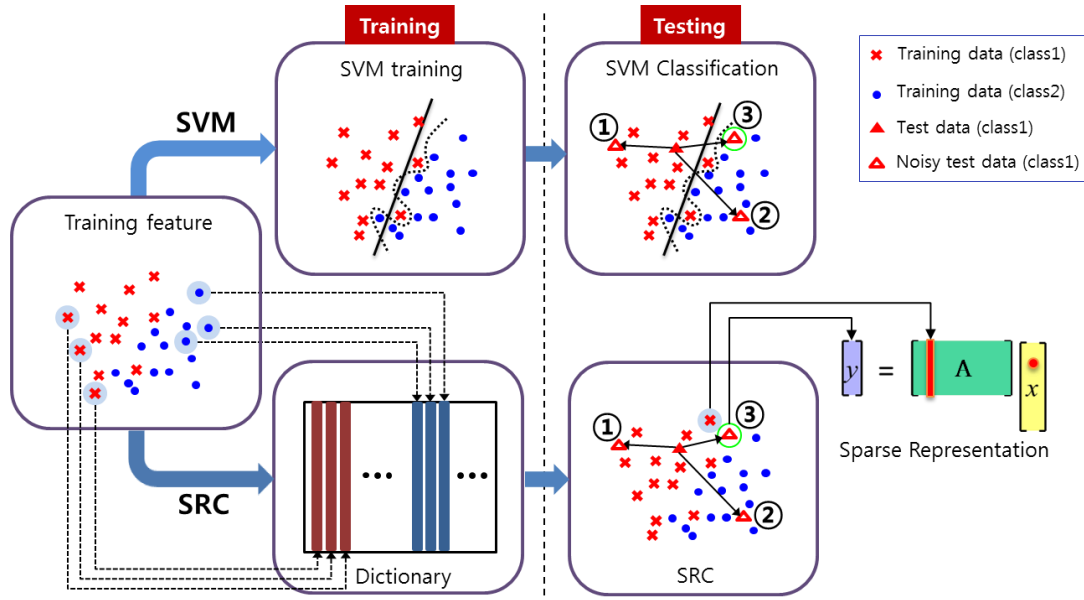


Fig. 12. Comparison of the classification procedure and characteristic of the SVM and SRC for the noisy test data. In the SVM part, black solid line and black dotted line indicate the decision boundaries for linear and RBF based SVM.

SRC had a better classification accuracy than the conventional SVM for motor imagery based EEG signal. In addition, SRC was more robust for polluted test data than SVM. In this section, we discuss the relationship between the classification performance and the difference in the classification mechanism for SRC and SVM methods.

Fig. 12 shows the concept of the classification strategy for the SVM and SRC using a toy example of polluted test data in two-dimensional feature space. In the SVM classification, decision hyperplane and non-linear decision boundary were presented for linear and RBF based SVM. For many conventional classifiers including SVM, the classifier was trained using training data; thus, the best decision rule was determined. Then, this classification rule was applied to each test data. However, as we have shown in Fig. 12, when the test data was polluted and shifted in feature space, the decision rule could not guarantee a satisfactory classification performance. On the other hand, in the SRC method, no classification rule was designed in the training part of SRC. Instead, a dictionary was formed by collecting feature vectors of the training data. Then, the sparse representation was performed for each test data using the dictionary. In addition, for the noisy test data, an independent classification task was performed in each classification by using all the training data instead of a fixed decision rule.

For a detailed analysis, we considered three possible cases of polluted test data that are presented by numbers ①, ②, and ③ in Fig. 12:

In the first case, test data was shifted away from the decision boundary and positioned at the same class feature space. In this case, both SVM and SRC correctly classified the noisy test data.

In the second case, the test data was positioned at a different class feature space of training data. Then, based on the decision boundary, the SVM classified the test data incorrectly. In the SRC method, the test data was more likely to be represented with different class training data. Thus, both classifiers were not working correctly.

Note that in the third case, similar to the second case, the test data was placed at a different class feature space. At the same time, the test data could be possibly positioned near the decision

boundary. In this case, based on the decision rule obtained from the training data, the SVM resulted in wrong classification.

When we used non-linear decision boundary, e.g., RBF SVM, as shown in black dotted line, this line was optimal for the training data. Thus, the classification error could be less than the linear decision hyperplane. However, for the polluted test data, the non-linear decision boundary was fixed. On the other hand, in the third case, SRC still had a chance for correct sparse representation with the same class training data as shown in Fig. 12. This was possible because the SRC method did not depend on a fixed decision rule that was obtained from the training data. Instead, for each classification of test data, the SRC method directly used all training data and performed sparse representation.

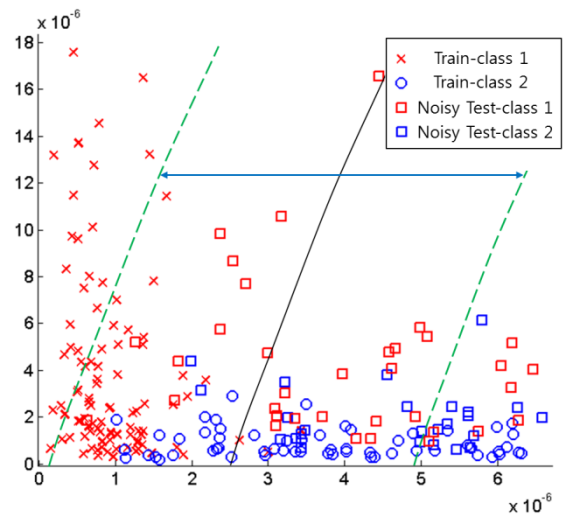


Fig. 13. Scatter plot of training data and noisy test data in two-dimensional feature space (2 CSP filters) for one subject data. Noisy test data are generated using background noise with 4 dB SNR.

To evaluate the validity of our analysis, we examined the same data shown in Fig. 10 in details. Fig. 13 shows an enlarged version of the scatter plot using the training and noisy test data. The black line indicates the obtained decision boundary from the RBF kernel based SVM. The region between the two green

dotted lines is chosen as the near area of the decision boundary. In this area, many miss-classification cases may occur for both classifiers. In addition, most of the polluted test data, which correspond to case ③ in Fig. 12 are located in this region.

For all noisy test data (i.e., 100 trials), the RBF SVM and SRC showed the classification accuracy of 56% and 62%, respectively. Because we used only two CSP filters for visualization, the classification accuracy was very low compared with the results given in Fig 8.

For the noisy test data, which are located between the green dotted lines, the RBF SVM showed 57% classification accuracy. However, the SRC showed an improved classification accuracy of 83%. In addition, when we only considered the noisy test data for case ③ examples, the RBF SVM had 18 miss-classification data. However, the SRC correctly classified 12 test data among 18 test data. Therefore, we confirmed that the noisy test data of case ③ were miss-classified from the fixed rule based SVM. On the other hand, for the same data, the SRC correctly classified many times with the effort of independent classification task for each test data using all training data.

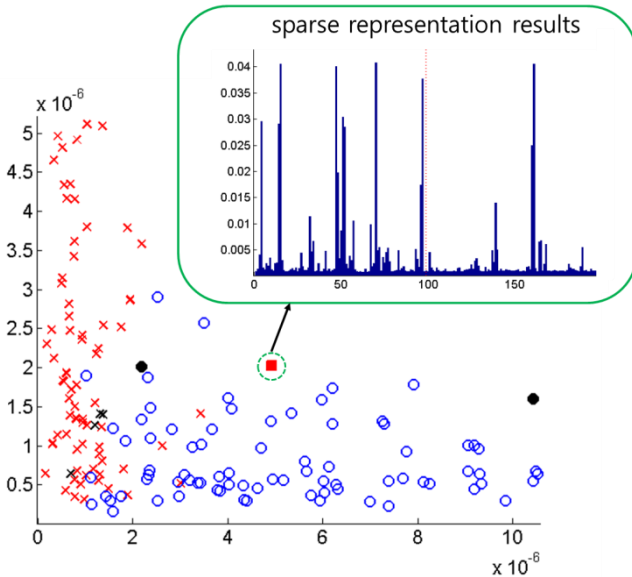


Fig. 14. Scatter plot of training data and noisy test data. The figure inside the green box indicates the sparse representation result of the noisy test data.

Fig. 14 shows one instance of the noisy test data that was not correctly classified by the SVM; however, was correctly classified by the SRC method. The test signal of class 1 is represented by a red square, which is located in the region between the green dotted lines shown in Fig. 13. The figure inside the green box shows the recovered coefficient \mathbf{x} from the SRC method. Using the trial numbers (x-axis of the figure inside the green box) with large coefficient values, we represented the corresponding trials by the black x-marks and circles in Fig. 14. Four largest coefficient values were selected for class 1. Two largest coefficient values were selected for class 2. As it can be seen, the noisy test trial of class 1 (red square) is located near the training trials of class 2. However, in the SRC method, using the coefficient \mathbf{x} , the test trial could be correctly classified from the minimum residual rule in Eq. (2). In addition, in each test trial, a different coefficient \mathbf{x} which represented the test data most compactly, was recovered by L1 minimization. Therefore, for the

case of time varying EEG signal classification, the SRC approach was much more appropriate to employ than the SVM method, which was based on the fixed decision rule.

An adaptive classification scheme for a conventional classifier such as LDA and SVM was studied to overcome the non-stationary problem of EEG signals [16,27,41]. In the adaptive techniques, typically decision boundary was updated (relearned) using collected labeled test data for a given duration. However, after designing new decision boundary, new test signal was dependent to the decision boundary. Thus, the adaptive scheme for the conventional classifier was still a decision rule based classification. Therefore, it could not be adaptively applied to each test signal. We think that some adaptation techniques for SRC [30–31], i.e., dictionary learning using collected signals, can be more efficient for real-time online BCI systems. Therefore, the comparison of the adaptive classification schemes between the SRC and conventional classifier is an interesting area for our future research.

4.3. Computation Time Analysis

In this section, we evaluate the computation time (running time) of the classification algorithms for the experimental datasets.

As it can be seen in Fig. 12, the most time consuming process of the SVM occurred while training the SVM. On the other hand, the most computation cost in the SRC algorithm occurred in L1 minimization step for sparse representation. Therefore, our evaluation for running time focused on the SVM training and L1 minimization step for the SRC algorithm. We used the *tic* and *toc* MATLAB commands to measure the start and end time of the SVM and SRC algorithms, respectively. We simulated all algorithms in the same environment using MATLAB 7.14 (R2012a) with 3.30 GHz processor and 8 GB memory.

For a single test trial, the average computation time for the SVM and SRC was 12.1 msec and 16.7 msec respectively. This computation time was averaged for 100 test trials of all subjects. However, in the case of online BCI classification, typically the SVM decision boundary was designed once using the training data. Then, all the test data was classified based on the decision boundary. On the other hand, independent classification task was performed for each test data in the SRC. Therefore, the computation time of the SRC method increased by the number of test trials. Thus, a robust classification performance of SRC included the cost of the computation time at each test trial.

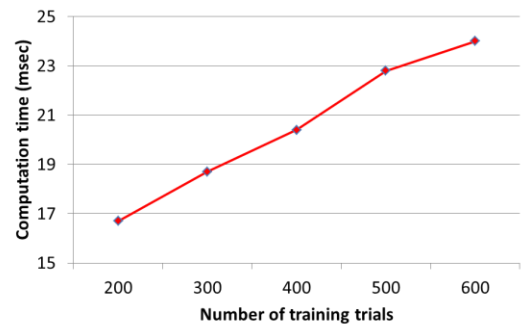


Fig. 15. Computation time of the SRC as a function for the number of training trials.

In this study, the size of the dictionary, i.e., the number of training trials, was 200. In this case, the computation time of the SRC was very small and negligible (16.7 msec). In addition, in Fig. 15, we display the average computation time as a function of the number of training trials. When the size of dictionary was increased, the difference of the computation time was just a few milliseconds. Therefore, this was not an important factor for an online classification in BCI systems. In addition, recently developed fast L1 minimization algorithms were used for the SRC method. In [24], authors showed that a few of the fast L1 minimization algorithms provided faster computation time than the conventional SRC method for large datasets of real face images.

Note that even though the computation time of the SVM was smaller than the SRC, the SVM required more effort to select a proper kernel and tune the model parameters for accurate classification results [6,25].

4.4. Online Data Analysis

In this study, we modeled noisy test data by adding two noise sources into the original trial data and controlled the noise power to evaluate the noise robustness of the SRC systematically. In this section, we aim to evaluate the SRC using online motor imagery experimental dataset. In this experiment, the training session and online test session were independently performed. Thus, in this evaluation, we used a non-stationary dataset from an online motor imagery experiment.

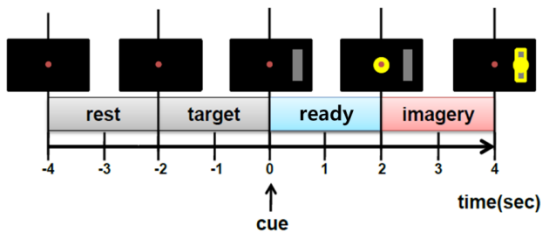


Fig. 16. Single trial procedure for online motor imagery experiment.

Five subjects participated in our online experiment. Right hand (R) and foot (F) motor imagery were performed for each subject. The sampling rate of these datasets was 512 samples per second, and the number of EEG channels was 64. The detailed experimental paradigm is illustrated in Fig. 16. The same paradigm was used for both training (calibration) and online testing (feedback) sessions. In each trial, the target bar was presented on 0 sec at the right or left side of the screen corresponding to the right or foot motor imagery. Two seconds after cue onset, the subject was instructed to perform the motor imagery task. During the training session, no feedback was provided. However, in the online testing session, the online feedback was provided in each trial. We collected 60 training trials and 75 online test trials for each class. After data segmentation from 2 to 4 sec, we performed the same preprocessing step that was used in section 2.2.

As shown in Fig. 17, using five online datasets, we evaluated the classification accuracy of the SRC and SVM_RBF. Even though size of the online dataset was small compared with the twenty offline datasets used in Fig. 5, we obtained consistent results. Thus, the SRC showed better mean classification accuracy than the SVM for the online datasets. Except one subject's dataset, which showed the same accuracy, the

classification accuracy of the SRC was better than the SVM_RBF method for four subjects.

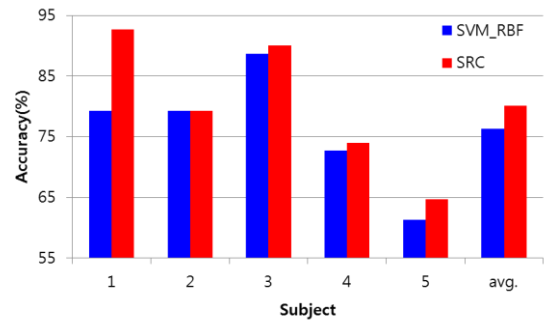


Fig. 17. Comparison of the classification accuracy of the SRC and SVM_RBF for online experimental dataset.

5. Conclusions

In this paper, we evaluated and analyzed the robustness of the SRC method against the non-stationarity of EEG signal classification. For this purpose, we generated noise corrupted EEG test signals using two noise sources such as random Gaussian noise and scalp recorded background noise. Then, we assessed the classification performance of the SRC when the noise power was varied. Using the experimental motor imagery based EEG and generated noisy test signals, we compared the classification results of the SRC with that of the SVM method, which has been considered as a robust classifier in many BCI studies. From the results, it was evident that the SRC showed superior noise robustness than the SVM for both Gaussian and background noise. Furthermore, the results of the online-experimental dataset showed that the classification accuracy of the SRC was better than the SVM. We analyzed that the robust classification accuracy of the SRC was due to a different classification approach compared with the conventional decision rule based SVM. Thus, the SRC showed an inherent adaptive classification mechanism for each test trial via optimal sparse representation of the training trials. In addition, we showed that the computation time of the SRC for a robust classification was on the order of milliseconds, which was acceptable for real time BCI systems.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (Do-Yak Research Program, No. 2010-0017944).

References

- [1] Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M., 2002. Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* 113 (6), 767–791.
- [2] Sellers, E., Donchin, E., 2006. A P300-based brain-computer interface: Initial tests by ALS patient. *Clin. Neurophysiol.* 117 (3), 538–548.
- [3] Pfurtscheller, G., Flotzinger, D., Kalcher, J., 1993. Brain-computer interface-a new communication device for handicapped persons. *J. Microcomput. Appl.* 16 (3), 293–299.
- [4] Wolpaw, J.R., McFarland, D.J., Neat, G.W., Forneris, C.A., 1991. An EEG-based brain-computer interface for cursor control. *Electroencephalogr. Clin. Neurophysiol.* 78 (3), 252–259.
- [5] Ramoser, H., Müller-Gerking, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng.* 8 (4), 441–447

- [6] Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain-computer interfaces. *J. Neural Eng.* 4 (2), R1–R13.
- [7] Dornhege, G., del, R. Millán J., Hinterberger, T., McFarland, D.J., Müller, K.R., 2007. *Toward Brain-Computer Interfacing*, The MIT Press, pp. 213–215.
- [8] Huang, K., Aviyente, S., 2006. Sparse representation for signal classification. *Adv. Neural Inf. Process. Syst.* 19, 609–616.
- [9] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2), 210–227.
- [10] Gemmeke, J.F., Virtanen, T., Hurmalainen, A., 2011. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Trans. Audio, Speech, Lang. Proc.* 19 (7), 2067–2080.
- [11] Younghak, S., Seungchan, L., Junho, L., Heung-No, L., 2012. Sparse representation-based classification scheme for motor imagery-based brain-computer interface systems *J. Neural Eng.* 9, 056002.
- [12] Yang, M. and Zhang, L., 2010. Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In *ECCV*.
- [13] Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T., and Yan, S., 2010. Sparse representation for computer vision and pattern recognition. *Proceedings of IEEE, Special Issue on Applications of Compressive Sensing & Sparse Representation*, 98(6):1031-1044.
- [14] Schlögl, A., Lee, F., Bischof, H., Pfurtscheller, G., 2005. Characterization of four-class motor imagery EEG data for the BCI-competition. *J. Neural Eng.* 2 (4), L14–L22.
- [15] Kaper, M., Meinicke, P., Grossekhoefer, U., Lingner, T., Ritter, H., 2004. BCI Competition 2003--Data set IIb: support vector machines for the P300 speller paradigm. *IEEE Trans. Biomed. Eng.* 51 (6), 1073-1076.
- [16] Shenoy, P., Krauledat, M., Blankertz, B., Rao, R.P.N., Müller, K.-R., Towards adaptive classification for BCI *J. Neural Eng.*, 3 (2006), pp. R13–R23
- [17] Donoho, D.L., 2006. Compressed sensing. *IEEE Trans. Inf. Theory* 52 (4), 1289–1306.
- [18] Baraniuk, R., 2007. Compressive sensing. *IEEE Signal Process. Mag.* 24 (4), 118–121.
- [19] Sergios, T., Aggelos, P., Konstantinos, K., Dionisis, C., 2010. *Introduction to Pattern Recognition :A Matlab Approach*, Academic Press, 43–57.
- [20] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.-R., 2008. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.* 25 (1), 41–56.
- [21] Candès, E., Romberg, J., Tao, T., 2006. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* 59 (8), 1207–1223.
- [22] Wasserman, L., 2010. *All of Statistics: A Concise Course in Statistical Inference*, Springer, 63–64.
- [23] MathWorks: <http://www.mathworks.co.kr/kr/help/stats/support-vector-machines-svm.html>
- [24] Yang, A.Y., Zhou, Z., Ganesh, A., Sastry, S.S., Ma, Y., 2013. Fast 11-minimization algorithms for robust face recognition. *IEEE Trans. Image process.* 22 (8), 3234–3246.
- [25] Hsu, C.-W., Chang, C.-C., and Lin, C.-J. 2003. A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University.
- [26] Kübler, A., Neumann, N., Wilhelm, B., Hinterberger, T., Birbaumer, N., 2004. Predictability of brain-computer communication. *J Psychophysiol.* 18:121–129.
- [27] Oskoei, M.A., Gan, J.Q., Huosheng, H., 2009. Adaptive schemes applied to online SVM for BCI data classification, *Inter. Conf. IEEE on Eng. in Medicine and Biology Society (EMBC)*, 2600-2603
- [28] Yu, H., Lu, H., Ouyang, T., Liu, H., and Lu, B. L., 2010. Vigilance detection based on sparse representation of EEG *Proc. 32nd Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2439–42
- [29] Blanchard, G., and Blankertz, B., 2004. Bci competition 2003-data set I: spatial patterns of self-controlled brain rhythm modulations. *IEEE Transactions on Biomedical Engineering*, 51(6):1062-1066.
- [30] Aharon, M., Elad, M., and Bruckstein, A., 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54 (11), 4311-4322.
- [31] Yang, M., Zhang, L., Feng, X., and Zhang, D., 2011. Fisher discrimination dictionary learning for sparse representation. *IEEE International Conference on Computer Vision*, 543-550.
- [32] Sanei, S., Chambers, J. A., 2007. *EEG Signal Processing* (John Wiley & Sons Inc.)
- [33] Morioka, H., Kanemura, A., Hirayama, J.-I., Shikauchi, M., Ogawa, T., Ikeda, S., Kawanabe, M., and Ishii, S., 2015. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111, 167-178.
- [34] Sellers, E. W., Kubler, A., Donchin, E., 2006. Brain-computer interface research at the University of South Florida Cognitive Psychophysiology Laboratory: the P300 Speller *IEEE Trans Neural Syst Rehabil Eng.* 14(2), 221–224.
- [35] Yang, L., Leung, H., Peterson, D. A., Sejnowski, T. J., Poizner, H., 2014. Toward a Semi-Self-Paced EEG Brain Computer Interface: Decoding Initiation State from Non-Initiation State in Dedicated Time Slots. *PLoS ONE* 9(2): e88915.
- [36] Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Nikulin, V., Müller, K. R., 2008. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. Platt, J., Koller, D., Singer, Y., Roweis, S., (Eds.), *Advances in Neural Information Processing Systems 20*, MIT Press, Cambridge, MA, 113–120.
- [37] Hohyun, C., Minkyu, A., Santae, A. and Sung, C., J., 2013. Strategy for Reducing Calibration Time With Invariant Common Spatio-Spectral Patterns. *Proceedings of the fifth international brain-computer interface meeting*, ID 123, 2013.
- [38] Samek, W., Vidaurre, C., Müller, K. R. and Kawanabe, M., 2012. Stationary common spatial patterns for brain-computer interfacing. *J. Neural Eng.*, 9(2), 026013.
- [39] Arvaneh, M., Guan, C., Ang, K., K. and Quek, C., 2013. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Trans. Networks and Learning Sys*, 24(4), 610-619.
- [40] Kawanabe, M., Samek, W., Müller, K. R. and Vidaurre, C., 2014. Robust common spatial filters with a maxmin approach. *Neural Comput.*, 26(2), 1-28.
- [41] Vidaurre, C., Kawanabe, M., von Büna, P., Blankertz, B. and Müller, K. R., 2011. Toward unsupervised adaptation of LDA for brain-computer interfaces *IEEE Trans. Biomed.Eng.* 58, 587–597.
- [42] Billinger, M., Brunner, C. and Müller-Putz G., R., 2013. Single-trial connectivity estimation for classification of motor imagery data. *J. Neural Eng.*, 10(4), 046006.

Holistic random encoding for imaging through multimode fibers

Hwanchol Jang,¹ Changhyeong Yoon,² Euiheon Chung,³ Wonshik Choi,²
and Heung-No Lee^{1,*}

¹*School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea*

²*Department of Physics, Korea University, Seoul 136-701, South Korea*

³*Department of Medical System Engineering and School of Mechatronics, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea*

*heungno@gist.ac.kr

Abstract: The input numerical aperture (NA) of multimode fiber (MMF) can be effectively increased by placing turbid media at the input end of the MMF. This provides the potential for high-resolution imaging through the MMF. While the input NA is increased, the number of propagation modes in the MMF and hence the output NA remains the same. This makes the image reconstruction process underdetermined and may limit the quality of the image reconstruction. In this paper, we aim to improve the signal to noise ratio (SNR) of the image reconstruction in imaging through MMF. We notice that turbid media placed in the input of the MMF transforms the incoming waves into a better format for information transmission and information extraction. We call this transformation as holistic random (HR) encoding of turbid media. By exploiting the HR encoding, we make a considerable improvement on the SNR of the image reconstruction. For efficient utilization of the HR encoding, we employ sparse representation (SR), a relatively new signal reconstruction framework when it is provided with a HR encoded signal. This study shows for the first time to our knowledge the benefit of utilizing the HR encoding of turbid media for recovery in the optically underdetermined systems where the output NA of it is smaller than the input NA for imaging through MMF.

©2015 Optical Society of America

OCIS codes: (110.0113) Imaging through turbid media; (100.3190) Inverse problems; (060.2350) Fiber optics imaging.

References and links

1. Y. Choi, C. Yoon, M. Kim, T. D. Yang, C. Fang-Yen, R. R. Dasari, K. J. Lee, and W. Choi, "Scanner-free and wide-field endoscopic imaging by using a single multimode optical fiber," *Phys. Rev. Lett.* **109**(20), 203901 (2012).
2. T. Čižmár and K. Dholakia, "Exploiting multimode waveguides for pure fibre-based imaging," *Nat. Commun.* **3**, 1027 (2012).
3. I. N. Papadopoulos, S. Farahi, C. Moser, and D. Psaltis, "Focusing and scanning light through a multimode optical fiber using digital phase conjugation," *Opt. Express* **20**(10), 10583–10590 (2012).
4. C. Yoon, Y. Choi, M. Kim, J. Moon, D. Kim, and W. Choi, "Experimental measurement of the number of modes for a multimode optical fiber," *Opt. Lett.* **37**(21), 4558–4560 (2012).
5. S. Bianchi, V. P. Rajamanickam, L. Ferrara, E. Di Fabrizio, C. Liberale, and R. Di Leonardo, "Focusing and imaging with increased numerical apertures through multimode fibers with micro-fabricated optics," *Opt. Lett.* **38**(23), 4935–4938 (2013).
6. I. N. Papadopoulos, S. Farahi, C. Moser, and D. Psaltis, "Increasing the imaging capabilities of multimode fibers by exploiting the properties of highly scattering media," *Opt. Lett.* **38**(15), 2776–2778 (2013).
7. Y. Choi, C. Yoon, M. Kim, J. Yang, and W. Choi, "Disorder-mediated enhancement of fiber numerical aperture," *Opt. Lett.* **38**(13), 2253–2255 (2013).
8. Y. Choi, T. D. Yang, C. Fang-Yen, P. Kang, K. J. Lee, R. R. Dasari, M. S. Feld, and W. Choi, "Overcoming the diffraction limit using multiple light scattering in a highly disordered medium," *Phys. Rev. Lett.* **107**(2), 023902 (2011).
9. I. M. Vellekoop, A. Lagendijk, and A. P. Mosk, "Exploiting disorder for perfect focusing," *Nat. Photonics* **4**(5), 320–322 (2010).

10. A. P. Mosk, A. Lagendijk, G. Lerosey, and M. Fink, "Controlling waves in space and time for imaging and focusing in complex media," *Nat. Photonics* **6**(5), 283–292 (2012).
11. H. Jang, C. Yoon, E. Chung, W. Choi, and H.-N. Lee, "Speckle suppression via sparse representation for wide-field imaging through turbid media," *Opt. Express* **22**(13), 16619–16628 (2014).
12. A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron, "Imaging with nature: compressive imaging using a multiply scattering medium," *Sci. Rep.* **4**, 5552 (2014).
13. A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.* **51**(1), 34–81 (2009).
14. D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization," *Proc. Natl. Acad. Sci. U.S.A.* **100**(5), 2197–2202 (2003).
15. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.* **54**, 4311–4322 (2006).
16. E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Trans. Inf. Theory* **52**(12), 5406–5425 (2006).
17. E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.* **31**(1), 59–73 (2011).
18. H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical Foundations and Numerical Methods for Sparse Recovery*, ser. Radon Series on Computational and Applied Mathematics, H. Rauhut, Ed. Berlin, Germany: De Gruyter, **9**, 1–94 (2010).
19. E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.* **59**(8), 1207–1223 (2006).
20. J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory* **50**(10), 2231–2242 (2004).
21. J. Yang and Y. Zhang, "Alternating direction algorithms for L_1 -problems in compressive sensing," *SIAM J. Sci. Comput.* **33**(1), 250–278 (2011).
22. G. Strang, in *Linear Algebra and Its Applications*, 4ed, (Thomson Brooks/Cole, 2006)
23. J. W. Goodman, "Some fundamental properties of speckle," *J. Opt. Soc. Am.* **66**(11), 1145–1150 (1976).
24. A. Goetschy and A. D. Stone, "Filtering random matrices: the effect of incomplete channel control in multiple scattering," *Phys. Rev. Lett.* **111**(6), 063901 (2013).
25. S. M. Popoff, G. Lerosey, R. Carminati, M. Fink, A. C. Boccara, and S. Gigan, "Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media," *Phys. Rev. Lett.* **104**(10), 100601 (2010).
26. I. Freund, M. Rosenbluh, and S. Feng, "Memory effects in propagation of optical waves through disordered media," *Phys. Rev. Lett.* **61**(20), 2328–2331 (1988).

1. Introduction

Multimode fibers (MMF) support *multimode* propagation of light such that the light travels not only along the cylindrical axis of the core, *single* mode if it does, but also along multiple different paths with non-zero wave vector components in traverse directions.

The use of MMF for imaging has drawn great interests recently [1–7]. Current endoscopic imaging systems in clinics are based on bundles of fibers where each fiber transfers the signal corresponding to a single pixel of the final image. The multiple propagation characteristic of MMF allows a complex image to be transferred through not a bundle of fibers but only with a single fiber. This enables the miniaturization of imaging systems. Thus, MMF is expected to become a significantly important part for minimally invasive endoscopic imaging where a fiber with needle-like dimensions can transfer complex images. However, there is an intrinsic limitation in imaging through MMF on the spatial image resolution which is imposed by the low numerical aperture (NA) of available MMF [4–7]; the typical NA of MMFs with large number of propagation modes range from 0.2 to 0.5 [5] whereas the NA of optical lenses reaches up to 0.95 in air and higher than 1.6 with special oil immersions.

It has been demonstrated that the problem of the low resolution given by the low NA of MMFs can be relaxed by the use of turbid media in conjunction with MMFs [6,7]. Wave propagation through turbid media, such as white paint, ground glass, and biological tissue, produces complex speckle patterns in the image plane due to multiple scattering of waves in the media. Multiple scattering of waves, referring to the phenomenon where the light waves are forced to deviate from a straight trajectory due to refractive index inhomogeneity through which they pass. It is obvious that this multiple scattering process hinders accurate transferring of images through turbid media. Recently, interesting results were reported that the multiple scattering process in turbid media can be used for overcoming the resolution limit determined by the NA of the optical systems [8,9]. The NA of an optical system sets the

maximum incident angle θ_{\max} of the incoming waves that can be accepted by the system. In wide-field imaging, the multiple scattering through turbid media changes the directions of input waves and some of the waves with large incident angles beyond the acceptance angle of a usual lens can be redirected to the detector [8]. Thus, the effective input NA of the lens becomes increased. In point scanning imaging, multiple scattering is combined with wave-front shaping and makes it possible to focus the light beam to a point smaller than the diffraction limit given by the NA [9]. These resolution improvements by turbid media can be made with the same principle in imaging through MMF when it is used with turbid media. It was found that waves with larger incident angles than the acceptance angle of the MMF can be transferred in the wide-field imaging [7], and a smaller focusing point can be made in point scanning imaging [6].

Here, we show that the redirection of waves is not the only useful characteristic of turbid media. Multiple scattering in turbid media scrambles the waves in a seemingly randomized, a deterministic but complex, manner and this brings forth other positive effects as well. Random scrambling of different modes of waves converts the object wave into a speckle pattern. The object cannot be directly observed looking at the speckle pattern with bare eyes, and thus some form of an inverse operation is necessary such as descrambling for imaging or wave-front shaping for focusing [8–12]. Thus, random scrambling gives an impression that it is only a hindrance; on the contrary, this traditional perspective can be challenged and improved.

Waves scattered from an object are made of the superposition of multiple different modes. Those modes travel through the turbid medium and become scrambled. We note that each mode of the scrambled waves in fact contains *holistic* information of the object. A signal mode is *holistic* in the sense that the mode contains the information of the whole modes of the incoming waves. Each mode of incoming waves propagates through the turbid medium is scattered and redirected into many different modes at the output of the turbid media. One mode is scattered into almost all the modes of the output waves. This is to mean that a single output mode is made out of the superposition of many if not all incoming waves. In addition, each mode of the incoming waves went through a propagation path completely different from other modes. Thus, each output mode offers an independent view of the same object. In this sense, the random scrambling can be viewed as a beneficial encoding process, which provides multiple independent outlooks of the whole waves from an object. We will refer to this signal transforming process in turbid media as *holistic random* (HR) encoding.

In this paper, we aim to make an efficient use of the HR encoding of the multiple scattering in turbid media and improve the image quality in wide-field imaging through MMF. We consider an imaging system through MMF where a turbid medium is placed at the input end of the MMF. In this case, the imaging system becomes underdetermined because the output NA of the MMF is smaller than the input NA of it. The information of object waves is transferred to a less number of wave modes than that of the input wave modes. Considering that the degree of freedom of a signal is reduced, it is not easy to transmit information of the object waves without loss. To compound the matter, the recovery of the object waves is not easy as well since the dimension (the number of elements) of the observed signal is smaller than that of the original signal to be estimated. Here, we show that the HR encoding of the turbid media enables much improved information transmission and signal reconstruction. We employ sparse representation (SR) framework [13–21] and show that the object information can be extracted at an improved fidelity when the signal is HR encoded. In many literatures, SR has been shown to provide superb estimation of the original signal from a smaller number of measurements than the dimension of the original signal [13–21]. Previously, SR was shown to be beneficial in imaging through turbid media [11,12]. In [11], SR was shown to suppress speckles in reconstructed images. It was shown in [12] that SR recovers the image well in the situation where the number of pixels in the CCD array is smaller than that the pixels of the original image. This paper provides the first result that SR can be used to improve the image reconstruction quality in imaging through MMF when the

imaging system is underdetermined due to the use of turbid media at the input end of the MMF.

This paper is organized as follows: Section 2 describes the system for imaging through MMF, and Section 3 investigates the HR encoding by the multiple scattering in turbid media and links it to the SR framework for image reconstruction. Experimental results are discussed in Section 4, and Section 5 concludes the paper.

2. System description

2.1 Experimental set-up

We consider an imaging system where the object wave propagates through a single MMF with 1 m of length (NA of 0.22; Thorlabs, M14L01) and recorded at the CCD array. The input facet of the MMF is randomly coated by ZnO nanoparticles.

The experimental schematic of the imaging is depicted in Fig. 1. Figure 1(a) describes the calibration stage where the transmission matrix (TM) of the coated MMF is measured. TM is a collection of responses of the coated MMF to a set of incoming plane waves ($\lambda = 633 \text{ nm}$) with N different incident angles to the input facet of the coated MMF. We used $N = 4000$ different angles for our TM. For preparation of the plane waves, no object is presented in the object plane. The incident angle of the plane wave is controlled by a galvanometer.

The transferring of an object wave is described in [Fig. 1(b)]. Once the TM is measured and becomes available, the response of the coated MMF to the object wave is measured at the CCD. The object wave is distorted in the coated MMF due to the multiple scattering in the turbid medium of ZnO nanoparticles and the interference among waves with different propagation modes inside the MMF. For the object, we use a sample similar to USAF target.

Now with the TM and the distorted object image, the object wave is recovered by computation. Here, all the measurements are post processed by the off-axis holography [8] to obtain the E-field images; the baseline method considered in this paper follows the turbid lens imaging (TLI) system in [8]. We use MMF with the NA of 0.22 in the experiment. We fix the NA of the sub-systems followed by the MMF slightly larger, 0.24, than 0.22. By doing this, we can capture the most of the signal from the MMF even though there are some experimental mismatches, for example, the error in the alignment on the optical axis.

2.2 Image recovery using transmission matrices

The object wave $U_o(x, y)$ is decomposed into a set of plane waves with different propagation directions as follows

$$U_o(x, y) = \sum_{\mathbf{k}} A_o(\mathbf{k})P(x, y; \mathbf{k}) \quad (1)$$

where $\mathbf{k} := k_x \mathbf{i}_x + k_y \mathbf{i}_y + k_z \mathbf{i}_z$ is the wave vector ($k_x / 2\pi = \sin \theta_x / \lambda$, $k_y / 2\pi = \sin \theta_y / \lambda$, $k_z / 2\pi = \sin \theta_z / \lambda$, and $(\theta_x, \theta_y, \theta_z)$ is the angle of propagation), \mathbf{i}_x , \mathbf{i}_y , and \mathbf{i}_z are unit vectors in x , y , and z directions (the optical axis is in the z direction), respectively, $P(x, y; \mathbf{k})$ is the plane wave with the propagation direction \mathbf{k} , and $A_o(\mathbf{k})$ is the angular spectrum of the object wave. The distorted object wave $U_r(x, y)$ after propagation through the coated MMF is expressed as

$$U_r(x, y) = \sum_{\mathbf{k}} A_o(\mathbf{k})F(x, y; \mathbf{k}) \quad (2)$$

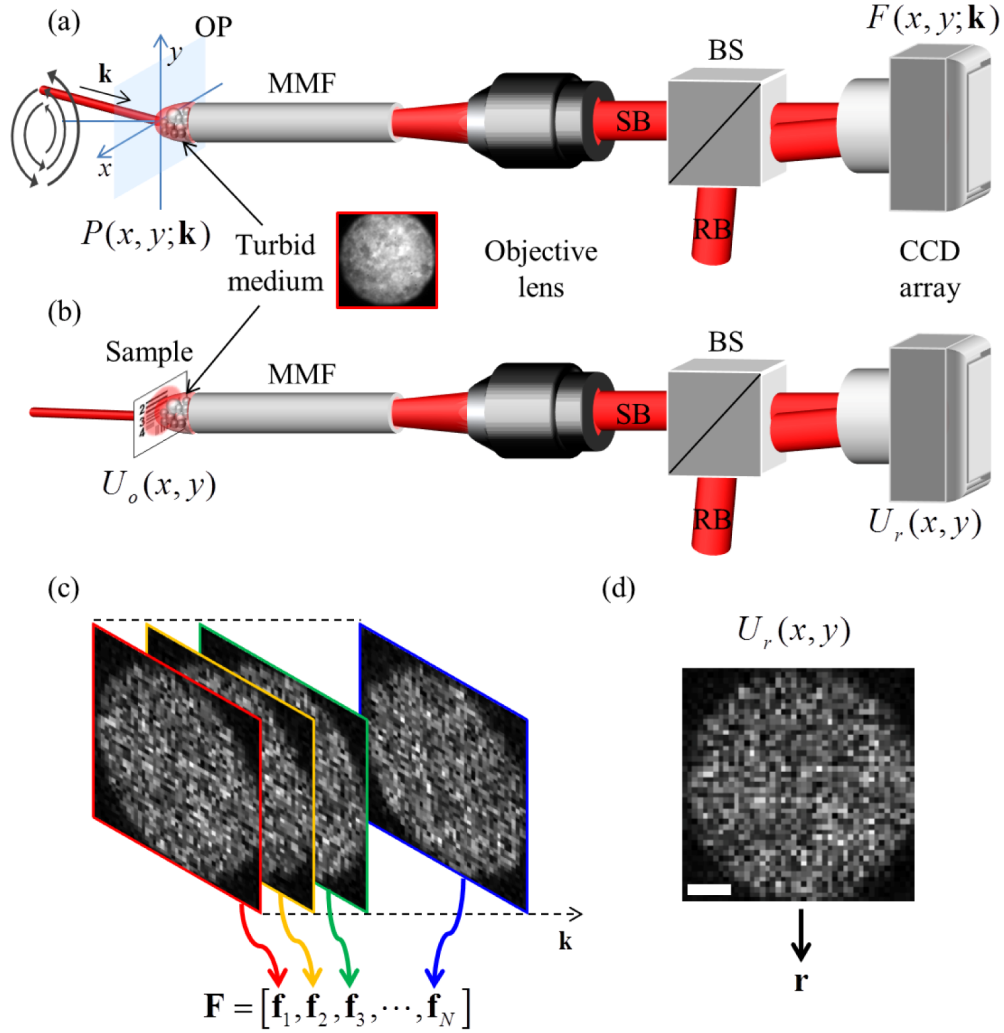


Fig. 1. Experimental schematics of imaging through turbid media coated MMF. (a) Recording of the TM. OP: object plane, MMF: multimode fiber, BS: beam splitter, SB: sample beam, RB: reference beam. The photo in the red box shows the image of the input surface of the turbid media coated MMF. (b) Recording of the distorted object wave. (c) The recorded TM. Only the intensities are shown in the image but the phases are estimated as well; thus the columns are complex valued vectors. (d) The recorded distorted object wave. Again, only the intensity is shown here. Scale bar: $10 \mu\text{m}$.

where $F(x, y; \mathbf{k})$ is the response wave of the coated MMF for a single plane wave $P(x, y; \mathbf{k})$. Now using the vector notations, the distorted object wave can be expressed as follows

$$\mathbf{r} = \mathbf{F}\mathbf{a} \quad (3)$$

where $\mathbf{r} \in \mathbb{C}^M$ and $\mathbf{a} \in \mathbb{C}^N$ are the vectorized versions of $U_r(x, y)$ and $A_o(\mathbf{k})$, and $\mathbf{F} \in \mathbb{C}^{M \times N}$ is the TM each column of which is the vectorized version of the $F(x, y; \mathbf{k})$; there are N different propagation directions (modes) \mathbf{k} considered.

From the distorted object wave, the angular spectrum of the object wave is estimated in [7] by using the pseudo inversion (PINV) method,

$$\hat{\mathbf{a}}_{\text{pinv}} = \mathbf{F}^{-1} \mathbf{r} \quad (4)$$

where \mathbf{F}^{-1} is the PINV matrix of \mathbf{F} . Using the estimated angular spectrum, the object wave in the object plane can be reconstructed.

3. Imaging through MMF in conjunction with turbid media

Recall that the input facet of the MMF is coated by a turbid medium. Due to the multiple scattering process through the turbid medium, some of those waves whose incident angles to the medium are larger than the acceptance angle determined by the NA of the MMF, $\theta_z > \theta_{\text{max}}$, are redirected, $\theta_z \rightarrow \theta'_z$, and are coupled to the MMF, $\theta'_z \leq \theta_{\text{max}}$. This introduces more modes of the object (the green arrows in [Fig. 2(b)]) to the detector. The NA of a MMF is proportional to the square root of the number of modes captured by the MMF. As a result, it was shown in experiments [7] that the effective NA of the input side of the MMF increases.

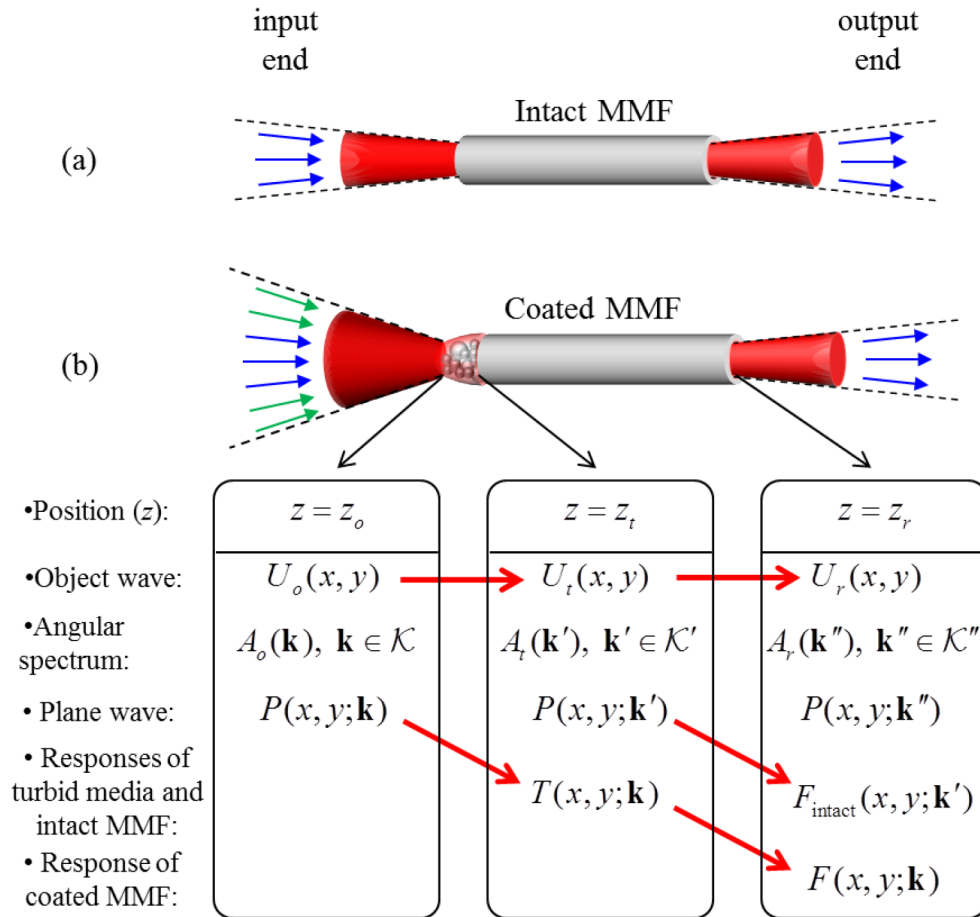


Fig. 2. Acceptance angle, exit angle, and their corresponding modes of waves for (a) intact MMF and (b) coated MMF. In (b), more modes of input waves (green arrows) can be captured in coated MMF. But, the number of modes of output waves is the same as that of intact MMF.

However, we notice that the increase in the effective NA causes the image capturing system to be underdetermined where the input NA is larger than the output NA (Fig. 2). Larger number of modes of the object waves than the available propagation modes in the MMF is captured. This gives a challenge in the recovery of the object. In general, the output signal of an underdetermined system does not convey all the information of the input signal.

Besides, even in the cases where there are no information losses, it is also well known from the linear algebra [22] that the correct estimation of a signal in the underdetermined systems is not easy for it has more than one solution. Note that the use of another turbid medium in the output facet does not change the situation because the number of propagation modes inside the MMF is still the same.

3.1 HR encoding

Multiple scattering in turbid media provides HR encoding to the incoming wave. HR encoding provides the two kinds of encoding, holistic encoding and random encoding. We define “holistic encoding” to be a signal transforming process which makes each component of the transformed signal to contain all the components of the original signal with a certain format, especially in this paper, a weighted summation of them. We also define “random encoding” to be a multiplexing process where the contributions of a component of the original signal to components of the output signal are modeled by independent random variables.

Now, we show that the waves $U_o(x, y)$ scattered from an object is HR encoded in the turbid medium. 1) Holistic encoding: Each input mode $P(x, y; \mathbf{k})$ of the incoming waves $U_o(x, y)$ experiences multiple scattering in the turbid medium. The object wave $U_i(x, y)$ after propagation through the turbid medium (refer to [Fig. 2(b)]) is expressed as

$$U_i(x, y) = \sum_{\mathbf{k} \in \mathcal{K}} A_o(\mathbf{k}) T(x, y; \mathbf{k}) \quad (5)$$

where \mathcal{K} is the set of modes \mathbf{k} which propagate through the coated MMF and reach to the detector and $T(x, y; \mathbf{k})$ is the response wave of the turbid medium to $P(x, y; \mathbf{k})$. The response wave $T(x, y; \mathbf{k})$ at the output of the turbid medium for a single plane wave $P(x, y; \mathbf{k})$ has many modes of waves. This is because the multiple scattering process in the turbid medium changes the directions of the waves in a randomized manner. $T(x, y; \mathbf{k})$ is expressed as

$$T(x, y; \mathbf{k}) = \sum_{\mathbf{k}' \in \mathcal{K}'} t(\mathbf{k}'; \mathbf{k}) P(x, y; \mathbf{k}') \quad (6)$$

where \mathbf{k}' is the wave vector after propagation through the turbid medium, \mathcal{K}' is the set of modes \mathbf{k}' which propagate through the MMF and reach to the detector, $t(\mathbf{k}'; \mathbf{k}) \in \mathbb{C}$ is the contribution of a mode \mathbf{k} of input waves to a mode \mathbf{k}' of output waves, and $P(x, y; \mathbf{k}')$ is the plane wave with the propagation direction \mathbf{k}' .

Here, $t(\mathbf{k}'; \mathbf{k})$ is well approximated by independent and identically distributed (i.i.d.) complex valued Gaussian random variable. It was found in [23,24] that the output waves of a turbid medium at a spatial plane (x', y') when a mode of waves is transmitted through the medium, $t(x', y'; \mathbf{k})$, are i.i.d. complex valued Gaussian random variables provided that the number of independent scatters is large; $t(x', y'; \mathbf{k})$ and $t(\mathbf{k}'; \mathbf{k})$ are a two-dimensional Fourier transform pair. This has been also supported in the experiments [12,25]. The distribution of the eigenvalues of the TM composed of $t(x', y'; \mathbf{k})$ was close to that of i.i.d. Gaussians [25]. The coherence (μ_0 in Sec 3.2) of the TM of $t(x', y'; \mathbf{k})$ behaved similarly to that of i.i.d. Gaussians [12]. We know that the Fourier transform of a Gaussian random matrix is another Gaussian random matrix. Thus, the contribution $t(\mathbf{k}'; \mathbf{k})$ follows i.i.d. complex valued Gaussian, too.

Using [Eq. (6)], the object wave $U_i(x, y)$ in [Eq. (5)] can be expressed as

$$U_i(x, y) = \sum_{\mathbf{k}' \in \mathcal{K}'} A_i(\mathbf{k}') P(x, y; \mathbf{k}') \quad (7)$$

where the angular spectrum $A_t(\mathbf{k}')$ of $P(x, y; \mathbf{k}')$ is

$$A_t(\mathbf{k}') = \sum_{\mathbf{k} \in \mathcal{K}} t(\mathbf{k}'; \mathbf{k}) A_o(\mathbf{k}). \quad (8)$$

Here, the probability that $t(\mathbf{k}'; \mathbf{k}) = 0$ is very small; the probability approaches zero for the event that a realization of complex Gaussian random variable equals to 0. Thus, the angular spectrum $A_t(\mathbf{k}')$ of a mode \mathbf{k}' of the output waves is now the combination of the angular spectrum $A_o(\mathbf{k})$ of all the modes \mathbf{k} . Therefore, a mode of the output waves contains holistic information of the input waves.

2) Random encoding: As it was discussed in the previous paragraph, $t(\mathbf{k}'; \mathbf{k})$ for $\mathbf{k} \in \mathcal{K}$ are approximated by independent random variables. Thus, the angular spectrum $A_t(\mathbf{k}')$ of each mode \mathbf{k}' of the output waves shows an independent view of the same object $U_o(x, y)$.

This HR encoding has a couple of desirable aspects to send information of object waves in underdetermined systems. i) The information of all the input modes is transferred no matter how few modes are in the output waves, if it is more than one. It is because each output mode captures information about all the input modes of the object wave. ii) It sends information of the object waves in an efficient manner. The information of a mode is not redundant to that of other modes as each output mode captures unique information about the object wave. Due to these two aspects, the HR encoded signal is expected to be appropriate for information transmission in underdetermined systems. In the literature, it is shown that HR encoded signals can transmit enough information for a certain kind of signals in underdetermined systems [16–18].

Now, let us see the effect of HR encoding on the object wave after propagation through the MMF. The HR encoded waves $U_t(x, y)$ at the output of the turbid medium are propagated through the MMF. The object wave $U_r(x, y)$ after propagation through the MMF ([Fig. 2(b)]) is expressed as

$$U_r(x, y) = \sum_{\mathbf{k}' \in \mathcal{K}'} A_t(\mathbf{k}') F_{\text{intact}}(x, y; \mathbf{k}') \quad (9)$$

where $F_{\text{intact}}(x, y; \mathbf{k}')$ is the response wave of the intact MMF to $P(x, y; \mathbf{k}')$. $F_{\text{intact}}(x, y; \mathbf{k}')$ usually has more than one propagation modes of waves in MMF. This is because that it is almost infeasible to match a mode of incoming waves to only a single propagation mode in MMF. $F_{\text{intact}}(x, y; \mathbf{k}')$ is expressed as

$$F_{\text{intact}}(x, y; \mathbf{k}') = \sum_{\mathbf{k}'' \in \mathcal{K}''} f_{\text{intact}}(\mathbf{k}''; \mathbf{k}') P(x, y; \mathbf{k}'') \quad (10)$$

where \mathbf{k}'' is the wave vector after propagation through the MMF, \mathcal{K}'' is the set of modes \mathbf{k}'' which are captured at the detector, $f_{\text{intact}}(\mathbf{k}''; \mathbf{k}')$ is the contribution of a mode \mathbf{k}' of input waves to a mode \mathbf{k}'' of output waves, and $P(x, y; \mathbf{k}'')$ is the plane wave with the propagation direction \mathbf{k}'' . Using [Eq. (8)], [Eq. (9)], and [Eq. (10)], the object wave $U_r(x, y)$ can be rewritten as

$$U_r(x, y) = \sum_{\mathbf{k}'' \in \mathcal{K}''} A_r(\mathbf{k}'') P(x, y; \mathbf{k}'') \quad (11)$$

where the angular spectrum $A_r(\mathbf{k}'')$ is

$$A_r(\mathbf{k}'') = \sum_{\mathbf{k} \in \mathcal{K}} C(\mathbf{k}''; \mathbf{k}) A_o(\mathbf{k}), \quad (12)$$

and the weight $C(\mathbf{k}''; \mathbf{k})$ is

$$C(\mathbf{k}''; \mathbf{k}) = \sum_{\mathbf{k}' \in \mathcal{K}'} t(\mathbf{k}'; \mathbf{k}) f_{\text{intact}}(\mathbf{k}'', \mathbf{k}'). \quad (13)$$

Please recall that a larger number of modes \mathbf{k} of the incoming waves than that of propagation modes \mathbf{k}'' in the intact MMF (equivalently \mathbf{k}') are captured with the help of the turbid medium in the coated MMF, $|\mathcal{K}| > |\mathcal{K}''|$. We see that the angular spectrum $A_o(\mathbf{k})$ for $\mathbf{k} \in \mathcal{K}$ of the object waves $U_o(x, y)$ are transferred through waves $P(x, y; \mathbf{k}'')$ with a smaller number of propagations modes $\mathbf{k}'' \in \mathcal{K}''$ in [Eq. (12)]. Thus, the reconstruction of the $A_o(\mathbf{k})$ (or equivalently $U_o(x, y)$) from the $A_r(\mathbf{k}'')$ (or $U_r(x, y)$) is an underdetermined problem.

We see that the signal is still holistically encoded in the output waves $U_r(x, y)$ in [Eq. (12)]. Every component of the angular spectrum $A_o(\mathbf{k})$ of the input waves is contained in each single components of the angular spectrum $A_r(\mathbf{k}'')$ of the output waves. The random encoding property of $U_i(x, y)$ is not inherited to $U_r(x, y)$ since $C(\mathbf{k}'', \mathbf{k})$ is not independent for different \mathbf{k}'' . But, we will see that the correlation between $C(\mathbf{k}'', \mathbf{k})$ with different \mathbf{k}'' is small in [Fig. 4(c)]. This is desirable for image recovery (Sec 3.2). A further explanation on $C(\mathbf{k}'', \mathbf{k})$ is given in Sec 3.3. Please note that we use $P(x, y; \mathbf{k}'')$ for the modes of the output waves of the MMF and for the propagation modes in the MMF interchangeably without distinction as they are one-to-one mapped.

3.2 Sparse representation

SR is a signal representation framework, which has received great interests since it can be used to estimate a signal even in the underdetermined systems [13–21]. HR encoding increases the oversampling ratio of the underdetermined systems in which a signal is estimated correctly [16–18]. It is well known that there are two conditions for successful application of the SR framework for recovery of original signal from its HR encoded one. First, the signal \mathbf{a} is compressible. A compressible signal means that the signal \mathbf{a} is well approximated with a small number of nonzero elements in \mathbf{a} , say K where $K \ll N$. The object signals of interest in this paper are natural signals and we have many research results showing that they are compressible. It is well known that most natural images are well approximated with only a few elements in the Wavelet domain [15]. Not only with the Wavelet domain, if the signal is represented with a few elements in any other orthogonal signal bases, the signal is compressible [16]. Second, the measurement matrix \mathbf{F} needs to be incoherent. We say a matrix is incoherent if the cross-correlations of columns of the matrix are small. This follows the conventional meaning for incoherence of a matrix in [14,17,20]. Note that this incoherence is different from that in optics which is typically a phase relationship among waves. The incoherence of a measurement matrix can be measured in its Gram matrix $\mathbf{D} = \mathbf{F}^* \mathbf{F}$. We assume that the norm of each column of \mathbf{F} is normalized to be one. The amplitude $|d_{ij}|$ of the off-diagonal elements of \mathbf{D} indicates the cross-correlation of different i^{th} and j^{th} columns of the measurement matrix where d_{ij} is the $(i, j)^{\text{th}}$ element of \mathbf{D} and $|\cdot|$ denotes the absolute value of the complex number.

Several different measures are used for incoherence of a measurement matrix. The simplest one is i) the largest off-diagonal element in the Gram matrix, $\mu_0 \triangleq \max_{i,j} |o_{ij}|$ where o_{ij} is the $(i, j)^{\text{th}}$ element of $\mathbf{O} = \mathbf{D} - \mathbf{I}$ [14,17,20]. But, this does not characterize the incoherence of a measurement matrix well for it only considers the most extreme case [20]. The other two measures are ii) the size of the smallest group of off-diagonal elements in a single row of the

Gram matrix which have the sum greater than one, $\mu_1 \triangleq \min_i \left(\min_{|J|} \left| \sum_{j \in J} |o_{ij}| \right| \geq 1 \right)$ where $|J|$ is the cardinality of an index set J [14], and iii) the maximum value of the summation of M off-diagonal elements in a single row of the Gram matrix, $\mu_2 \triangleq \max_{|J|=M} \max_{i \in J} \sum_{j \in J} |o_{ij}|$ [20]. Those two measures provide somewhat more of the general behavior of a measurement matrix. However, they still do not provide an overall behavior of the measurement matrix as they also reflect extreme cases by taking the minimum size or the maximum sum; actually, the two measures are designed to provide theoretical bounds of M which guarantee the successful estimation of sparse signals, which are signals with small numbers of nonzero elements, in the SR framework.

In this paper, we aim to use $\mu_3 \triangleq \left| \left\{ |o_{ij}| > \alpha \right\} \right| / (N^2 - N)$ the fraction of the off-diagonal elements of a Gram matrix whose absolute values are comparable, $|o_{ij}| > \alpha$ where α is the degree of comparableness ($0 < \alpha < 1$), to the value of the diagonal elements as the measure of incoherence. The rationale for this is i) that it does not take any extreme values, ii) that it is coherent to the previous three measures, μ_0 , μ_1 , and μ_2 , as more number of large off-diagonal elements is likely to lead to values of the three measures which indicate the matrix is less incoherent, and iii) that the use of fractions is appropriate to compare the incoherence of the two different Gram matrices with different size; this is the case we consider in this paper. It is desirable to have smaller μ_3 . Construction of an incoherent measurement matrix in a deterministic manner for an underdetermined system is known to be difficult. Fortunately, in the literature, it is shown analytically or empirically that many kinds of randomly generated matrices are incoherent [16–19]. That is, we can make good measurement matrices by using random generation, without a careful precise matrix design.

Now we aim to explain how SR framework can be utilized to recover the signal successfully in our turbid lens based MMF imaging. The current state-of-the-art SR systems can recover a signal with K nonzero elements, the so-called K -sparse signal, correctly with just $M = O(K \log(N/K))$ number of random measurements [17].

The estimation in SR can be done by finding solution of the following problem [17]

$$\hat{\mathbf{a}}_{\text{SR}} = \arg \min_{\mathbf{a}} \left\| \Psi^* \mathbf{a} \right\|_1 \text{ subject to } \mathbf{r} = \mathbf{F} \mathbf{a}, \quad (14)$$

where Ψ is the sparsifying basis in which the signal \mathbf{a} can be approximated with just a small number of nonzero elements, $(\cdot)^*$ denotes the conjugate transpose of a matrix, and $\|\cdot\|_1$ denotes the L1 norm, that is, the sum of the absolute values of the vector elements.

3.3 Effect of the random scrambling of turbid media on TM

As it was discussed in Sec 3.1, turbid media provide HR encoding to the object waves. In this subsection, we show that HR encoding provides incoherent TM to MMF. We focus on how improved the coated MMF is compared to the intact MMF in terms of incoherence. As we have discussed in Sec 3.2, we will use μ_3 as the measure of incoherence.

TM is a collection of responses $F(x, y; \mathbf{k})$ of the MMF to a set of incoming plane waves $P(x, y; \mathbf{k})$ with N different modes \mathbf{k} (Sec 2.2). Let us consider first the TM of an intact MMF without a turbid medium deposited. For an intact MMF, the response of the MMF $F(x, y; \mathbf{k})$ is $F_{\text{intact}}(x, y; \mathbf{k})$ in [Eq. (10)]; here, \mathbf{k} and \mathbf{k}' are the same. The TM of the MMF consists of $F_{\text{intact}}(x, y; \mathbf{k})$ with different modes \mathbf{k} . As it was told in Section 3.1, $F_{\text{intact}}(x, y; \mathbf{k})$

is a superposition of waves $P(x, y; \mathbf{k}'')$ for more than one propagation modes \mathbf{k}'' in MMF. $F_{\text{intact}}(x, y; \mathbf{k})$ in [Eq. (10)] can be written as

$$F_{\text{intact}}(x, y; \mathbf{k}) = \sum_{\mathbf{k}'' \in S_{\text{intact}}(\mathbf{k})} f_{\text{intact}}(\mathbf{k}''; \mathbf{k}) P(x, y; \mathbf{k}'') \quad (15)$$

where $S_{\text{intact}}(\mathbf{k}) := \{\mathbf{k}'' | f_{\text{intact}}(\mathbf{k}''; \mathbf{k}) \neq 0\}$ is the set of excited propagation modes in the MMF when plane waves $P(x, y; \mathbf{k})$ with the mode \mathbf{k} are inserted; $S_{\text{intact}}(\mathbf{k}) \subset \mathcal{K}''$. We found that a small number of propagation modes in the MMF are excited when waves with a single mode are inserted ([Fig. 3(c)]); this is serious especially when the incident angles θ_z of the incoming waves are small. With a small number of propagation modes $P(x, y; \mathbf{k}'')$, there are not many incoherent $F_{\text{intact}}(x, y; \mathbf{k})$ available because they are the combinations of the few plane waves whose modes \mathbf{k}'' are included in $S_{\text{intact}}(\mathbf{k})$; it is analogous to generating incoherent vectors which are linear combinations of a small number of basis columns. Thus, it is not easy to fill up the columns of the TM with incoherent $F_{\text{intact}}(x, y; \mathbf{k})$ for \mathbf{k} with small incident angles. This results in a not incoherent TM.

We show in [Fig. 3(a)] some of the $F_{\text{intact}}(x, y; \mathbf{k})$ with several \mathbf{k} , respectively. We consider discrete incident angles \mathbf{k}_i for $1 \leq i \leq 2000$. Among the responses $F_{\text{intact}}(x, y; \mathbf{k}_i)$ of all the considered incident angles \mathbf{k}_i , we show \mathbf{k}_i with $i = 1, 101, 201, \text{ and } 1001$. The incident angle changes with the pattern of a spiral starting from the center. A small index i of the response means that it is for that with a small incident angle θ_z . We see that the responses $F_{\text{intact}}(x, y; \mathbf{k}_i)$ of smaller incident angles ($i = 1, 101, 201$) are not complex due to lack of the available propagation modes $P(x, y; \mathbf{k}'')$ ([Fig. 3(c)]). Here, we mean by complex that the value in each pixel (x, y) of the response $F_{\text{intact}}(x, y; \mathbf{k}_i)$ changes enough from those in its neighboring pixels. This can be seen in the autocorrelation of $F_{\text{intact}}(x, y; \mathbf{k}_i)$,

$$R_{FF; \text{intact}}(x, y; \mathbf{k}_i) \triangleq \left| \sum_{x'} \sum_{y'} F_{\text{intact}}(x', y'; \mathbf{k}_i) F_{\text{intact}}^*(x' + x, y' + y; \mathbf{k}_i) \right| \quad \text{in [Fig. 3(b)].}$$

It looks complex around $i = 1001$. But it is also made of plane waves with only several different modes, not all the modes considered. Compared to $R_{FF}(x, y; \mathbf{k}_i)$ in [Fig. 4(b)], $R_{FF; \text{intact}}(x, y; \mathbf{k}_i)$ has more non-ignorable values at those with $x \neq 0$ or $y \neq 0$.

Intact MMF

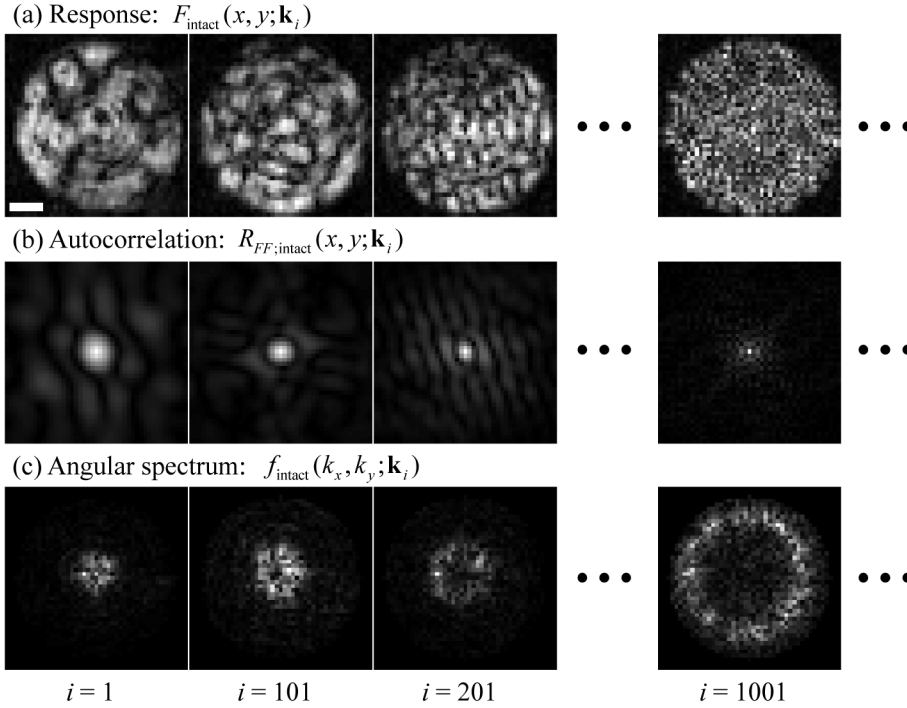


Fig. 3. (a) Recorded responses $F_{\text{intact}}(x, y; \mathbf{k}_i)$, (b) Amplitude of the autocorrelation for the response, $R_{FF;\text{intact}}(x, y; \mathbf{k}_i)$, and (c) Angular spectrum of the response, $f_{\text{intact}}(\mathbf{k}''; \mathbf{k}_i)$, of intact MMF. Among the responses of all the incident angles covering the NA of 0.22, the 1st, the 101st, the 201st, and the 1001st of them are presented. Only the intensities are shown here. Scale bar: 10 μm .

Now consider the TM of the coated MMF with a turbid medium deposited in the input facet. The response $F(x, y; \mathbf{k})$ of the coated MMF to $P(x, y; \mathbf{k})$ can be obtained in two steps. First, the response $T(x, y; \mathbf{k})$ of the turbid medium to $P(x, y; \mathbf{k})$ is obtained. Second, the response $F(x, y; \mathbf{k})$ of the intact MMF to $T(x, y; \mathbf{k})$ is obtained. $T(x, y; \mathbf{k})$ is available in [Eq. (6)]. $F(x, y; \mathbf{k})$ is derived as

$$\begin{aligned}
 F(x, y; \mathbf{k}) &= \sum_{\mathbf{k}' \in \mathcal{K}'}^{(a)} t(\mathbf{k}'; \mathbf{k}) F_{\text{intact}}(x, y; \mathbf{k}') \\
 &= \sum_{\mathbf{k}'' \in \mathcal{K}''} \sum_{\mathbf{k}' \in \mathcal{K}'}^{(b)} t(\mathbf{k}'; \mathbf{k}) f_{\text{intact}}(\mathbf{k}''; \mathbf{k}') P(x, y; \mathbf{k}'') \\
 &= \sum_{\mathbf{k}'' \in \mathcal{S}_{\text{coated}}(\mathbf{k})} C(\mathbf{k}''; \mathbf{k}) P(x, y; \mathbf{k}'')
 \end{aligned} \tag{16}$$

where (a) is from [Eq. (6)] and the fact that $F_{\text{intact}}(x, y; \mathbf{k}')$ is the response wave of the intact MMF to $P(x, y; \mathbf{k}')$, (b) is from [Eq. (10)], and $\mathcal{S}_{\text{coated}}(\mathbf{k}) := \{\mathbf{k}'' | C(\mathbf{k}''; \mathbf{k}) \neq 0\}$ denotes the set of excited propagation modes in the MMF when plane waves $P(x, y; \mathbf{k})$ with the mode \mathbf{k} are inserted. Here, different from that of the intact MMF, the number of excited modes in the coated MMF for $F(x, y; \mathbf{k})$ is not small ([Fig. 4(c)]). We can easily see in [Eq. (13)] that $\mathcal{S}_{\text{coated}}(\mathbf{k}) = \bigcup_{\mathbf{k}'} \mathcal{S}_{\text{intact}}(\mathbf{k}')$, the set of excited modes in the coated MMF is the union of all the

sets of excited modes in the intact MMF. This is because the random scrambling in the turbid medium varies the directions of the waves. This makes the MMF to have incoming waves with a variety of incident angles, and the propagation modes corresponding to those incident angles are all excited. Now, $F(x, y; \mathbf{k})$ are made by combining many $P(x, y; \mathbf{k}'')$ with a variety of \mathbf{k}' . Thus, there are many possible incoherent $F(x, y; \mathbf{k})$ patterns. It is easier to find many incoherent interference patterns out of them. As a result, it becomes easier to compose the TM with many incoherent interference patterns. In [Fig. 4(a)] and [Fig. 4(c)], we see that the responses of the coated MMF to plane waves with the considered incident angles θ_z are complex enough to be speckle patterns.

Turbid media coated MMF

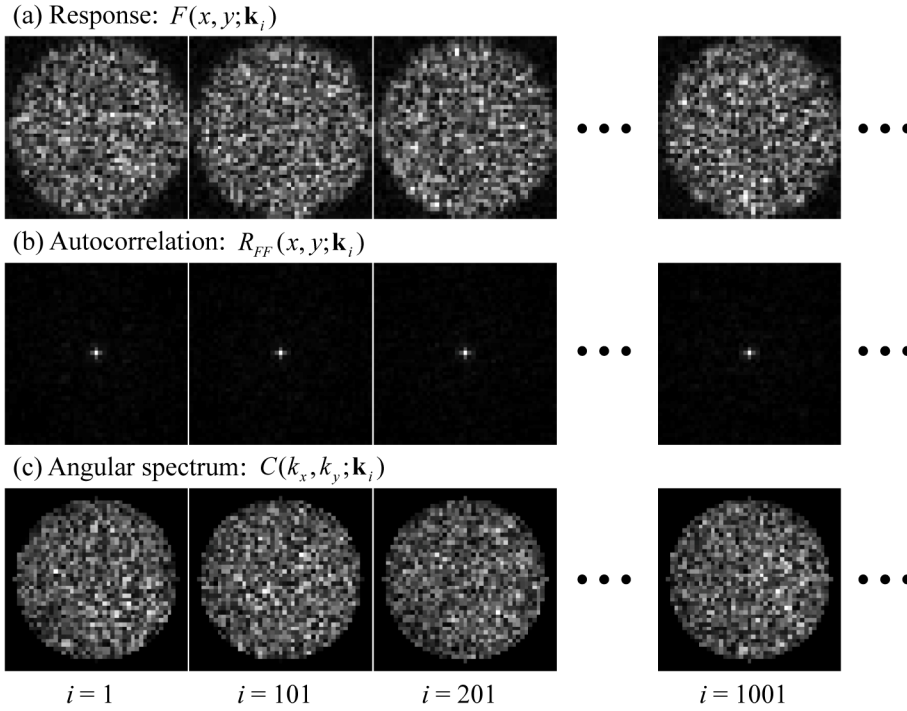


Fig. 4. (a) Recorded responses $F(x, y; \mathbf{k}_i)$, (b) Amplitude of the autocorrelation for the response, $R_{FF}(x, y; \mathbf{k}_i)$, and (c) Angular spectrum of the response, $C(\mathbf{k}''; \mathbf{k}_i)$, of coated MMF. Among the responses of all the incident angles covering the NA of 0.22, the 1st, the 101st, the 201st, and the 1001st of them are presented. Only the intensities are shown here. Scale bar: 10 μm .

Having more propagation modes in the coated MMF surely provides a better situation for the TM to be incoherent. But, this does not always mean that the TM would be incoherent. For an incoherent TM, the way of combining of the propagation modes $C(\mathbf{k}''; \mathbf{k})$ needs to be incoherent for the beams of light with different incident angles \mathbf{k} . It would serve no point if the way of combining the propagation modes was the same for all the incident beams considered, then, as the TM would be completely coherent. We can see in [Eq. (13)] that the way of combining becomes different if $t(\mathbf{k}'; \mathbf{k})$ is different from each other for \mathbf{k} . For two different modes \mathbf{k}_i and \mathbf{k}_j ($\mathbf{k}_i \neq \mathbf{k}_j$), it is found that the contributions $t(\mathbf{k}'; \mathbf{k}_i)$ and $t(\mathbf{k}'; \mathbf{k}_j)$ are uncorrelated if the angle difference of the two modes is not too small,

$\cos^{-1}(\mathbf{k}_i \cdot \mathbf{k}_j) \geq \delta$ for a certain small δ [26]. Please be reminded that a random matrix generation gives an incoherent TM (Sec. 3.2). Now, with the incoherent combinations of the coated MMF, we expect the TM to be more incoherent than that of the intact MMF.

We compare the incoherence μ_3 ($0 \leq \mu_3 \leq 1$) of the TMs of the intact MMF and the coated MMF in Fig. 5. The respective number of rows and that of columns of TM are 2025 and 2000 for the intact MMF and 2025 and 4000 for the coated MMF. We consider several values of α for $1 \geq \alpha \geq 0$. Please note that it is desirable to have smaller μ_3 for superior incoherence. μ_3 of the intact MMF starts to have nonzero value from $\alpha = 0.75$ and becomes larger as α decreases. The start of nonzero value of μ_3 for the coated MMF is at $\alpha = 0.65$ and μ_3 becomes larger as α decreases. For $0.65 \geq \alpha \geq 0.45$, the ratios of the μ_3 of the intact MMF and that of the coated MMF tend to increase (4 at $\alpha = 0.66$ and 6.55 at $\alpha = 0.45$). For $0.4 \geq \alpha \geq 0.1$, the ratios tend to decrease (6.09 at $\alpha = 0.4$ and 4.5 at $\alpha = 0.1$). Regardless of the tendency, the ratios are considerable for $0.65 \geq \alpha \geq 0.1$. At $\alpha = 0.05$, the μ_3 for the coated MMF is larger than that of the intact MMF, and they become the same at $\alpha = 0$. Though the μ_3 for the intact MMF is better at $\alpha = 0.05$, the incoherence here is not meaningful because $\alpha = 0.05$ is not comparable to 1. In all the values of α which are reasonably comparable to 1 ($0.75 \geq \alpha \geq 0.1$), it is found that the coated MMF has superior incoherence than the intact MMF does.

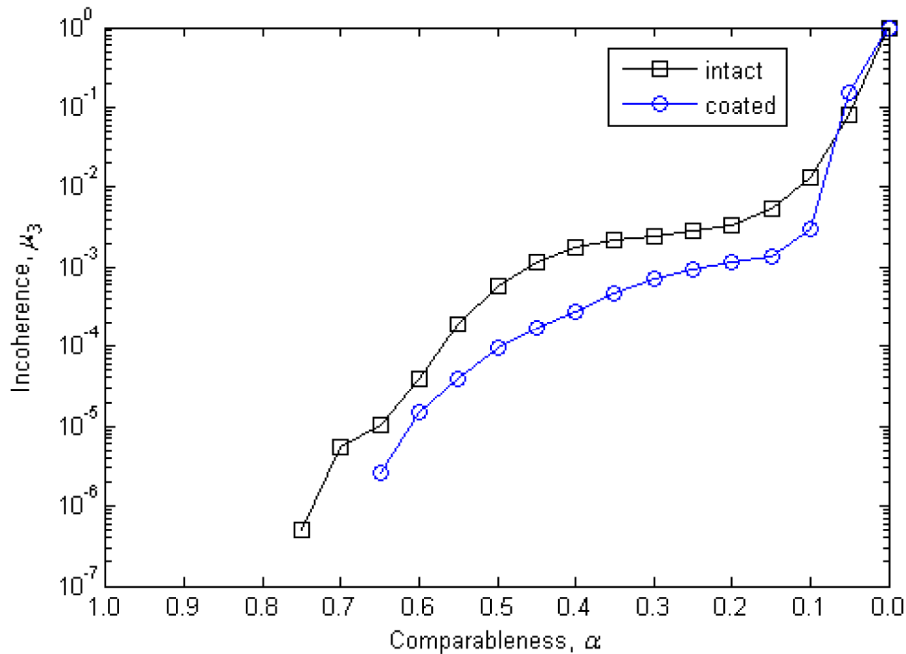


Fig. 5. The incoherence of TMs for intact MMF ($M = 2025$ and $N = 2000$) and coated MMF ($M = 2025$ and $N = 4000$). The incoherence of them is plotted in log scale.

4. Results

We now aim to compare the object image reconstruction capabilities with and without efficiently utilizing the HR encoding effect. For the conventional object wave reconstruction, we use PINV (Section 2.2). For the proposed reconstruction, we use the SR framework which is reported to take advantage of the HR encoding effect in a satisfactory manner [19]. For the sparse recovery in the SR framework, we use the alternating direction method [21] for its

efficiency. For the sparsifying basis in the sparse recovery, we use the Fourier basis directly, $\Psi = \mathbf{I}$. For a fair comparison between the two reconstructions, the reconstructed images are normalized so that their norms become one. For the TM measurements, $N = 2000$ and $N = 4000$ incident angles of the incoming waves are considered and the responses of them are captured respectively for the intact MMF and the coated MMF. The TMs cover the NA of 0.22 ($0^\circ \leq \theta_z \leq 12.71^\circ$) and the NA of 0.4 ($0^\circ \leq \theta_z \leq 23.58^\circ$). The dimension of M , the number of pixels in the CCD used in our experiment, corresponding to the output NA of 0.24 (Sec. 2.1) is 2025 ($M = 2025$). Thus, the TMs have the dimensions ($M \times N$) 2025×2000 and 2025×4000 for the intact fiber and for the coated fiber, respectively.

Intact MMF

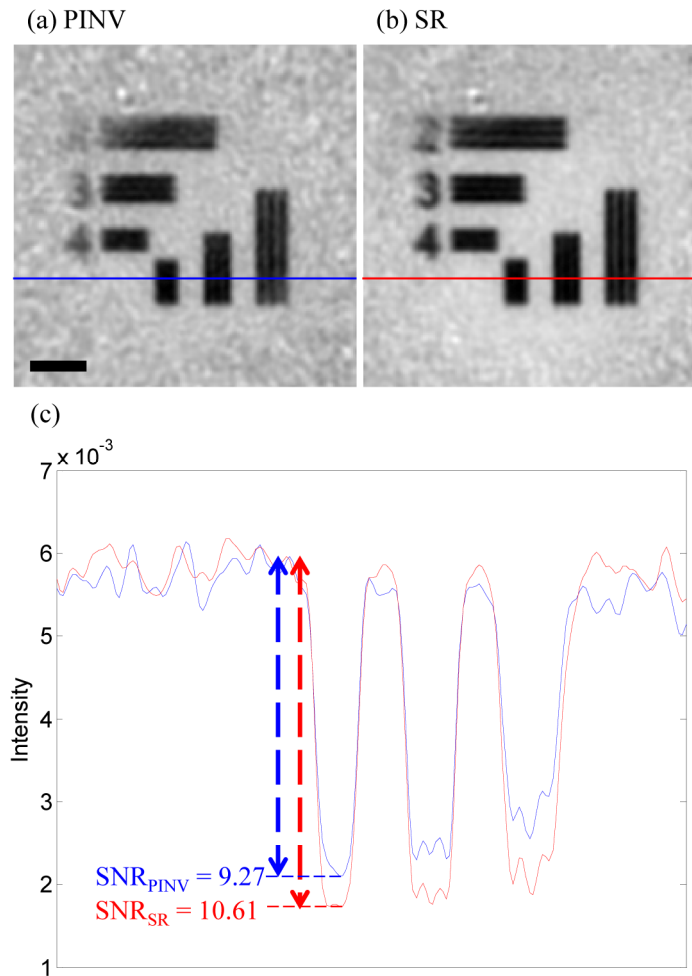


Fig. 6. Reconstruction for imaging through intact MMF. (a) Recovered amplitude image using PINV. (b) Recovered amplitude image using SR. (c) Cross sections of them. Images are averaged over 1000 samples. Scale bar: $10 \mu\text{m}$. SNRs are calculated in the cross sections.

Figure 6 shows the reconstructed images when the image is transferred through the MMF without depositing the turbid medium on the input facet of the fiber, hence there is no HR encoding effect as it should be. It is seen that the smallest structures in the object image cannot be resolved for both reconstructions. This is reasonable since higher modes of object waves which have incident angles beyond the acceptance angle of the intact MMF are not

captured. Here, both PINV and SR are shown to reconstruct the given signal well without significant perturbations; the system is not underdetermined. We found some image quality improvements when the SR is employed. For example, pay attention to the quality of reconstruction of the alphanumeric numbers 2, 3, and 4 at the image shown in Fig. 6. The number 2 at the upper left corner of the image becomes reconstructed and the numbers 3 and 4 becomes clearer. However, the improvement overall by SR is not very significant. The signal-to-noise ratio (SNR) is increased from 9.27 to 10.61; just a 14% increment. Here, the SNR was calculated as $SNR = |\bar{S}_{sig} - \bar{S}_{bg}| / \sigma_{bg}$ where \bar{S}_{sig} is the mean of the signal patterns, \bar{S}_{bg} is the mean of the background, and σ_{bg} is the standard deviation of the background.

Turbid media coated MMF

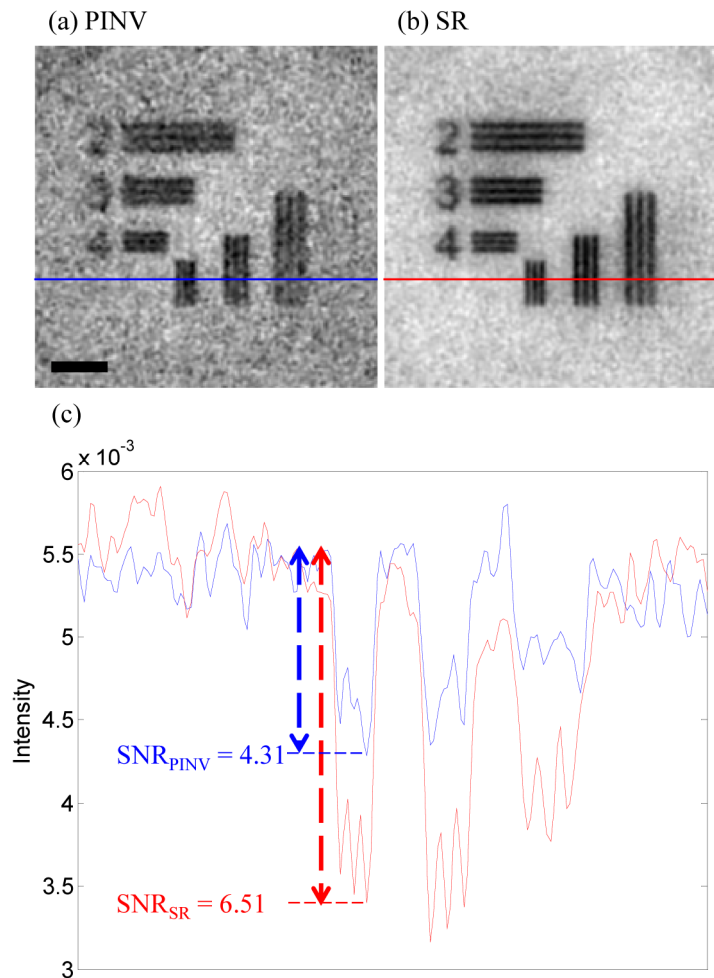


Fig. 7. Reconstruction for imaging through coated MMF. (a) Recovered amplitude image using PINV. (b) Recovered amplitude image using SR. (c) Cross sections of them. Images are averaged over 2000 samples. Scale bar: 10 μ m. SNRs are calculated in the cross sections.

Now, consider the recovered images when the turbid medium is used. It is found in Fig. 7 that the structures in the patterns are significantly improved in terms of resolution. This is true regardless of the use of recovery routines PINV or SR. Both methods deliver much improved reconstruction fidelity compared to those of the intact MMF. This is because higher mode

waves are introduced through the turbid medium. However, as it was already discussed, the use of turbid medium makes the system underdetermined. This makes the reconstruction of the object image difficult. As a result, it is shown that the reconstructed image by PINV becomes significantly perturbed by speckles. The SNR also becomes reduced considerably compared to that with the intact MMF.

We now show the results employing the SR framework for object wave recovery. As discussed earlier, the point here is to see if it will bring forth improved quality in reconstruction via efficient utilization of the HR encoding process offered by the use of turbid medium. As expected, in contrast to the case of PINV, SR is shown to improve the reconstruction significantly well. The speckle is successfully removed in the reconstructed image. We can also see all the small scale structures in the recovered image. The SNR becomes increased from 4.31 to 6.51; a 51% increment. With these, we have verified that the SR framework can exploit the HR encoding process of the turbid medium and improve the quality of the image reconstruction.

5. Conclusion

In conclusion, we demonstrated that the random scattering in the turbid media can be exploited for improving the quality of image reconstruction in MMF imaging. Random scattering through turbid medium provides random encoding of the object signal in holistic and incoherent manner. This encoding can be efficiently utilized in the signal recovery process within the proposed sparse representation framework. As a result, the perturbation is significantly reduced, the image contrast becomes sharper, and the fine details within the image can be captured.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (Do-Yak Research Program, No. 2013-035295).

Bayesian Hypothesis Test using Nonparametric Belief Propagation for Noisy Sparse Recovery

Jaewook Kang, *Member, IEEE*, Heung-No Lee, *Senior Member, IEEE*, and Kiseon Kim, *Senior Member, IEEE*

Abstract—This paper proposes a low-computational Bayesian algorithm for noisy sparse recovery (NSR), called BHT-BP. In this framework, we consider an LDPC-like measurement matrices which has a tree-structured property, and additive Gaussian noise. BHT-BP has a joint detection-and-estimation structure consisting of a sparse support detector and a nonzero estimator. The support detector is designed under the criterion of the minimum detection error probability using a nonparametric belief propagation (nBP) and composite binary hypothesis tests. The nonzeros are estimated in the sense of linear MMSE, where the support detection result is utilized. BHT-BP has its strength in noise robust support detection, effectively removing quantization errors caused by the uniform sampling-based nBP. Therefore, in the NSR problems, BHT-BP has advantages over CS-BP [13] which is an existing nBP algorithm, being comparable to other recent CS solvers, in several aspects. In addition, we examine impact of the minimum nonzero value of sparse signals via BHT-BP, on the basis of the results of [27],[28],[30]. Our empirical result shows that variation of x_{min} is reflected to recovery performance in the form of SNR shift.

Index Terms—Noisy sparse recovery, compressed sensing, nonparametric belief propagation, composite hypothesis testing, joint detection-and-estimation

I. INTRODUCTION

A. Background

Robust reconstruction of sparse signals against measurement noise is a key problem in real-world applications of compressed sensing (CS) [1]-[3]. We refer to such signal recovery problems as *noisy sparse signal recovery* (NSR) problems. The NSR problems can be directly defined as an l_0 -norm minimization problem [4],[5]. Solving the l_0 -norm task is very limited in practice when the system size (M, N) becomes large. Therefore, several alternative solvers have been developed to relax computational cost of the l_0 -norm task, such as l_1 -norm minimization solvers, *e.g.*, Dantzig selector (l_1 -DS) [6] and Lasso [7], and greedy type algorithms, *e.g.*, OMP [8] and COSAMP [9]. Another popular approach to the computational relaxation is based on the Bayesian philosophy [11]-[17]. In the Bayesian framework, the l_0 -norm task is described as maximum a posteriori (MAP) estimation problem, and sparse solution then is sought by imposing a

certain sparsifying prior probability density function (PDF) with respect to the target signal [10].

Recently, Bayesian solvers applying *belief propagation* (BP) have been introduced and caught attention as a low-computational approach to handle the NSR problems in a large system setup [13]-[17]. These BP-based solvers reduce computational cost of the signal recovery by removing unnecessary and duplicated computations using statistical dependency within the linear system. Such BP solvers are also called message-passing algorithms because their recovery behavior is well explained by passing statistical messages over a tree-structured graph representing the statistical dependency [18].

For implementation of BP, two approaches have been mainly discussed according to message representation methods: parametric BP (pBP) [15]-[17],[39],[40] where the BP-message is approximated to a Gaussian PDF; hence, only the mean and variance are used for message-passing, and nonparametric BP (nBP) [13],[14],[19]-[23] where the BP-message is represented by samples of the corresponding PDF. When the pBP approach is used, there are errors from the Gaussian approximation; these errors decrease as problem size (N, M) increases. If the nBP approach is used, there is an approximation error which generally depends upon the choice of message sampling methods.

B. Contribution

In this paper, a low-computational Bayesian algorithm is developed based on the nBP approach. We refer to the proposed algorithm as *Bayesian hypothesis test using nonparametric belief propagation* (BHT-BP)¹. Differently from the pBP-based solvers, BHT-BP can precisely infer the multimodally distributed BP-messages via an uniform sampling-based nBP. Therefore, BHT-BP can be applied to any types of sparse signals in the CS framework by adaptively choosing a signal prior PDF. In addition, the proposed algorithm uses *low-density parity-check codes* [24] (LDPC)-like sparse measurement matrices as works in [13],[15],[16]. Although such sparse matrices perform worse than the dense matrices do in terms of compressing capability in the CS framework, they can highly speed up the generation of the CS measurements [27].

Most CS algorithms to date for the NSR problems have been developed under the auspices of signal estimation rather than support detection. However, recently studies have indicated that the existing estimation-based algorithms, such as Lasso [7], lead to a potentially large gap with respect to the theoretical limit for the noisy support recovery [28]-[30]. Motivated by such theoretical investigation, the proposed BHT-BP takes a joint detection-and-estimation structure [31],[41], as shown in Fig.3, which consists of a sparse support detector

Manuscript received October 10, 2013; revised August 4, 2014; accepted December 8, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Raviv Raich. This work was supported in part by National Research Foundation of Korea under Grant of Do-yak Program 2013-035295, in part by Gwangju Institute of Science and Technology under a Grant of the Basic Research Project Program, and in part by Ministry of Science, ICT and Future Planning (2009-00422) under Grant K20901002277-12E0100-06010.

Portions of this work were presented at IEEE Statistical Signal Processing Workshop 2012 [14].

The authors are with Department of Information and Communication, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea (Email: {jwkkang,heungno,kskim}@gist.ac.kr)

¹The MATLAB code of the proposed algorithm is available at our webpage, <https://sites.google.com/site/jwkwang10/>

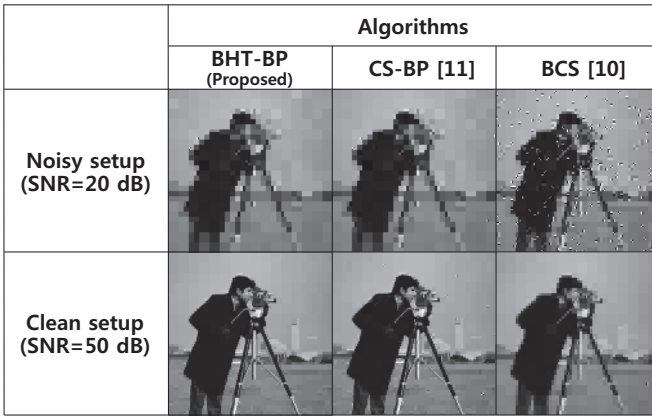


Fig. 1. An illustrative recovery example of BHT-BP (the proposed), CS-BP [13] and BCS [12] in the presence of noise. The original image, known as the Cameraman, of size $N = 128^2$, is transformed via three step discrete wavelet transform. For this example, we pad zeros for the coefficients having values below 100 in wavelet domain, and we recover these images from $M/N = 0.5$ undersampled measurements. From this example, we see that the recovered image via BCS includes flicker noise but which is not shown in that of BHT-BP in the noisy setup. In the clean setup, BHT-BP more clearly recovers the image than those of CS-BP and BCS.

and a nonzero estimator. The support detector uses uniform sampling-based nBP and composite binary hypothesis tests to the CS measurements \mathbf{Z} at hand for the sparse support finding. Given the detected support, the underdetermined CS problem is reduced to an overdetermined problem. Then, the nonzero estimator is applied under the criterion of *linear minimum mean-square-error* (LMMSE) [33]. Then, let us state the detailed novel points of the proposed algorithm. In the CS framework considering reconstruction of a sparse signal $\mathbf{X} \in \mathbb{R}^N$ from noisy measurements $\mathbf{Z} \in \mathbb{R}^M$, BHT-BP is novel in terms of

- 1) Providing robust support detection against additive measurement noise based on the *criterion of the minimum detection error probability*,
- 2) Removing MSE degradation caused by the message sampling of the uniform sampling-based nBP using a joint detection-and-estimation structure,
- 3) Handling sparse signals whose minimum nonzero value is regulated by a parameter $x_{\min} \geq 0$, proposing a signal prior PDF for such signals,
- 4) Providing fast sparse reconstruction with recovery complexity $\mathcal{O}(N \log N + KM)$ where K is the signal sparsity.

For the support detection of BHT-BP, we use a hypothesis-based detector designed under the criterion of the minimum detection error probability [32]. BHT-BP represents the signal support using a binary vector, scalarwisely applying the hypothesis testing to each binary element for the support finding. This hypothesis test is “composite” because the likelihood for the test is associated with the value of each scalar X_i . Therefore, we calculate the likelihood under the Bayesian paradigm; then, the likelihood for the test is a function of the signal prior and the marginal posterior of X_i . This is the reason why we refer to our support detection as *Bayesian hypothesis test* (BHT) detection. BHT-BP has noise robustness, outperforming the conventional algorithms, such as CS-BP [13], in the support detection. In this BHT detection, the nBP

part takes a role to provide the marginal posterior of X_i . Therefore, the advantage of BHT-BP in support detection can be claimed when the BP convergence is achieved with the sampling rate, $\frac{M}{N}$, above a certain threshold.

Typically, recovery performance of the nBP-based algorithms is dominated by the message sampling methods. In the case of CS-BP [13], its performance is corrupted by quantization errors because CS-BP works with the uniform sampling-based nBP such that the signal estimate is directly obtained from a sampled posterior. The joint detection-and-estimation structure of BHT-BP overcomes this weakpoint of CS-BP, improving MSE performance. The key behind the improvement is that the sampled posterior is only used for the support detection in BHT-BP. Furthermore, BHT-BP closely approaches to the oracle performance². in high SNR regime if the rate $\frac{M}{N}$ are sufficiently maintained for the signal sparsity K . Fig.1 is an illustration intended to see a motivational evidence of the recovery performance among the proposed BHT-BP, CS-BP [13] and BCS [12].

The importance of the minimum nonzero value x_{\min} of sparse signals \mathbf{X} in the NSR problems was highlighted by Wainwright *et al.* in [27],[28] and Fletcher *et al.* in [30], where they proved that the perfect support recovery is very difficult even with arbitrarily large *signal-to-noise ratio* (SNR) if x_{\min} is very small. Following these works, in the present work, we consider recovery of \mathbf{X} whose minimum nonzero value is regulated by x_{\min} . In addition, we propose to use a signal prior including the parameter x_{\min} , called *spike-and-dented slab* prior, investigating how helpful the knowledge of x_{\min} for the performance is. We empirically show in the BHT-BP recovery³ that variation of x_{\min} is reflected to the recovery performance in the form of SNR shift. In addition, we support this statement with a success rate analysis for the BHT support detection under the identity measurement matrix assumption, *i.e.*, $\Phi = \mathbf{I}$.

The recovery complexity of BHT-BP is $\mathcal{O}(N \log N + KM)$ which includes the cost $\mathcal{O}(KM)$ of the LMMSE estimation and that of the BHT support detection $\mathcal{O}(N \log N)$. This is advantageous compared to that of the l_1 -norm solvers $\Omega(N^3)$ [6],[7] and BCS $\mathcal{O}(NK^2)$ [12], being comparable to that of the recent BP-based algorithms using sparse measurement matrices $\mathcal{O}(N \log N)$: CS-BP [13] and SuPrEM [16].

C. Organization

The remainder of the paper is organized as follows. We first provide basic setup for our work in Section II. In Section III, we discuss our solution approach to the NSR problem. Section IV describes a nonparametric implementation of the BHT support detector and its computational complexity. Section V provides experimental validation to show performance and several aspects of the proposed algorithm, compared to the other related algorithms. Finally, we conclude this paper in Section VI.

²Here, the oracle performance means the performance of the LMMSE estimator having the knowledge of the sparse support set of the signal \mathbf{X} .

³To the best of our knowledge, we have not seen CS algorithms using x_{\min} as an input parameter.

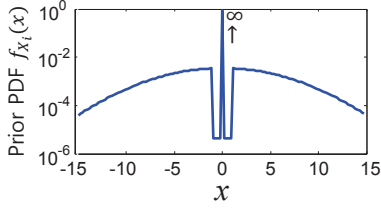


Fig. 2. Example of spike-and-dented slab PDF in log-scale where the prior is drawn with the parameters, $q = 0.05, \sigma_X = 5, x_{\min} = \frac{\sigma_X}{4}, \lambda = 10^{-4}$, and normalized to be $\int_{X_i} f_{X_i}(x) dx = 1$.

II. BASIC SETUP

In this section, we introduce our signal model, and a factor graphical model for linear systems used in this work.

A. Signal Model

Let $\mathbf{x}_0 \in \mathbb{R}^N$ denote a sparse vector which is a deterministic realization of a random vector \mathbf{X} . Here, we assume that the elements of \mathbf{X} are *i.i.d.*, and each X_i belongs to the support set with a sparsity rate $q \in [0, 1)$. To indicate the supportive state of \mathbf{X} , we use a state vector $\mathbf{S} \in \{0, 1\}^N$ whose each element S_i is Bernoulli random with the rate q as following

$$S_i = \begin{cases} 1, & \text{if } X_i \neq 0 \text{ with } q \\ 0, & \text{if } X_i = 0 \text{ with } 1 - q \end{cases}. \quad (1)$$

Then, the signal sparsity, $K = \|\mathbf{S}\|_0$, becomes Binomial random with $\mathcal{B}(k; N, q)$. In the present work, we consider the signal \mathbf{x}_0 whose minimum nonzero value is regulated by a parameter $x_{\min} \geq 0$. For such signal generation,

- We first draw a state vector \mathbf{s} by generating N *i.i.d.* Bernoulli numbers of (1).
- Then, we assign zero value to the signal scalars corresponding to $s_i = 0$, *i.e.*, $x_{0,i} = 0$.
- For the signal scalar corresponding to $s_i = 1$, a Gaussian number is drawn from $\mathcal{N}(x; 0, \sigma_X^2)$ and assigned to the signal scalar $x_{0,i}$ if $|x_{0,i}| \geq x_{\min}$; otherwise, the number is redrawn until a realization with $|x_{0,i}| \geq x_{\min}$ occurs.

For such signals with x_{\min} , we propose to use a *spike-and-dented slab* prior which is a variant of the spike-and-slab prior [36]. According to (1), the signal prior of X_i can be described as a two-state mixture PDF with the state S_i , *i.e.*,

$$f_{X_i}(x) = (1 - q)f_{X_i}(x|S_i = 0) + qf_{X_i}(x|S_i = 1). \quad (2)$$

Then, the spike-and-dented slab prior includes the conditional priors as following

$$\begin{aligned} f_{X_i}(x|S_i = 0) &= \delta(x), \\ f_{X_i}(x|S_i = 1) &\propto \begin{cases} \mathcal{N}(x; 0, \sigma_X^2), & |x| \geq x_{\min} \\ \lambda, & |x| < x_{\min} \end{cases} \end{aligned} \quad (3)$$

where $\delta(x)$ is the Dirac delta PDF and $\lambda > 0$ is a near-zero constant. Fig.2 shows an example of the spike-and-dented slab prior where the prior is drawn with the parameters, $q = 0.05, \sigma_X = 5, x_{\min} = \frac{\sigma_X}{4}, \lambda = 10^{-4}$, and normalized to be $\int_{X_i} f_{X_i}(x) dx = 1$.

The goal of the proposed algorithm is to recover the signal vector \mathbf{x}_0 from a noisy measurement vector

$$\mathbf{z} = \Phi \mathbf{x}_0 + \mathbf{w} \in \mathbb{R}^M, \quad (4)$$

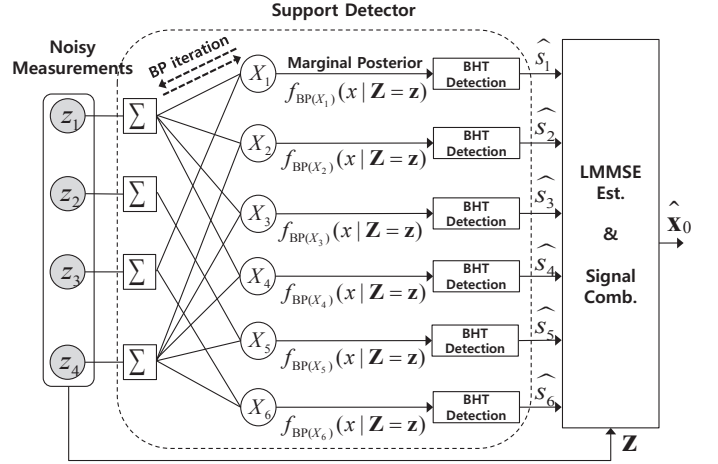


Fig. 3. Diagrammatic representation of the proposed algorithm (when $N = 6, M = 4, L = 2$) where the inputs of the proposed algorithm is the measurement $\mathbf{z} = [z_1, z_2, z_3, z_4]$ and the output is a signal estimate $\hat{\mathbf{x}}_0$. The proposed algorithm first detects the signal support $\hat{\mathbf{s}} = [\hat{s}_1, \dots, \hat{s}_6]$ from the measurements \mathbf{z} at hand, and then applies linear MMSE estimation to find the signal estimate $\hat{\mathbf{x}}_0$ given the detected support $\hat{\mathbf{s}}$.

given a fat measurement matrix $\Phi \in \{0, 1, -1\}^{M \times N}$ ($M < N$), where the vector $\mathbf{w} \in \mathbb{R}^M$ is a realization of a Gaussian random vector $\mathbf{W} \sim \mathcal{N}(\mathbf{w}; 0, \sigma_W^2 \mathbf{I})$; therefore, the vector $\mathbf{z} \in \mathbb{R}^M$ is drawn from a mean shifted Gaussian random vector conditioned on $\mathbf{X} = \mathbf{x}_0$, *i.e.*, $\mathbf{Z} \sim \mathcal{N}(\mathbf{z}; \Phi \mathbf{x}_0, \sigma_W^2 \mathbf{I})$. For the measurement matrix Φ , we consider an LDPC-like sparse matrix which has very low matrix sparsity (typically less than 1% matrix sparsity) and the tree-structured property [25],[26]. We regulate the matrix sparsity by the fixed column weight L such that $\mathbb{E}[\|\phi_{\text{column}}\|_2^2] = L$. This regulation enables the matrix Φ to span the measurement space with column vectors having equal energy.

B. Factor Graphical Modeling of Linear Systems

Factor graphs effectively represent such sparse linear systems in (4) [18]. Let $\mathcal{V} := \{1, \dots, N\}$ denote a set of variable nodes corresponding to the signal elements, $\mathbf{x}_0 = [x_{0,1}, \dots, x_{0,N}]$, and $\mathcal{C} := \{1, \dots, M\}$ denote a set of factor nodes corresponding to the measurement elements, $\mathbf{z} = [z_1, \dots, z_M]$. In addition, we define a set of edges connecting \mathcal{V} and \mathcal{C} as $\mathcal{E} := \{(j, i) \in \mathcal{C} \times \mathcal{V} \mid \phi_{ji} = 1\}$ where ϕ_{ji} is the (j, i) -th element of Φ . Then, a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{C}, \mathcal{E})$ fully describes the neighboring relation in the sparse linear system. For convenience, we define the neighbor set of \mathcal{V} and \mathcal{C} as $N_{\mathcal{V}}(i) := \{j \in \mathcal{C} \mid (j, i) \in \mathcal{E}\}$ and $N_{\mathcal{C}}(j) := \{i \in \mathcal{V} \mid (j, i) \in \mathcal{E}\}$, respectively. Note that the column weight of the matrix Φ is expressed as $L = |N_{\mathcal{V}}(i)|$ in this graph model.

III. SOLUTION APPROACH OF PROPOSED ALGORITHM

The proposed algorithm, BHT-BP, has a joint detection-and-estimation structure where we first detect the sparse support by a combination of BP and BHT, then estimating nonzeros in the detected support by an LMMSE estimator, as shown in Fig.3. In this section, we provide our solution approach to the support detection and the nonzero estimation under the joint structure.

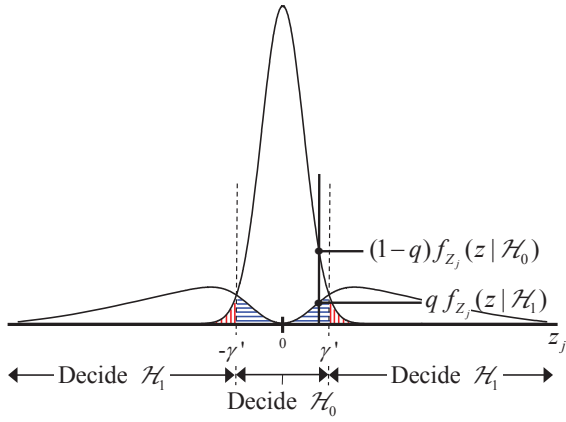


Fig. 4. Fig.4 illustrates the scalar state detection by BHT under an assumption of $\Phi = \mathbf{I}$. Under this assumption, “the hypothesis test given a vector \mathbf{z} ” is simplified to “the test given a scalar z_j ”, described in (8), where the threshold γ' is derived from the equality condition of (9). In the figure, the horizontal-lined region (blue) represents $\Pr\{\hat{s}_i \neq s_i|\mathcal{H}_1\}$ and the vertical-lined (red) region does $\Pr\{\hat{s}_i \neq s_i|\mathcal{H}_0\}$.

A. Support Detection using Bayesian Hypothesis Testing

1) *Support detection in BHT-BP*: The support detection problems can be decomposed to a sequence of binary state detection problems given the marginal posterior $f_{X_i}(x|\mathbf{Z} = \mathbf{z})$ of each signal scalar. Our state detection problem is to choose one between the two hypotheses:

$$\mathcal{H}_0 : S_i = 0 \text{ and } \mathcal{H}_1 : S_i = 1,$$

given the measurements \mathbf{z} . Our methodology to this problem is related to *Bayesian composite hypothesis testing* [32, p. 198]. In contrast to the simple hypothesis test where the PDFs under both hypothesis are perfectly specified, the composite hypothesis test must consider associated random variables. In our problem, the associated random variable is X_i . Then, the binary state detector decides \mathcal{H}_1 if

$$\frac{f_{\mathbf{z}}(\mathbf{z}|\mathcal{H}_1)}{f_{\mathbf{z}}(\mathbf{z}|\mathcal{H}_0)} = \frac{\int f_{\mathbf{z}}(\mathbf{z}|\mathcal{H}_1, X_i = x) f_{X_i}(x|\mathcal{H}_1) dx}{\int f_{\mathbf{z}}(\mathbf{z}|\mathcal{H}_0, X_i = x) f_{X_i}(x|\mathcal{H}_0) dx} > \gamma, \quad (5)$$

where γ is a threshold for the test. The PDF $f_{\mathbf{z}}(\mathbf{z}|X_i = x)$ is simplified to $f_{\mathbf{z}}(\mathbf{z}|X_i = x)$ since the hypothesis \mathcal{H}_{s_i} and the measurements \mathbf{Z} are conditionally independent given X_i . Therefore, finally, the binary hypothesis test in (5) can be rewritten as

$$T_{\text{BHTBP}}(\mathbf{z}) := \frac{\int \frac{f_{X_i}(x|S_i=1)}{f_{X_i}(x)} f_{X_i}(x|\mathbf{Z} = \mathbf{z}) dx}{\int \frac{f_{X_i}(x|S_i=0)}{f_{X_i}(x)} f_{X_i}(x|\mathbf{Z} = \mathbf{z}) dx} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma \quad (6)$$

where the Bayesian rule is applied to $f_{\mathbf{z}}(\mathbf{z}|X_i = x) = \frac{f_{X_i}(x|\mathbf{Z}=\mathbf{z})f_{\mathbf{z}}(\mathbf{z})}{f_{X_i}(x)}$, and obviously $f_{X_i}(x|\mathcal{H}_{s_i}) = f_{X_i}(x|S_i = s_i)$ holds from the prior knowledge of (2).

In some detection problem under Bayesian paradigm, one can reasonably assign prior probabilities to the hypotheses. In the present work, we assign the sparsity rate q to the hypotheses, *i.e.*, $\Pr\{\mathcal{H}_0\} = 1 - q$ and $\Pr\{\mathcal{H}_1\} = q$. Then, we can define the state error rate (SER) of the scalar state detection (6) [32, p. 78]

$$P_{\text{SER}} := \Pr\{\hat{s}_i \neq s_i|\mathcal{H}_0\}(1 - q) + \Pr\{\hat{s}_i \neq s_i|\mathcal{H}_1\}q. \quad (7)$$

It is well known that the threshold γ of (6) can be optimized under *the criterion of the minimum detection error probability* with the SER expression (7). By the criterion, we assign the threshold to $\gamma = \gamma^* := \frac{1-q}{q}$. We omit the derivation for this threshold optimization here, referring interested readers to [32, p. 90]. We call this binary hypothesis test (6) with the threshold γ^* as *Bayesian hypothesis test (BHT)* detection. The proposed algorithm generates a detected support $\hat{\mathbf{s}} \in \{0, 1\}^N$ according to the results of a sequence of BHTs. Therefore, given a marginal posterior of each X_i , BHT-BP can robustly detect the signal support even when the measurements are noisy.

Fig.4 illustrates the scalar state detection of BHT-BP when the matrix is $\Phi = \mathbf{I}$ such that the measurement channel can be decoupled to N scalar Gaussian channels, *i.e.*, $Z_j = X_i + W_j$, ($i = j$). Under this assumption, “the hypothesis test given a vector \mathbf{z} ” can be scalarwise to “the test given a scalar z_j ”, being simplified

$$\forall j \in \mathcal{C} : |z_j| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma' \quad (8)$$

where the threshold γ' is derived from the equality condition with the two scalar likelihood and the threshold $\gamma^* = \frac{1-q}{q}$,

$$\frac{f_{Z_j}(z|\mathcal{H}_1)}{f_{Z_j}(z|\mathcal{H}_0)} = \gamma^*. \quad (9)$$

Hence, the threshold γ' is a function of σ_X , σ_W , x_{\min} , and q (see Appendix II). With this threshold γ' , we can find the conditional SER, $\Pr\{\hat{s}_i \neq s_i|\mathcal{H}_{s_i}\}$, for the case $\Phi = \mathbf{I}$. In Fig.4, the horizontal-lined region (blue) represents $\Pr\{\hat{s}_i \neq s_i|\mathcal{H}_1\}$ and the vertical-lined (red) region does $\Pr\{\hat{s}_i \neq s_i|\mathcal{H}_0\}$. The corresponding SER analysis will be provided in Appendix II. Although Fig.4 does not show typical behavior of the BHT detection given a vector measurement \mathbf{z} , the figure helps intuitive understanding of the BHT detection.

In addition, it is noteworthy in Fig.4 that the shape of $f_{Z_j}(z|\mathcal{H}_1)$ is dented near $z_j = 0$. This is caused by the use of the spike-and-dented slab prior, given in (3), where the dented part varies with the parameter x_{\min} .

2) *Support detection of CS-BP*: Support detection is not performed in practical recovery of CS-BP, but we describe it here for a comparison purpose. CS-BP estimates the sparse solution $\hat{\mathbf{x}}_0$ directly from a BP approximation of the signal posterior, through MAP or MMSE estimation. Let us consider CS-BP using the MAP estimation. Then, given the marginal posterior $f_{X_i}(x|\mathbf{Z} = \mathbf{z})$, the scalar state detection of CS-BP is equivalent to choose one of the two peaks at $x = 0$ and $x = \hat{x}_{\text{MAP},i} := \arg \max_x f_{X_i}(x|\mathbf{Z} = \mathbf{z})$. Namely, the binary state detector of CS-BP decides \mathcal{H}_1 if

$$T_{\text{CSBP}}(\mathbf{z}) := \frac{\Pr\{\hat{x}_{\text{MAP},i} - \Delta x < X_i \leq \hat{x}_{\text{MAP},i} + \Delta x|\mathbf{Z} = \mathbf{z}\}}{\Pr\{0 - \Delta x < X_i \leq 0 + \Delta x|\mathbf{Z} = \mathbf{z}\}} = \frac{\int_{\hat{x}_{\text{MAP},i} - \Delta x}^{\hat{x}_{\text{MAP},i} + \Delta x} f_{X_i}(x|\mathbf{Z} = \mathbf{z}) dx}{\int_{0 - \Delta x}^{0 + \Delta x} f_{X_i}(x|\mathbf{Z} = \mathbf{z}) dx} > 1, \quad (10)$$

where Δx is a small quantity that we eventually let approach to 0. When $\hat{x}_{\text{MAP},i} = 0$, the test cost becomes one; then, the detector immediately decides \mathcal{H}_0 . Hence, in CS-BP, the

detected support $\hat{\mathbf{s}}$ is just a by-product of the signal estimate $\hat{\mathbf{x}}_{\text{MAP}}$, which is not robust support detection against additive measurement noise.

B. Conditions for BP Convergence

In the proposed algorithm, marginal posterior of each X_i , $f_{X_i}(x|\mathbf{Z} = \mathbf{z})$, for the BHT detection is computed by BP. It was known that BP efficiently computes such marginal posteriors, achieving its convergence if the conditions in Note 1 are satisfied [42]. Given the BP convergence, each approximate marginal posterior converges to a PDF peaked at a unique value \hat{x}_i during the iteration. In noiseless setup, the unique value is exactly the true value, *i.e.*, $\hat{x}_i = x_{0,i}$.

Note 1 (Conditions for BP convergence):

- The factor graph, which corresponds to the relation between \mathbf{X} and \mathbf{Z} , has a tree-structure.
- Sufficiently large number of iterations l is maintained such that BP-messages have been propagated along every link of the tree, and a variable node has received messages from all the other variables nodes.

Although the second condition in Note 1 is practically demanding, it has been reported that BP provides a good approximation of marginal posteriors even with factor graphs including cycles, which is called loopy BP [25],[39],[40].

A related argument for BP was stated by Guo and Wang in the context of the multiuser detection problem of CDMA systems, where the problem is actually equivalent to solve a linear system [37],[38]. In the works, Guo and Wang showed that the marginal posterior computed by BP is almost exact in a large linear system ($M, N \rightarrow \infty$) if the factor graph corresponding to the matrix Φ is asymptotically cycle-free and the sampling rate $\frac{M}{N}$ is above a certain threshold⁴. Namely, Guo and Wang showed that

$$\lim_{l \rightarrow \infty} \limsup_{M, N \rightarrow \infty} \left| f_{\text{BP}(X_i)}^{(l)}(x|\mathbf{Z} = \mathbf{z}) - f_{X_i}(x|\mathbf{Z} = \mathbf{z}) \right| = 0, \quad (11)$$

where $f_{\text{BP}(X_i)}^{(l)}(x|\mathbf{Z} = \mathbf{z})$ is an approximate marginal posterior of each X_i by l iterations of BP.

According to the literature, in the linear system with the LDPC-like matrix Φ , the sampling rate $\frac{M}{N}$ is the only obstacle for the BP convergence. The asymptotic condition used in (11) is not always necessary if the tree-structured property is guaranteed for the matrix Φ because the main reason to use the asymptotic condition in the works of [37],[38] is to make the system ‘‘asymptotically cycle-free’’, which is equivalent to having an ‘‘asymptotically tree-structured’’ matrix Φ ⁵. Thus, we claim the advantage of BHT-BP over CS-BP in support detection with a certain threshold of the rate $\frac{M}{N}$. Given $\frac{M}{N}$ below the threshold, the BP convergence is not achieved such that the likelihood $f_{\mathbf{Z}}(\mathbf{z}|\mathcal{H}_{s_i})$ is not properly calculated for the BHT detection. We will empirically find the threshold using information entropy of the approximate marginal posterior, $f_{\text{BP}(X_i)}^{(l)}(x|\mathbf{Z} = \mathbf{z})$, in Section V-A. Although we do not provide an analytical threshold of $\frac{M}{N}$ for the BP convergence in this paper, simulation results with $\frac{M}{N}$ above the empirical threshold are quite favorable, as shown in Section V-B and -C.

⁴In [37],[38], the authors considered the sampling rate $\frac{M}{N}$ above one.

⁵If the graph corresponding to the matrix Φ has at least one cycle, the BP convergence cannot be rigorously guaranteed.

C. LMMSE Estimation of Nonzero Values

Given the support information by the BHT detection, the rest of the work is reduced to the nonzero estimation problem, represented as

$$\hat{\mathbf{x}}_0 = \mathbb{E}[\mathbf{X}|\mathbf{S} = \hat{\mathbf{s}}, \mathbf{Z} = \mathbf{z}], \quad (12)$$

and it can straightforwardly solved by combining the nonzero position by $\hat{\mathbf{s}}$ and the nonzero values given by the LMMSE estimate [33, p. 364]

$$\hat{\mathbf{x}}_{0,\hat{\mathbf{s}}} = \left(\frac{1}{\sigma_X^2} \mathbf{I} + \frac{1}{\sigma_W^2} \Phi_{\hat{\mathbf{s}}}^T \Phi_{\hat{\mathbf{s}}} \right)^{-1} \frac{1}{\sigma_W^2} \Phi_{\hat{\mathbf{s}}}^T \mathbf{z}, \quad (13)$$

where $\Phi_{\hat{\mathbf{s}}} \in \{0, 1\}^{M \times K}$ denotes a submatrix of Φ that contains only the columns corresponding to the detected support $\hat{\mathbf{s}}$, σ_X^2 are the variance of an nonzero scalar X_i .

The estimate $\hat{\mathbf{x}}_0$ from the proposed joint detection-and-estimation structure is not optimal. As we have seen, our support detector (6) is based on the criterion of minimum detection error probability. Even with this detector, however, we cannot guarantee the estimation optimality since the LMMSE estimator of (12) is not designed from the cost function involving the detection part [31],[41]. Nevertheless, worth mentioning here is that the proposed joint structure has advantages as given in Note 2.

Note 2 (Claims from the joint detection-and-estimation structure):

- Removing the MSE degradation caused by the uniform sampling-based nBP.
- Achieving the oracle performance in the high SNR regime with the sufficiently high rate $\frac{M}{N}$ for the BP convergence.

We will empirically validate this claim in Section V-C.

IV. NONPARAMETRIC IMPLEMENTATION OF BHT SUPPORT DETECTOR

This section describes a nonparametric implementation of the proposed support detector consisting of BP and the BHT detection. We discuss our nonparametric approach of the BP part first, and then explain the BHT detection part. This BHT support detection is summarized in Algorithm 1.

A. Nonparametric BP using Uniform Sampling

In the BHT support detector, the BP part provides the marginal posterior of X_i for the hypothesis test in (6). Since the signal \mathbf{x}_0 is real valued, each BP-message takes the form of a PDF, and the BP-iteration becomes a density-message-passing process. To implement the density-message-passing, we take the nBP approach [19]-[22]. Many nBP algorithms have been proposed according to several message sampling methods such as discarding samples having low probability density [20], adaptive sampling [21], Gibbs sampling [19], rejection sampling [22], or importance sampling [23].

Our nBP approach is to use an uniform sampling for the message representation where we set the sampling step T_s on the basis of *the three sigma-rule* [35] such that

$$T_s = \frac{2 \cdot 3\sigma_X}{N_d} \quad (14)$$

where N_d is the number of samples to store a BP-message. Then, we define the uniform sampling of a density-message $f(x)$ as

$$\text{Samp}\{f(x); T_s\} := f(mT_s - 3\sigma_X) \\ = f[m], \text{ for } m \in \{0, 1, \dots, N_d - 1\}, \quad (15)$$

where $\text{Samp}\{\cdot; T_s\}$ denotes the uniform sampling function with the step size T_s . Hence, the sampled message $f[m]$ can be treated as a vector with size N_d by omitting the index m . The main strength of the uniform sampling-based nBP is adaptivity to various signal prior PDFs. In addition, we note that the calculation of uniformly sampled messages can be accelerated using the *Fast fourier transform* (FFT).

Consider the factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{C}, \mathcal{E})$ depicted in the support detection part of Fig.3 where a signal element X_i corresponds to a variable node $i \in \mathcal{V}$ and a measurement element Z_j corresponds to a factor node $j \in \mathcal{C}$. At every iteration, messages are first passed from each variable node $i \in \mathcal{V}$ to its neighboring factor nodes $N_{\mathcal{V}}(i)$; each factor nodes $j \in \mathcal{C}$ then calculates messages to pass back to the neighboring variable nodes $N_{\mathcal{C}}(j)$ based on the previously received messages. These factor-to-variable (FtV) messages include *extrinsic* information of X_i , and will then be employed for the computation of updated variable-to-factor (VtF) messages in the next iteration. (For the detail, see the paper [18]).

Let $\mathbf{a}_{i \rightarrow j}^{(l)} \in [0, 1]^{N_d}$ denote a sampled VtF message at the l -th iteration in the vector form, given as

$$\mathbf{a}_{i \rightarrow j}^{(l)} = \eta \left[\mathbf{p}_{X_i} \times \prod_{k \in N_{\mathcal{V}}(i) \setminus \{j\}} \mathbf{b}_{k \rightarrow i}^{(l-1)} \right] \quad \forall (j, i) \in \mathcal{E}, \quad (16)$$

where all product operations are elementwise, the vector $\mathbf{p}_{X_i} \in [0, 1]^{N_d}$ denotes the sampled signal prior, *i.e.*, $\mathbf{p}_{X_i} := \text{Samp}\{f_{X_i}(x), T_s\}$, and $\eta[\cdot]$ is a normalization function to make $\sum \mathbf{a}_{i \rightarrow j}^{(l)} = 1$. The sampled FtV message at the l -th iteration, $\mathbf{b}_{j \rightarrow i}^{(l)} \in [0, 1]^{N_d}$, is defined as

$$\mathbf{b}_{j \rightarrow i}^{(l)} = \mathbf{p}_{Z_j|\mathbf{X}} \otimes \left(\bigotimes_{k \in N_{\mathcal{C}}(j) \setminus \{i\}} \mathbf{a}_{k \rightarrow j}^{(l)} \right) \quad \forall (j, i) \in \mathcal{E}, \quad (17)$$

where \otimes is the operator for the linear convolution of vectors, and the vector $\mathbf{p}_{Z_j|\mathbf{X}} \in [0, 1]^{N_d}$ is the sampled measurement PDF, *i.e.*, $\mathbf{p}_{Z_j|\mathbf{X}} := \text{Samp}\{\mathcal{N}(z_j; (\Phi\mathbf{X})_j, \sigma_W^2), T_s\}$.

The convolution operations in (17) can be efficiently computed by using FFT. Accordingly, we can rewrite the FtV message calculation as

$$\mathbf{b}_{j \rightarrow i}^{(l)} = \mathcal{F}^{-1} \left\{ \mathcal{F} \mathbf{p}_{Z_j|\mathbf{X}} \times \left(\prod_{k \in N_{\mathcal{C}}(j) \setminus \{i\}} \mathcal{F} \mathbf{a}_{k \rightarrow j}^{(l)} \right) \right\} \quad (18)$$

where \mathcal{F} denotes the FFT operation. Therefore, for efficient use of FFT, the sampling step T_s should be appropriately chosen such that N_d is power of two. In fact, the use of FFT brings a small calculation gap since the FFT-based calculation performs a circular convolution. However, this gap can be ignored, especially when the messages take the form of bell-shaped PDFs such as Gaussian PDFs.

The sampled approximation of the marginal posterior of each X_i , *i.e.*, $\mathbf{p}_{X_i|\mathbf{Z}}^{(l)} := \text{Samp}\{f_{\text{BP}(X_i)}^{(l)}(x|\mathbf{Z} = \mathbf{z}), T_s\}$, is

Algorithm 1 BHT support detection

Inputs: Noisy measurements \mathbf{z} , measurement matrix Φ , sparsity rate q , sampled prior PDF \mathbf{p}_X , sampled measurement PDF $\mathbf{p}_{Z_j|\mathbf{X}}$, The number of samples N_d , Termination condition ε .

Outputs: Reconstructed signal $\hat{\mathbf{x}}_0$, Detected support vector $\hat{\mathbf{s}}$.

1) Belief propagation:

set $\mathbf{b}_{j \rightarrow i}^{(l=0)} = \mathbf{1}$ for all $(j, i) \in \mathcal{E}$

while $\frac{1}{N} \sum_{i=1}^N \frac{\|\mathbf{p}_{X_i|\mathbf{Z}}^{(l)} - \mathbf{p}_{X_i|\mathbf{Z}}^{(l-1)}\|_2^2}{\|\mathbf{p}_{X_i|\mathbf{Z}}^{(l)}\|_2^2} > \varepsilon$ **do**

$\forall (j, i) \in \mathcal{E}$:

set $\mathbf{a}_{i \rightarrow j}^{(l)} = \eta \left[\mathbf{p}_{X_i} \times \prod_{k \in N_{\mathcal{V}}(i) \setminus \{j\}} \mathbf{b}_{k \rightarrow i}^{(l-1)} \right]$

set $\mathbf{b}_{j \rightarrow i}^{(l)} = \mathbf{p}_{Z_j|\mathbf{X}} \otimes \left(\bigotimes_{k \in N_{\mathcal{C}}(j) \setminus \{i\}} \mathbf{a}_{k \rightarrow j}^{(l)} \right)$

$\forall i \in \mathcal{V}$:

set $\mathbf{p}_{X_i|\mathbf{Z}}^{(l)} = \eta \left[\mathbf{a}_{i \rightarrow j^*}^{(l)} \times \mathbf{b}_{j^* \rightarrow i}^{(l-1)} \right]$

end while

2) BHT detection:

$\forall i \in \mathcal{V}$:

if $\log \frac{\sum \mathbf{r}_1 \times \mathbf{p}_{X_i|\mathbf{Z}}}{\sum \mathbf{r}_0 \times \mathbf{p}_{X_i|\mathbf{Z}}} > \log \frac{1-q}{q}$ **then** set $\hat{s}_i = 1$

else set $\hat{s}_i = 0$

end if

produced by using the FtV message (17) for every $i \in \mathcal{V}$. Namely,

$$\mathbf{p}_{X_i|\mathbf{Z}}^{(l)} = \eta \left[\mathbf{p}_{X_i} \times \prod_{k \in N_{\mathcal{V}}(i)} \mathbf{b}_{k \rightarrow i}^{(l-1)} \right] \quad \forall i \in \mathcal{V}, \quad (19)$$

To terminate the BP loop, we test the condition at every iteration, which is given as

$$\frac{1}{N} \sum_{i=1}^N \frac{\|\mathbf{p}_{X_i|\mathbf{Z}}^{(l)} - \mathbf{p}_{X_i|\mathbf{Z}}^{(l-1)}\|_2^2}{\|\mathbf{p}_{X_i|\mathbf{Z}}^{(l)}\|_2^2} \leq \varepsilon \quad (20)$$

where $\varepsilon > 0$ is a constant for the termination condition. If the condition given in (20) is satisfied, the BP loop will be terminated. After the BP termination, we can simply express the marginal posterior of X_i by dropping out the iteration index l , *i.e.*, $\mathbf{p}_{X_i|\mathbf{Z}}$.

B. BHT Detection using Sampled Marginal Posterior

We perform the hypothesis test in (6) by scaling it in logarithm. Using the sampled marginal posterior obtained from the BP part, a nonparametric implementation of the hypothesis test in (6) is given as

$$\log \frac{\sum \mathbf{r}_1 \times \mathbf{p}_{X_i|\mathbf{Z}}}{\sum \mathbf{r}_0 \times \mathbf{p}_{X_i|\mathbf{Z}}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \log \frac{1-q}{q} \quad (21)$$

where \times is elementwise multiplication of vectors, and $\mathbf{r}_0, \mathbf{r}_1 \in \mathbb{R}^{N_d}$ are reference vectors from the signal prior knowledge, defined as

$$\mathbf{r}_0 := \frac{\mathbf{p}_{X_i|S_i=0}}{\mathbf{p}_{X_i}}, \quad \mathbf{r}_1 := \frac{\mathbf{p}_{X_i|S_i=1}}{\mathbf{p}_{X_i}}. \quad (22)$$

This BHT-based detector is only compatible with the nBP approach because the BHT detection requires full information

TABLE I
LIST OF ALGORITHMS IN THE PERFORMANCE VALIDATION

Algorithms	Complexity	Type of Φ	Type of Prior PDFs	Utilized Techniques
BHT-BP (Proposed)	$\mathcal{O}(N \log N + KM)$	LDPC-like	Spike-and-dented slab	nBP
CS-BP [13]	$\mathcal{O}(N \log N)$	LDPC-like	Spike-and-dented slab	nBP
SuPrEM [16]	$\mathcal{O}(N \log N)$	LDF	Two-layer Gaussian with Jeffery	EM, pBP
BCS [12]	$\mathcal{O}(NK^2)$	LDPC-like	Two-layer Gaussian with Gamma	EM
l_1 -DS [6]	$\Omega(N^3)$	Std. Gaussian	-	CVX opt.

on the multimodally distributed posterior of X_i which cannot be provided through the pBP approach.

C. Computational Complexity

In our uniform sampling-based nBP, the density-messages are vectors with size N_d . Therefore, the decoder requires $\mathcal{O}(LN_d)$ flops to calculate a VtF message $\mathbf{a}_{i \rightarrow j}^{(l)}$ and $\mathcal{O}(\frac{N LN_d}{M} \log N_d)$ flops for a FtV message $\mathbf{b}_{j \rightarrow i}^{(l)}$ per iteration. In addition, the cost of the FFT-based convolution given in (18) spends $\mathcal{O}(N_d \log N_d)$ flop if we assume the row weight is NL/M in average sense. Hence, the per-iteration cost of the uniform sampling-based nBP is $\mathcal{O}(N LN_d + M \frac{N LN_d}{M} \log N_d) \approx \mathcal{O}(N LN_d \log N_d)$ flops. For the BHT detection, the decoder requires $\mathcal{O}(N_d)$ flops to generate the likelihood ratio of (21), which is much smaller than that of the BP part. Therefore, the cost for the BHT detection can be ignored.

For the linear MMSE estimation to find nonzeros on the support, the cost can be reduced upto $\mathcal{O}(KM)$ flops by applying QR decomposition [34]. Thus, the total complexity of the proposed algorithm is $\mathcal{O}(l^* \times N LN_d \log N_d + KM)$ flops and it is further simplified to $\mathcal{O}(l^* \times N + KM)$ since L and N_d are fixed constants. In addition, it is known that the message-passing process is applied recursively until messages have been propagated along with every edge in the tree-structured graph, and every signal element has received messages from all of its neighborhood, which requires $l^* = \mathcal{O}(\log N)$ iterations [13],[25],[42]. Therefore, we finally obtain $\mathcal{O}(N \log N + KM)$ for the complexity of the proposed algorithm, BHT-BP.

V. PERFORMANCE VALIDATION

We validate performance of the proposed algorithm, BHT-BP, with extensive experimental results. Four types of experimental results are discussed in this section, as given below:

- 1) Threshold $(\frac{M}{N})^*$ for BP convergence,
- 2) Support detection performance over SNR,
- 3) MSE comparison to recent algorithms over SNR,
- 4) Empirical calibration of BHT-BP over N_d and L .

The support detection performance is evaluated in terms of the success rate of perfect support detection, defined as

$$P_{\text{succ}} := \Pr\{\widehat{\mathbf{s}} = \mathbf{s} | \mathbf{Z} = \mathbf{z}\}, \quad (23)$$

and the MSE comparison to the other algorithms is performed in terms of normalized MSE, given as

$$\text{MSE} := \frac{\|\widehat{\mathbf{x}}_0 - \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2}. \quad (24)$$

We generate all the experimental results by averaging the measures, given in (23) and (24), with respect to the signal

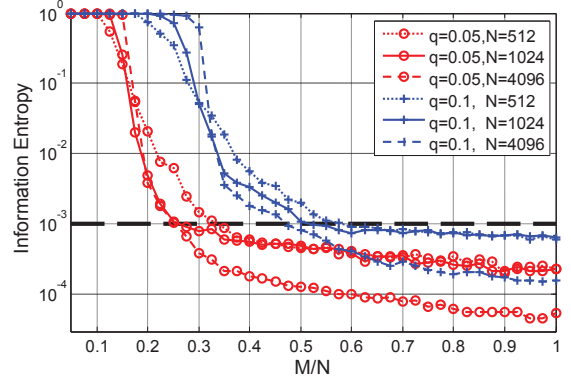


Fig. 5. The entropy phase transition curve over the sampling rate $\frac{M}{N}$ for a variety of the signal length N and the sparsity rate q where we set the threshold $(\frac{M}{N})^*$ to the point achieving $\frac{1}{N} \sum_{i=1}^N h(X_i | \mathbf{Z} = \mathbf{z}) \leq 10^{-3}$, which is given in Table II. These curves are information entropy of the approximate marginal posterior, $f_{\text{BP}(X_i)}^{(l)}(x | \mathbf{Z} = \mathbf{z})$, drawn with the parameters $\sigma_X = 5$, $x_{\min} = \sigma_X/4$, $N_d = 256$, $\varepsilon = 10^{-5}$, $\lambda = 10^{-4}$ and a noiseless setup. In addition, we set the column weight of the matrix Φ to $L = 4$ for $N = 512, 1024$, and $L = 5$ for $N = 4096$, in this experiment.

\mathbf{x}_0 and the additive noise \mathbf{w} using Monte Carlo method⁶. In addition, we define a SNR measure used in the experiment as

$$\text{SNR} := 10 \log_{10} \frac{\mathbb{E} \|\Phi \mathbf{X}\|_2^2}{M \sigma_W^2} \quad (\text{dB}). \quad (25)$$

For a comparison purpose, in this validation, we include several recent Bayesian algorithms, CS-BP [13], BCS [12], and SuPrEM [16], as well as an l_1 -norm based algorithm, l_1 -DS [6]⁷. We provide brief introduction to the Bayesian algorithms in Appendix I for interested readers. In this validation, BHT-BP and CS-BP use the spike-and-dented slab prior, given in (3), by applying the uniform sampling, *i.e.*, $\mathbf{p}_{X_i} := \text{Samp}\{f_{X_i}(x), T_s\}$. Worth mentioning here is that the nBP-based solvers, such as BHT-BP and CS-BP, are only compatible with such an unusual signal prior, like the spike-and-dented slab prior, which is one main advantage of the nBP solvers. For the measurement matrix Φ , we basically consider a LDPC-like matrix in BHT-BP, CS-BP and BCS. In case of SuPrEM, a LDF matrix is used for the measurement generation⁸, and l_1 -DS is performed with the standard Gaussian matrix as a benchmark of the CS recovery. For fair comparison, all types of the matrices Φ are equalized to have

⁶At every Monte Carlo trial, we realize \mathbf{x}_0 and \mathbf{w} to produce a measurement vector \mathbf{z} given the matrix Φ .

⁷The source codes of those algorithms are obtained from each author's webpage. For CS-BP, we implemented it by applying the uniform sampling-based nBP introduced in Section IV-A.

⁸SuPrEM is only compatible with the LDF matrix which was autonomously proposed in the work [16].

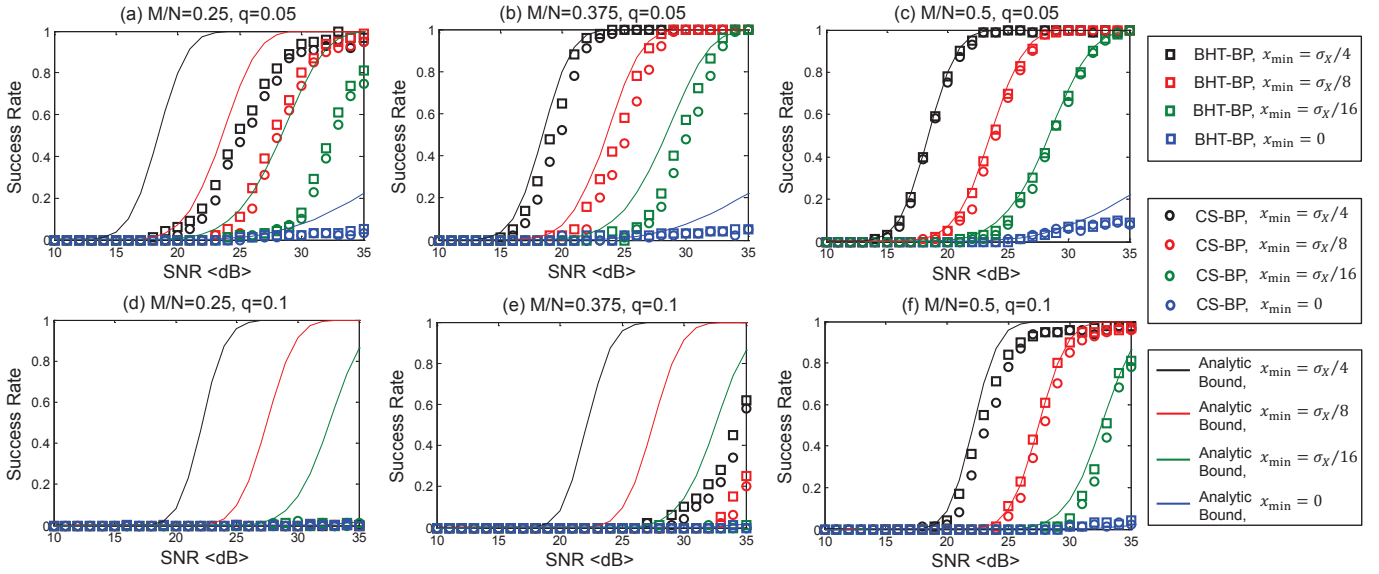


Fig. 6. Experimental result for the success rate of support detection over SNR for a variety of x_{\min} where we consider the case of $N = 1024$, $L = 5$, and $\sigma_X = 5$. In Fig.6, we plot the success rate of of BHT-BP (proposed) and CS-BP [13] together with the analytic bound for the case of $\Phi = \mathbf{I}$, where the nBP part of the both algorithms is implemented with $N_d = 256$, $\varepsilon = 10^{-5}$, $\lambda = 10^{-4}$.

TABLE II
EMPIRICAL THRESHOLD $(M/N)^*$ FOR THE BP CONVERGENCE

Sparsity rate	$N = 512$	$N = 1024$	$N = 4096$
$q = 0.05$	0.325	0.25	0.25
$q = 0.1$	0.575	0.50	0.475

the same column energy, *i.e.* $\mathbb{E} \left[\|\phi_{\text{column}}\|_2^2 \right] = L$; therefore, each entry ϕ_{ji} of the standard Gaussian matrix is drawn from $\mathcal{N}(\phi_{ji}; 0, \frac{L}{M})$. Table I summarizes all the algorithms included in this performance validation.

A. Threshold $(\frac{M}{N})^*$ for BP Convergence,

We claimed the advantage of BHT-BP over CS-BP in support detection with the rate $\frac{M}{N}$ above a certain threshold $(\frac{M}{N})^*$ in Section III-B. Given the rate $\frac{M}{N} \geq (\frac{M}{N})^*$, a BP approximation of the marginal posterior $f_{\text{BP}(X_i)}^{(l)}(x|\mathbf{Z} = \mathbf{z})$ contains sufficiently less uncertainty on the true value $x_{0,i}$. We empirically find the threshold $(\frac{M}{N})^*$ in a noiseless setup using the average information entropy, $\frac{1}{N} \sum_{i=1}^N h(X_i|\mathbf{Z} = \mathbf{z})$ which measures uncertainty of $f_{\text{BP}(X_i)}^{(l)}(x|\mathbf{Z} = \mathbf{z})$. The empirical entropy curves in Fig.5 show sharp phase transition as $\frac{M}{N}$ increases. From the result, we set the threshold to the point achieving $\frac{1}{N} \sum_{i=1}^N h(X_i|\mathbf{Z} = \mathbf{z}) \leq 10^{-3}$, which is given in Table II for a variety of the signal length N and the sparsity rate q . We also note from Fig.5 that the entropy phase transition becomes sharper as N increases.

B. Support Detection Performance over SNR

Fig.6 depicts an experimental comparison of the success rate, defined in (23), between BHT-BP and CS-BP over SNR for a variety of x_{\min} . According to the threshold $(\frac{M}{N})^*$ given in Table II, the BP convergence is achieved only for the cases of (a),(b),(c),(f) in Fig.6. Therefore, we confine our discussion

here to such cases, claiming the advantage of BHT-BP over CS-BP in support detection.

1) *SNR gain by BHT support detection* : The empirical results of Fig.6 validate our claim that BHT-BP has more robust support detection ability against noise, than CS-BP. Indeed, Fig.6 shows that BHT-BP enjoys a remarkable SNR gain from CS-BP in the low SNR regime. This SNR gain is from difference of the detection criterion as discussed in Section III-A. As SNR increases, the success rate of the both algorithms gradually approach to one. In the high SNR regime, BHT-BP and CS-BP do not have notable difference in the performance.

We support the advantage of BHT-BP over CS-BP with Fig.7. This figure depicts an exemplary marginal posterior, obtained from the BP part, according to two different SNR levels, SNR=10 and 30dB, where the true value of X_i is $x_{0,i} = -4.0$; hence $s_i = 1$.

- When SNR is sufficiently high such as the SNR=30 dB case, both of the algorithms can successfully detect the state S_i from the posterior since the probability mass is concentrated on the true value $x_{0,i}$.
- When SNR is low such as the SNR=10 dB case, however, CS-BP may result in misdetection because the point-mass at $x = 0$ is higher than the point-mass at $x_{0,i} = -4.0$ due to the additive noise, leading to $\hat{s}_i = 0$. In contrast, the BHT detector decides the state S_i by incorporating all the spread mass due the noise. This is based on that the likelihoods $f_{\mathbf{z}}(\mathbf{z}|\mathcal{H}_{s_i})$, which construct the hypothesis test of (5), is associated with the entire range of the x -axis rather than a specific point-mass. Therefore, BHT-BP can generate $\hat{s}_i = 1$ and success in the detection even when SNR is low.

2) *Analytic Bound of BHT detection when $\Phi = \mathbf{I}$* : Fig.6 includes an analytic bound of the BHT detection for the case that the measurement matrix is an identity matrix, *i.e.* $\Phi = \mathbf{I}$, such that there is no performance degradation from lack of measurements. Therefore, this bound provides a performance

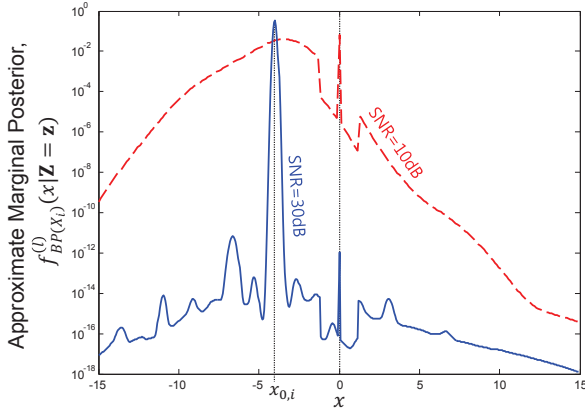


Fig. 7. Example of the approximate marginal posterior $f_{BP(X_i)}^{(l)}(x|\mathbf{Z}=\mathbf{z})$, obtained from the nBP part, for two different SNRs: 10 dB and 30 dB, where the true value of X_i is $x_{0,i} = -4.0$, and the other parameters are set to $M/N = 0.5$, $q = 0.05$, $\sigma_X = 5$, $l^* = 30$, and the minimum value is $x_{min} = \sigma_X/4$.

benchmark of the BHT detection when $\frac{M}{N} \geq (\frac{M}{N})^*$, because exact marginal posteriors are given to the BHT detector under the assumption of $\Phi = \mathbf{I}$. We refer to Appendix II for the detailed derivation of the analytic bound. This derivation reveals that the bound is a function of q, x_{min} , and SNR. In Fig.6, it is clearly shown that the empirical points are fit into the analytic bounds as $\frac{M}{N}$ increases.

3) *Support detection with x_{min}* : Fig.6 also shows the support detection behavior according to x_{min} , confirming that x_{min} is a key parameter in the NSR problem. From Fig.6, we have the observation as given in Note 3.

Note 3 (Empirical observations for x_{min}):

- All the success rate curve shift toward high SNR region as x_{min} decreases.
- Extremely, when $x_{min} = 0$, the experimental points stay near zero even with $\frac{M}{N} \geq (\frac{M}{N})^*$ and high SNR.

These empirical observations intuitively tells us that contribution of x_{min} is as significant as SNR in the NSR problem, implicating that we need $\text{SNR} \rightarrow \infty$ for the perfect support recovery if the signal has $x_{min} \rightarrow 0$. Note that our interpretation on the result here shows good agreement with not only our analytic bound under the assumption of $\Phi = \mathbf{I}$, but also the information-theoretical results [27],[28],[30] showing that support recovery is arbitrarily difficult by sending $x_{min} \rightarrow 0$ even as SNR becomes arbitrarily large.

C. MSE Comparison to Recent Algorithms over SNR

In Fig.8 and Fig.9, we provide an MSE comparison among the algorithms listed in Table I and the support-aware oracle estimator over SNR for a variety of $(\frac{M}{N}, q)$, where MSE^* denotes the performance of the support-aware oracle estimator, given as

$$\text{MSE}^* := \frac{\text{Tr} \left[\left(\frac{1}{\sigma_X^2} \mathbf{I} + \frac{1}{\sigma_W^2} \Phi_s^T \Phi_s \right)^{-1} \right]}{\|\mathbf{x}_0\|_2^2}. \quad (26)$$

In this section, we discuss the comparison result by categorizing the setup of $(\frac{M}{N}, q)$ into two cases: the “region of

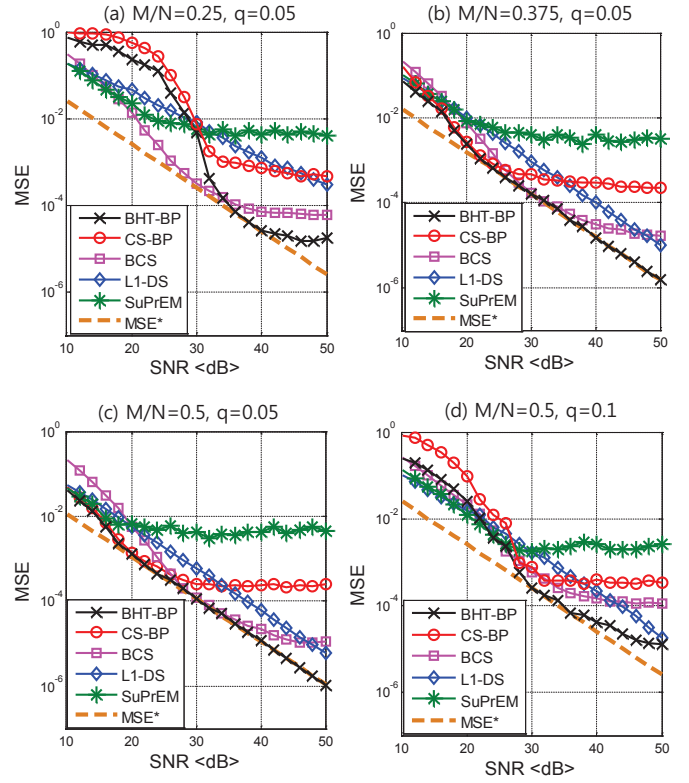


Fig. 8. MSE comparison among the algorithms (BHT-BP (Proposed), CS-BP [13], SuPrEM [16], BCS [12], l_1 -DS [6]) over SNR where we consider signal recovery with $\frac{M}{N} \geq (\frac{M}{N})^*$. We simulate the MSE performance under $N = 1024$, $L = 5$, $\sigma_X = 5$, $x_{min} = \sigma_X/4$. The nBP part embedded in BHT-BP (proposed) and CS-BP is implemented with $N_d = 128$ and $\varepsilon = 10^{-5}$, $\lambda = 10^{-4}$.

$\frac{M}{N} \geq (\frac{M}{N})^*$ ” and the “region of $\frac{M}{N} < (\frac{M}{N})^*$ ” cases, according to the empirical threshold $(\frac{M}{N})^*$ given in Table II, where we fix the parameters $N = 1024$, $L = 5$, $\sigma_X = 5$, $x_{min} = \sigma_X/4$.

1) *MSE performance in region of $\frac{M}{N} \geq (\frac{M}{N})^*$* : With Fig.8, we argue that in the region of $\frac{M}{N} \geq (\frac{M}{N})^*$, BHT-BP catches up with the oracle performance, MSE^* , beyond the SNR point allowing the accurate support finding. Fig.8-(b) and -(c) validate our claim by showing that the BHT-BP curve coincides very closely with the MSE^* curve beyond a certain SNR point. Worth mentioning here is that the SNR point, which starts to achieve the oracle MSE^* , nearly corresponds to the point which attains the perfect support detection with $P_{succ} \approx 1.0$ in Fig.6. For the cases of Fig.8-(a) and -(d), the BHT-BP curve does not fit to the oracle MSE^* at the high SNR region. The reason is coming from lack of measurements for the BP convergence. Indeed, it is observed from Fig.5 that the entropy points corresponding to $(\frac{M}{N}, q)$ of Fig.8-(a) and -(d) is in not a steady region but a transient region. This means that the corresponding posterior includes residual uncertainty on X_i . Although this residual uncertainty does not remarkably work in the low SNR region due to noise effect, it is gradually exposed as SNR increases, degrading the MSE performance in the high SNR region.

In Fig.8, the CS-BP curve forms an error floor as SNR increases, leading to a MSE gap from BHT-BP in the high SNR regime. This MSE gap is mainly caused by the quantization error of the nBP. Since CS-BP obtains its estimate

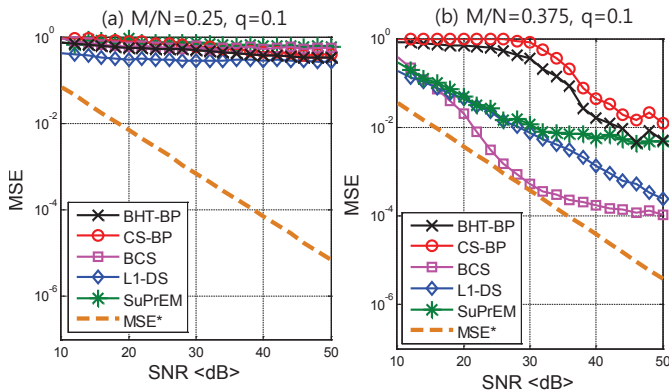


Fig. 9. MSE comparison among the algorithms (BHT-BP (Proposed), CS-BP [13], SuPrEM [16], BCS [12], l_1 -DS [6]) over SNR where we consider signal recovery with $\frac{M}{N} < \left(\frac{M}{N}\right)^*$. We simulate the MSE performance under $N = 1024$, $L = 5$, $\sigma_X = 5$, $x_{min} = \sigma_X/4$. The nBP part embedded in BHT-BP (proposed) and CS-BP is implemented with $N_d = 128$ and $\varepsilon = 10^{-5}$, $\lambda = 10^{-4}$.

directly from the sampled posteriors, the quantization error is unavoidable, leading to an error floor. The level of the floor can be approximately predicted by the MSE degradation of the quantization, given as

$$\frac{\mathbb{E} \|Q_{T_s}[\mathbf{X}_S] - \mathbf{X}_S\|_2^2}{\mathbb{E} \|\mathbf{X}_S\|_2^2} = \frac{T_s^2/12}{\sigma_X^2} = \frac{3}{N_d^2} \quad (27)$$

where $Q_{T_s}[\cdot]$ is the quantization function with the step size T_s given in (14), and \mathbf{X}_S is a random vector on the signal support \mathbf{S} . Under our joint detection-and-estimation structure, the LMMSE estimator (13) enables BHT-BP to go beyond the error floor.

For SuPrEM, the performance is poor to the other algorithms in the experimental results of Fig.8. But, it is not surprising since SuPrEM is basically for signals having fixed signal sparsity K .⁹ Indeed, the SuPrEM algorithm requires the sparsity K as an input parameter. However, in many cases, the signal sparsity K is unknown and random. In our basic setup, recall that we assumed signals having Binomial random sparsity, *i.e.*, $K \sim \mathcal{B}(k; N, q)$. Therefore, naturally SuPrEM underperforms the other algorithms in this experiment. l_1 -DS and BCS are comparable to BHT-BP, but l_1 -DS has a certain SNR loss from the BHT-BP over all range of SNR, and BCS shows an error floor at high SNR region.

2) *MSE performance in region of $\frac{M}{N} < \left(\frac{M}{N}\right)^*$* : In Fig.9, we investigate the MSE comparison in the region of $\frac{M}{N} < \left(\frac{M}{N}\right)^*$. Under the setup of $\frac{M}{N} = 0.25$, $q = 0.1$, every algorithm generally does not work as shown in Fig.9-(a). From the setup of $\frac{M}{N} = 0.375$, $q = 0.1$, all the algorithms begin to find signals but, BHT-BP underperforms BCS, l_1 -DS, and SuPrEM in this setup, as shown in Fig.9-(b). The reason is that in the region of $\frac{M}{N} < \left(\frac{M}{N}\right)^*$, the BP does not converge properly due lack of the measurements such that probability mass on the true value $x_{0,i}$ is not dominant in the approximate marginal posteriors, as we discussed in Section III-B. From the results, we conclude that BHT-BP is not advantageous

⁹We empirically confirmed that when K is fixed, SuPrEM works as comparable to BHT-BP even though we does not include that result in this paper.

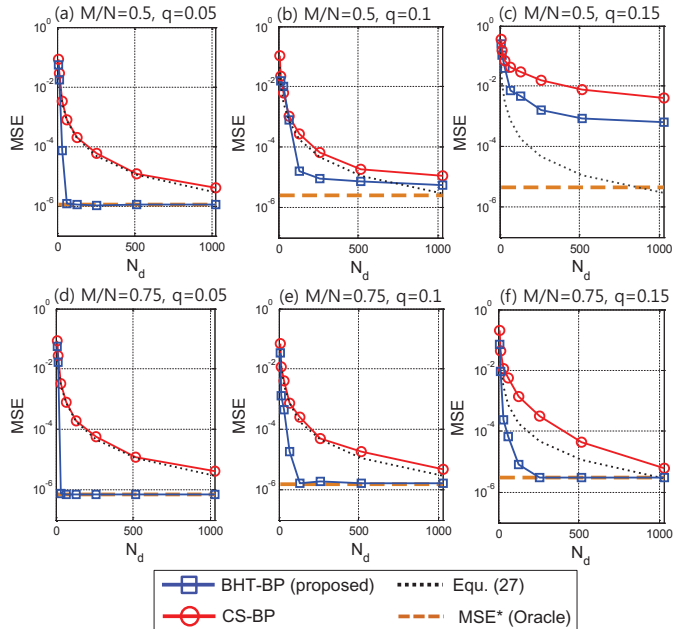


Fig. 10. MSE comparison of BHT-BP (proposed) and CS-BP [13] with clean measurements (SNR=50dB) over $N_d \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$ for a variety of $\frac{M}{N}$, q , where we plot the MSE curves together with the MSE degradation by the quantization error, given in (27). In this experiment, we consider a case of $N = 1024$, $L = 5$, $x_{min} = \sigma_X/4$, $\varepsilon = 10^{-5}$, $\lambda = 10^{-4}$.

over the other algorithms excluding CS-BP when sufficient measurements is not maintained for the signal sparsity.

D. Empirical Calibration of BHT-BP over N_d and L

1) *Number of samples N_d for nBP*: From the discussion in Section IV-C, one can argue that the complexity of BHT-BP is highly sensitive to the number of samples N_d ; therefore, BHT-BP cannot be low-computational in a certain case. It is true, but we claim that the effect of N_d is limited in the BHT-BP recovery. To support our claim, Fig.10 compares MSE performance of BHT-BP, CS-BP, and the support-aware oracle estimator as a function of N_d in a clean setup (SNR=50 dB) where we plot the MSE curves together with the MSE degradation by the quantization error, given in (27). From Fig.10, we confirm that BHT-BP can achieve the oracle performance if N_d is beyond a certain level and $\left(\frac{M}{N}, q\right)$ belongs to the success phase, whereas CS-BP cannot provide the oracle performance even as N_d increases. Consequently, N_d does not significantly contribute to the MSE of the BHT-BP recovery once N_d exceeds a certain level. This result implies that the complexity of the BHT-BP recovery can be steady with a constant N_d in practice. Therefore, BHT-BP can holds the low-computational property given by the BP philosophy. In addition, we confirm from Fig.10 that the MSE of CS-BP is bounded by (27).

2) *Column weight L of LDPC-like matrices*: Another interesting question is how to determine the column weight L of the LDPC-like matrix Φ for BHT-BP. Fig.11 provides an answer for this question by showing the MSE of the BHT-BP recovery as a function of L , where we consider the recovery from clean measurements (SNR= 50 dB). When $\frac{M}{N}$ is sufficiently large, for example $\frac{M}{N} = 0.75$ as shown in Fig.11-(b), the BHT-BP recovery generally becomes accurate

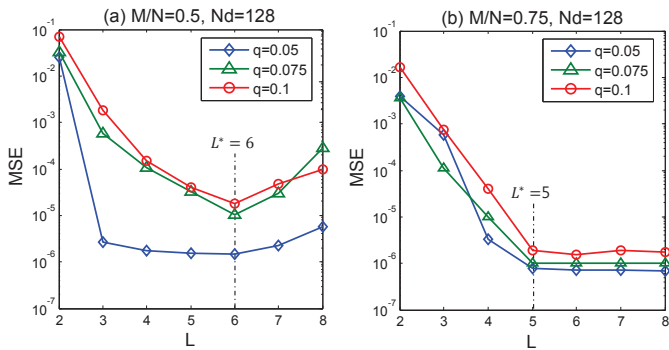


Fig. 11. MSE performance of the BHT-BP recovery over the column weight L of the measurement matrix Φ where we set $N = 1024$, $N_d = 128$, $x_{min} = \sigma_X/4$, $\varepsilon = 10^{-5}$, $\lambda = 10^{-4}$ and consider the recovery from clean measurements (SNR= 50dB).

as L increases. Then, the accuracy is almost constant after a certain point $L = L^*$. On the other hands, when $\frac{M}{N}$ is not sufficient, for example $\frac{M}{N} = 0.5$ as shown in Fig.11-(a), the recovery accuracy rather can be degraded beyond a certain point L^* . The reason is that when $\frac{M}{N}$ is small, the large L spoils the tree-structured property of the matrix Φ , reducing the accuracy of the marginal posterior approximation by BP [25],[26]. Therefore, L should keep as small as possible once the desirable recovery accuracy is achieved. In the case of Fig.11, we empirically set $L^* = 6, 5$ for $\frac{M}{N} = 0.5, 0.75$ respectively.

From the calibration shown in Fig.10 and Fig.11, we support our claim that the computational cost of BHT-BP can be $\mathcal{O}(N \log N + KM)$ in practice by fixing L and N_d , as we discussed in Section IV-C.

VI. CONCLUSION

The theoretical and empirical research in this paper demonstrated that BHT-BP is powerful as not only a low-computational solver, but also a noise-robust solver. In BHT-BP, we employed a joint detection-and-estimation structure consisting of the BHT support detection and the LMMSE estimation for the nonzeros on the signal support. We have shown that the BHT-BP detects the signal support based on a sequence of binary hypothesis tests, which is related to the criterion of the minimum detection error probability. This support detection approach brings SNR gain of BHT-BP from CS-BP [13], which is an existing nBP-based algorithm, for the support detection. In addition, we noted the fact that BHT-BP effectively removes the quantization error of the nBP approach in the signal recovery. We have claimed that our joint detection-and-estimation strategy prevents from degrading the MSE by the quantization error. We have supported the claim based on an empirical result that the performance of BHT-BP achieves the oracle performance when sufficient measurements is maintained for the signal sparsity. Furthermore, we confirm the impact of x_{min} on the noisy sparse recovery (NSR) problem via BHT-BP. Based on the empirical evidence, we showed that exact sparse recovery with small x_{min} is very demanding unless sufficiently large SNR is provided, which is an agreement with the result of [27],[28],[30] that emphasizes the importance of x_{min} in the NSR problem.

APPENDIX I BRIEF INTRODUCTION TO RECENT BAYESIAN ALGORITHMS

In this appendix, we provide a brief introduction to some previously proposed Bayesian algorithms for the NSR problem: BCS [12], CS-BP [13], SuPrEM [16]. These algorithms have been developed by applying several types of signal prior PDFs and statistical techniques. These algorithms are included for simulation based comparison in Section V.

A. BCS Algorithm

Ji *et al.* proposed a Bayesian algorithm based on the *sparse Bayesian learning* (SBL) framework, called BCS [12]. In the SBL framework, a two-layer hierarchical Gaussian model has been invoked for signal estimation. Namely, the signal prior PDF takes the form of

$$f_{\mathbf{x}}(\mathbf{x}|a, b) = \prod_{i=1}^N \int_0^{\infty} \mathcal{N}(x_i; 0, \gamma_i^{-1}) f_{\Gamma}(\gamma_i|a, b) d\gamma_i, \quad (28)$$

where $f_{\Gamma}(\gamma_i|a, b)$ is a hyper-prior following the Gamma distribution with its parameters a, b . Then, the MAP estimate $\hat{\mathbf{x}}_0$ of the signal can be analytically expressed as a function of the hyperparameter $\Gamma = [\gamma_1, \dots, \gamma_N]$, the measurement matrix Φ , and the noisy measurements \mathbf{z} .

In BCS, the hyperparameter Γ is estimated by performing a type-II *maximum likelihood* (ML) procedure [11]. Specifically, the type-II ML finds the hyperparameter Γ maximizing the evidence PDF, *i.e.*, $f_{\mathbf{Y}}(\mathbf{y}|\Gamma) = \int f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\Gamma) d\mathbf{x}$. The expectation maximization (EM) algorithm can be an efficient approach for the type-II ML procedure. The strategy of EM is to derive a lower bound on the log evidence PDF, $\log f_{\mathbf{Y}}(\mathbf{y}|\Gamma)$, at the E-step, and optimize that lower bound to find Γ at the M-step. The E-step and M-step are iterated until the lower bound becomes tighter.

The BCS algorithm is input parameter-free, which means this algorithm is adaptive to any types of signals and noise level since BCS properly catches the hyperparameter γ and the noise variance σ_W^2 during the recovery. In addition, the BCS algorithm is well compatible with any type of the measurement matrices.

B. CS-BP Algorithm

Baron *et al.* for the first time proposed the use of BP to the sparse recovery problem with LDPC-like measurement matrices [13]. The algorithm is called CS-BP. Signal model of CS-BP is a compressible signal which has a small number of large elements and a large number of near-zero elements. The authors associated this signal model with a two-state mixture Gaussian prior, given as

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^N [q\mathcal{N}(x_i; 0, \sigma_{X_1}^2) + (1-q)\mathcal{N}(x_i; 0, \sigma_{X_0}^2)], \quad (29)$$

where $q \in [0, 1)$ denotes the probability that an element has the large value, and $\sigma_{X_1} \gg \sigma_{X_0}$. Therefore, the prior is fully parameterized with σ_{X_0} , σ_{X_1} , and q . CS-BP performs MAP or MMSE estimation using marginal posteriors obtained from BP similarly to the proposed algorithm, where the authors applied both nBP and pBP approaches for the BP implementation. The

recovery performance is not very good when measurement noise is severe since the CS-BP was basically designed to work under noiseless setup.

C. SuPrEM Algorithm

Most recently, Akcakaya *et al.* proposed SuPrEM under a framework similar to BCS which uses the two-layer hierarchical Gaussian model for the signal prior. SuPrEM was developed under the use of a specific type of hyper-prior called the Jeffreys' prior $f_{\mathcal{J}}(\tau_i) = 1/\beta_i, \beta_i \in [T_i, \infty] \forall i \in \mathcal{V}$. This hyper-prior reduces the number of input parameters while sparsifying the signal. The overall signal prior PDF is given as

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^N \int_0^{\infty} \mathcal{N}(x_i; 0, \beta_i) f_{\mathcal{J}}(\beta_i) d\beta_i. \quad (30)$$

SuPrEM utilizes the EM algorithm to find each hyperparameter β_i like the BCS algorithm. However, differently from BCS that calculates the signal estimate $\hat{\mathbf{x}}_0$ using matrix operations which include matrix inversion, SuPrEM elementwisely calculates the signal estimate from β_i via a pBP algorithm. Therefore, SuPrEM can be more computationally efficient than BCS.

The measurement matrix used in SuPrEM is restricted to an LDPC-like matrix which has fixed column and row weights, called low-density-frames (LDF). They are reminiscent of the regular LDPC codes [24]. In addition, the signal model is confined to K -sparse signals consisting of K nonzeros and $N - K$ zeros since SuPrEM includes a sparsifying step which chooses the K largest elements at each end of iteration. The noise variance σ_W^2 is an optional input to the algorithm. Naturally, if the noise variance is provided, SuPrEM will produce an improved recovery performance.

APPENDIX II

SUCCESS RATE ANALYSIS OF THE BHT DETECTION WHEN $\Phi = \mathbf{I}$

Under the assumption of $\Phi = \mathbf{I}$, the measurement channel can be decoupled to N scalar Gaussian channels which are $Z_j = X_i + W_j \forall i, j \in \mathcal{V}$ where clearly $i = j$ holds. Accordingly, the success rate, given in (23), can be represented as the product of the complementary probability of the state error rate (SER) given in (7), *i.e.*, $P_{\text{succ}} = (1 - P_{\text{SER}})^N$. Then, the problem is reduced to the analysis of the rate P_{SER} (see Fig.4). The conditional SER given the hypothesis \mathcal{H}_{s_i} is calculated from the likelihood PDF $f_{Z_j}(z|\mathcal{H}_{s_i})$ as following:

$$P_{\text{SER}|\mathcal{H}_{s_i}} := \Pr\{\hat{s}_i \neq s_i | \mathcal{H}_{s_i}\} = \int_{\bar{D}_{\mathcal{H}_{s_i}}} f_{Z_j}(z|\mathcal{H}_{s_i}) dz, \quad (31)$$

where we define the decision regions with a threshold γ' as

$$D_{\mathcal{H}_0} := \{|z| < \gamma'\} \text{ and } D_{\mathcal{H}_1} := \{|z| \geq \gamma'\}, \quad (32)$$

and $\bar{D}_{\mathcal{H}_0} = D_{\mathcal{H}_1}$ vice versa. The likelihood PDFs can be obtained from

$$f_{Z_j}(z|\mathcal{H}_{s_i}) = \int f_{Z_j}(z|X_i = x) f_{X_i}(x|\mathcal{H}_{s_i}) dx \quad (33)$$

as we have done in (5), where $f_{Z_j}(z|X_i = x) = \mathcal{N}(z; x, \sigma_W^2)$ under the scalar Gaussian channel. Then, the likelihood given

\mathcal{H}_0 simply becomes $f_{Z_j}(z|\mathcal{H}_0) = \mathcal{N}(z; 0, \sigma_W^2)$. In contrast, the likelihood conditioning \mathcal{H}_1 is not straightforward due to the dented slab part of our prior in (3), which is given by

$$\begin{aligned} f_{Z_j}(z|\mathcal{H}_1) &\propto \int_{|x| \geq x_{\min}} \mathcal{N}(z; x, \sigma_W^2) \mathcal{N}(x; 0, \sigma_X^2) dx \quad (34) \\ &+ \lambda \int_{|x| < x_{\min}} \mathcal{N}(z; x, \sigma_W^2) dx \\ &= \mathcal{N}(z; 0, \sigma_W^2 + \sigma_X^2) \left(1 - \frac{1}{2} \operatorname{erf} \left(\frac{A(z)}{\sqrt{2}} \right) - \frac{1}{2} \operatorname{erf} \left(\frac{B(z)}{\sqrt{2}} \right) \right) \\ &+ \frac{\lambda}{2} \left(\operatorname{erf} \left(\frac{x_{\min} - z}{\sigma_W \sqrt{2}} \right) + \operatorname{erf} \left(\frac{x_{\min} + z}{\sigma_W \sqrt{2}} \right) \right) \end{aligned}$$

where normalization is required to satisfy $\int f_{Z_j}(z|\mathcal{H}_1) dz = 1$, and the functions $A(z), B(z)$ are respectively described as

$$A(z) := \frac{x_{\min} \left(\frac{1}{\sigma_W^2} + \frac{1}{\sigma_X^2} \right) - \frac{z}{\sigma_W}}{\sqrt{\frac{1}{\sigma_W^2} + \frac{1}{\sigma_X^2}}}, \quad B(z) := \frac{x_{\min} \left(\frac{1}{\sigma_W^2} + \frac{1}{\sigma_X^2} \right) + \frac{z}{\sigma_W}}{\sqrt{\frac{1}{\sigma_W^2} + \frac{1}{\sigma_X^2}}}.$$

In this problem, an analytical expression of γ' is unattainable from the equality condition (9) since the PDF $f_{Z_j}(z|\mathcal{H}_1)$ involves the error function terms as shown in (34). Therefore, we utilize a root-finding algorithm to compute γ' . We use the SNR definition given in (25) such that $\text{SNR} = 10 \log_{10} \frac{qL\sigma_X^2}{\sigma_W^2}$ under the assumption of $\Phi = \mathbf{I}$. We specify the decision regions (32) with γ' , finalizing this analysis by computing the condition SERs, which are given as

$$P_{\text{SER}|\mathcal{H}_0} = 1 - \operatorname{erf} \left(\frac{\gamma'}{\sigma_W \sqrt{2}} \right), \quad (35)$$

$$P_{\text{SER}|\mathcal{H}_1} = 2 \int_0^{\gamma'} f_{Z_j}(z|\mathcal{H}_1) dz \quad (36)$$

where the calculation of $P_{\text{SER}|\mathcal{H}_1}$ requires a numerical integration owing to the error function terms in $f_{Z_j}(z|\mathcal{H}_1)$. Using (7), (35), and (36), we can evaluate the SER, then obtaining the success rate of the BHT detection when $\Phi = \mathbf{I}$. We compare this analysis result to the empirical results in Section V-B.

REFERENCES

- [1] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, issue 6, pp. 1182-1195, Dec. 2007.
- [2] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83-91, Mar. 2008.
- [3] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Devices and Systems*, vol. 5, issue 1, pp. 8-20, Jan. 2011.
- [4] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 6-18, Jan. 2006.
- [5] J. A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1030-1051, 2006.
- [6] E. Candes and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313-2351, 2007.
- [7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., Ser. B*, vol. 58, no. 1, pp. 267-288, 1996.
- [8] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655-4666, Dec. 2007.
- [9] D. Needell, J. Tropp, "COSAMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. and Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301-321, 2008.

- [10] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34-81, Feb. 2009.
- [11] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211-244, 2001.
- [12] Shihao Ji, Ya Xue, and Lawrence Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346-2356, June 2008.
- [13] D. Baron, S. Sarvotham, and R. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 269-280, Jan. 2010.
- [14] J. Kang, H.-N. Lee, and K. Kim, "Bayesian hypothesis test for sparse support recovery using belief propagation," *Proc. of IEEE Statistical Signal Processing Workshop (SSP)*, pp. 45-48, Aug. 2012.
- [15] X. Tan and J. Li, "Computationally efficient sparse Bayesian learning via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2010-2021, Apr. 2010.
- [16] M. Akcakaya, J. Park, and V. Tarokh, "A coding theory approach to noisy compressive sensing using low density frame," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5369-5379, Nov. 2011.
- [17] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, pp. 18914-18919, Nov. 2009.
- [18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 498-519, Feb. 2001.
- [19] E. Sudderth, A. Ihler, W. Freeman, and A. S. Willsky, "Nonparametric belief propagation," *Communi. of the ACM* vol 53, no. 10, pp. 95-103, Oct. 2010.
- [20] J. M. Coughlan and S. J. Ferreira, "Finding deformable shapes using loopy belief propagation," *Proc. of 12th Euro. Conf. on Comp. Vision (ECCV)*, pp. 453-468, 2002.
- [21] M. Isard, J. MacCormick, and K. Achan, "Continuously-adaptive discretization for message-passing algorithms," *Proc. of the Adv. in Neural Inform. Process. Sys. (NIPS)*, 2009.
- [22] N. Noorshams, and M. J. Wainwright, "Quantized stochastic belief propagation: efficient message-passing for continuous state spaces," *Proc. of IEEE Int. Symp. Inform. Theory (ISIT)*, pp. 1246-1250, July, 2012.
- [23] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE Journal on Sel. Areas in Communi.*, vol 23, no.4, pp. 809-819, Apr. 2005.
- [24] R. G. Gallager, *Low-Density Parity Check Codes*, MIT Press: Cambridge, MA, 1963.
- [25] D. J. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003, Available from www.inference.phy.cam.ac.uk/mackay/itila/
- [26] T. Richardson, and R. Urbanke, "The capacity of low-density parity check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 599-618, Feb. 2001.
- [27] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inform. Theory*, vol. 56, no. 6, pp. 2967-2979, Jun. 2010.
- [28] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5728-5741, Dec. 2009.
- [29] M. Akcakaya and V. Tarokh, "Shannon-theoretic limit on noisy compressive sampling," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 492-504, Jan. 2010.
- [30] A. Fletcher, S. Rangan, and V. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5758-5772, Dec. 2009.
- [31] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signal in noise," *IEEE Trans. Inform. Theory*, vol. 14, no. 3, pp. 434-444, May. 1968.
- [32] S. Kay, *Fundamentals of Statistical Signal Processing Volume I: Detection Theory*, Prentice Hall PTR, 1993.
- [33] S. Kay, *Fundamentals of Statistical Signal Processing Volume II: Estimation Theory*, Prentice Hall PTR, 1993.
- [34] Ake Bjorck, *Numerical Methods for Least Squares Problems*, SIAM: PA, 1996.
- [35] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, May. 1994.
- [36] H. Ishwaran and J. S. Rao, "Spike and slab variable selection : Frequentist and Bayesian strategies," *Ann. Statist.*, vol.33, pp. 730-773, 2005.
- [37] D. Guo and S. Verdu, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 1983-2010, Jun. 2005.
- [38] D. Guo and C. C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Comm.*, vol. 26, no. 3, pp. 421-431, Mar. 2008.
- [39] Frey, B. J. and D. J. MacKay, "A revolution: Belief propagation in graphs with cycles," *Proc. of the 11th Annual Conference on Neural Inform. Proces. Sys., (NIPS)*, pp. 479-485, Dec. 1997.
- [40] K.P. Murphy, Y. Weiss, and M.I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," *Proc. of the 5th conf. on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 467-475, 1999.
- [41] G. Moustakides, G. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: optimum tests and applications," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4215-4229, June 2012.
- [42] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer: NY, 2006.



signal processing.



include information theory, signal processing, communications/networking theory, and their application to wireless communications and networking, compressive sensing, future internet, and brain-computer interface.



wideband digital communications system design, sensor network design, analysis and implementation both, at the physical layer and at the resource management layer.

Jaewook Kang (M'14) received the B.S. degree in information and communications engineering (2009) from Konkuk University, Seoul, Republic of Korea, and the M.S. degree in information and communications engineering (2010) from the Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea. He is currently pursuing the Ph.D. degree in information and communications engineering at the GIST. His research interests lie in the board areas of compressed sensing, machine learning, wireless communications, and statistical

Heung-No Lee (SM'13) received the B.S., M.S., and Ph.D. degrees from the University of California, Los Angeles, CA, USA, in 1993, 1994, and 1999, respectively, all in electrical engineering. From 1999 to 2002, he was with HRL Laboratories, LLC, Malibu, CA, USA, as a Research Staff Member from 1999 to 2002. From 2002 to 2008, he was with the University of Pittsburgh, Pittsburgh, PA, USA, as an Assistant Professor. He joined Gwangju Institute of Science and Technology (GIST), Korea, where he is currently a Professor. His general areas of research

Kiseon Kim (SM'98) received the B.Eng. and M.Eng. degrees, in electronics engineering, from Seoul National University, Korea, in 1978 and 1980, and the Ph.D. degree in electrical engineering systems from University of Southern California, Los Angeles, in 1987. From 1988 to 1991, he was with Schlumberger, Houston, Texas. From 1991 to 1994, he was with the Superconducting Super Collider Lab, Texas. He joined Gwangju Institute of Science and Technology (GIST), Korea, in 1994, where he is currently a Professor. His current interests include

Controlling Band Gap and Refractive Index in Dopant-Free α -Fe₂O₃ Films

Pawan Kumar,¹ Nitin Rawat,² Da-Ren Hang,³ Heung-No Lee,² and Rajesh Kumar^{1,2,*}

¹Jaypee University of Information Technology, Wagnaghat, Solan 173234, Himachal Pradesh, India

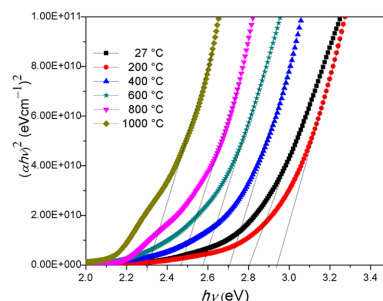
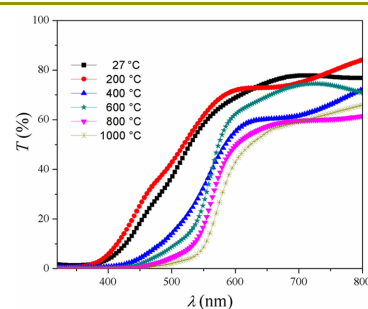
²Gwangju Institute of Science and Technology, Gwangju 500712, Korea

³Department of Materials and Optoelectronic Science, National Sun Yat-sen University, Kaohsiung 804, Taiwan

(received date: 3 January 2014 / accepted date: 17 July 2014 / published date: 10 January 2015)

Dopant-free hematite (α -Fe₂O₃) films are formed at a liquid-vapor interface by means of an easy method in order to control the band gap and refractive index of the films. The α -Fe₂O₃ films after being transferred to a glass substrate are studied for their structural and optical properties. Control over the thickness of the films in the range from 75 to 400 nm and the constituent nanocrystallite size from 3 to 46 nm is achieved by controlling the synthesis parameters. By controlling the film thickness, crystallite size, and crystallinity of dopant-free α -Fe₂O₃ films, the optical band gap is increased significantly (by ≈ 0.64 eV) from 2.30 to 2.94 eV, along with increase in the refractive index from 1.35 to 2.8. The observed increase in the optical band gap is explained on the basis of change in lattice symmetry (via change in the c/a ratio) of α -Fe₂O₃ crystallites.

Keywords: oxide materials, optical materials, α -Fe₂O₃ thin films, band gap, and optical properties



1. INTRODUCTION

The band gap and band-edge positions of semiconductors are of importance in photoelectrochemical and photocatalytic applications.^[1] Iron oxide, particularly α -Fe₂O₃ has several advantages over other semiconductor materials when used to realize devices with an optical band gap of approximately 2.00 eV. It possesses excellent chemical stability over a broad range of pH values, an absorption spectrum in the wavelength region between 600 and 295 nm,^[2] is abundantly available in the earth's crust, and is inexpensive and non toxic.^[3] This makes α -Fe₂O₃ an attractive candidate for photoelectrochemical [PEC] water splitting,^[4] optical limiting,^[5] and optoelectronic applications.^[6] Most of these applications require a tunable optical band gap for improved performance, e.g., an optical band gap of around 2.46 eV is necessary for

water photocatalysis while using α -Fe₂O₃ without the application of any bias voltage.^[7] In this light, realizing a blue shift in the band gap of hematite by an energy of about 0.3 to 0.6 eV can make hematite an ideal anode material for photocatalytic oxidation of water as well.^[1,7]

In applications wherein the optical band gap of α -Fe₂O₃ requires to be greater than 2.00 eV, control of the crystallite size/thickness can enable tuning of the optical band gap. Similar to the optical band gap, the refractive index of materials is also an important factor in several optical designs/applications.^[8,9] The performance of many solid state devices such as integrated optical emissive displays, optical sensors, integrated optical circuits, and light-emitting diodes can be improved by applying a high refractive index film/coating on the light emitting/sensing portion of the devices.^[10-15] In fact, both the optical band gap and refractive index depend upon the crystallite size and thickness of the film.

In the backdrop of controlling crystallite size, nanostructures

*Corresponding author: rajesh.kumar@juit.ac.in
©KIM and Springer

of α -Fe₂O₃ have been synthesized by numerous methods,^[16-23] but the practical application of these methods has been restricted because of the high cost of synthesis equipment, limitations in achieving a large surface area, and uniform deposition of film.^[24-26] Thus, a facile and cost effective method providing easy tuning of the optical band gap and refractive index of α -Fe₂O₃ film is highly desirable. The techniques reported earlier^[27-29] for band gap engineering requires doping of other elements or the fabrication of nanocomposites, which are disadvantageous in terms of stability and cost effectiveness. Previously, we have reported a novel technique for the synthesis of undoped α -Fe₂O₃ films on the surface of a precursor solution at low temperature.^[30] Here, the same method is adopted for the tuning of the optical properties of the band gap and refractive index of undoped α -Fe₂O₃ films. The optical properties of the films depend upon their thickness and crystallite size, and this method enables easy control over the thickness and crystallite size of the film and thus on the optical properties. In the present study, unlike the case of the quantum confinement effect on the band gap,^[1,29] we observed that the variation in the optical band gap of film is dependent upon the change in lattice symmetry caused by lattice modification. When compared with reported band gap values,^[31-33] a larger variation in the optical band gap of undoped α -Fe₂O₃ film is observed, which is attributed to the small crystallite size and partial amorphous nature of the film. The variation in refractive index is explained in terms of the packing density

of α -Fe₂O₃ films, which is easily controlled by the synthesis parameters.

2. EXPERIMENTAL PROCEDURE

Floating films of α -Fe₂O₃ were formed on a liquid-vapor interface. A mixed solution containing 24.0 mM of FeCl₂ (purity 99.99%, Sigma Aldrich) and 22.0 mM of FeCl₃·6H₂O (purity 99.99%, Sigma Aldrich) was used as the precursor solution.

The floating films were transferred to glass substrates that were annealed in a horizontal tube furnace in presence of argon gas. The variation in optical properties of the films was studied with the following variations in the synthesis parameters, i.e., (i) dose (vol. %) of NH₃, (ii) concentration of polyvinyl alcohol (PVA), and (iii) the annealing temperature. The dose of NH₃ vapor was varied from 2% (40 cm³) to 4% (80 cm³) and then to 6% (120 cm³) at a fixed (32 μ M) concentration of PVA. The concentration of PVA was varied from 8 to 32 and then to 80 μ M for a fixed dose of NH₃ at 6% (120 cm³). The films obtained in these two sets of experiments were annealed at 500°C. In the third set, the films formed for a fixed concentration (32 μ M) of PVA and a fixed dose of NH₃ (6% (120 cm³)) were annealed at 200°C, 400°C, 600°C, 800°C, and 1000°C. These films were characterized for a study of their structural and optical properties. The structural properties were examined using an x-ray diffractometer (XRD, PANalytical's X'Pert-PRO) and

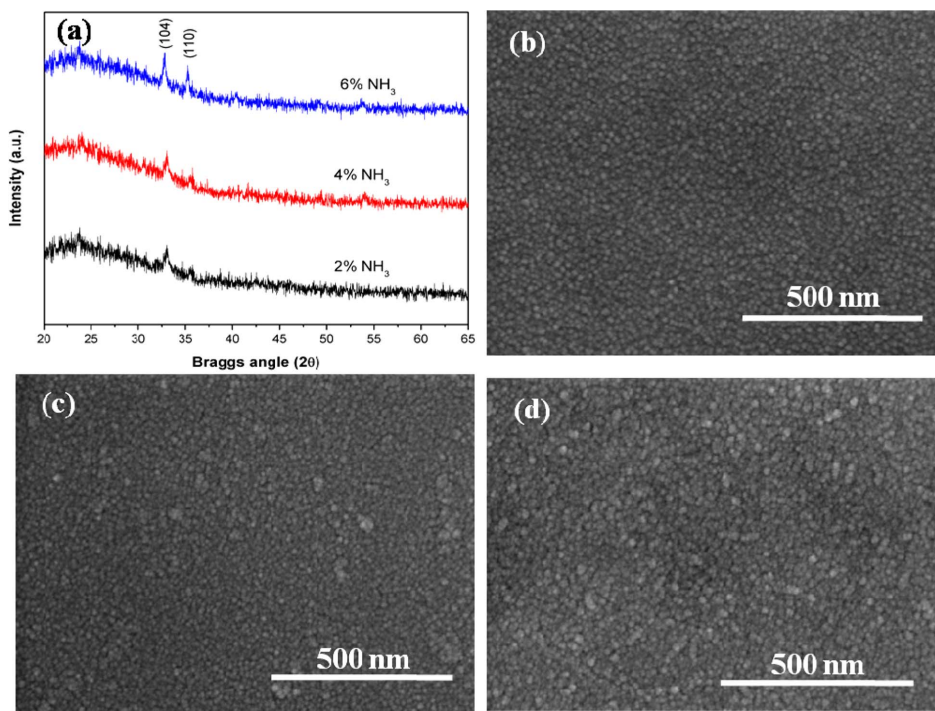


Fig. 1. (a) XRD patterns of α -Fe₂O₃ films obtained with 2%, 4%, and 6% NH₃ doses. The films were annealed at 500°C. (b), (c), and (d) SEM images of films formed with 2, 4, and 6% NH₃ doses, respectively.

a transmission electron microscope (TEM, JEOL, JEM 2100), the morphological properties were studied via a scanning electron microscope (SEM, Hitachi, S-4700), and the film thickness was examined using a Stylus profilometer. The optical properties were studied using a UV-Vis-NIR double-beam spectrophotometer (Perkin Elmer Lambda-750) in the 250 - 900 nm wavelength range.

3. RESULTS AND DISCUSSION

3.1 Variation in optical properties with NH₃ dosage

For all three sets of experiments, we analyzed the morphological and structural changes in the films. We first discuss the films formed under the condition when the NH₃ dose was varied. Figure 1(a) shows the XRD images of films formed for 2%, 4%, and 6% doses of NH₃. The films were investigated using XRD with a Cu K α (1.54 Å) source and scanning angles ranging from 20° to 65° with a step size of 0.01 at room temperature. The XRD plot shows diffraction peaks corresponding to α -Fe₂O₃ (according to JCPDS-ICCD PDF card No. 33-0664). Crystalline peaks around 32.4° and 35.4° correspond to the (104) and (110) planes of α -Fe₂O₃, thereby indicating its hexagonal (corundum-type) structure. The XRD shows that the intensity of the crystalline peaks increases with NH₃ dose which may be due to increasing thickness of the α -Fe₂O₃ film as obtained by profilometer data (Table 1). The increase in thickness of the α -Fe₂O₃ film with increasing doses of NH₃ is due to the presence of a large number of NH₃ molecule within the reaction chamber that

react with a large number of precursor ions (Fe³⁺/Fe²⁺) on the solution surface, thereby resulting increased film thickness. The average crystallite size (D) in α -Fe₂O₃ films is estimated using Scherrer's formula,^[34] $D = 0.9\lambda/\beta \cos\theta$, where β denotes the full width at half maximum and λ the wavelength of x-rays. The average crystallite sizes with the corresponding lattice parameters are listed in Table 1. As the film thickness increases, the size of crystallites in the film also increases as can be observed from Table 1, and this behavior is in accordance with other reports.^[35] Figures 1(b) to 1(d) show the SEM images of α -Fe₂O₃ films prepared with 2, 4, and 6% NH₃ doses respectively. The increasing thickness of the film with increase in the NH₃ dosage gives rise to clustering of α -Fe₂O₃ particles, which leads to an increase in the film's roughness as indicated in SEM images (Figs. 1(b) to 1(d)).

As regards the optical properties, a UV-Vis-NIR spectrophotometer was used to observe the variation in the optical band gap and refractive index of α -Fe₂O₃ films. The obtained transmission (T) spectra with respect to variation in the NH₃ dosage are shown in Fig. 2(a). There is a decrease in the transmission of α -Fe₂O₃ films with increase in NH₃ dosage. This decrease in transmission is attributed to increase in the size of the clustered nanocrystals and the thickness of the film. Due to clustering of nanocrystallites, the increased roughness of the films enhances the scattering of light and a consequent reduced transmittance.^[36] From the transmission spectra, the optical absorption coefficient α was calculated using^[37] $\alpha = (1/t) \ln (1/T)$, where t denotes the thickness of the film. Further, the optical band gap (E_g) was calculated

Table 1. Values of NH₃ dosage, PVA concentration, average thickness (t), crystallite size (D), lattice parameters ($a = b, c$), optical band gap energy (E_g), refractive index (n), and relative density (ρ_f/ρ_b) for α -Fe₂O₃ films.

Dose of NH ₃	PVA Concentration	Thickness (t) (nm)	D (nm) from XRD	$a = b$ (Å)	c (Å)	c/a (Å)	E_g (eV)	n (at 589 nm)	ρ_f/ρ_b
2%	32 μ M	75	14.24 \pm 0.81	5.06	13.92	2.75098	2.72	1.35	0.265
4%	32 μ M	155	15.05 \pm 1.01	5.05	13.88	2.74851	2.56	1.54	0.379
6%	32 μ M	350	19.70 \pm 2.30	5.05	13.86	2.74455	2.35	2.32	0.692

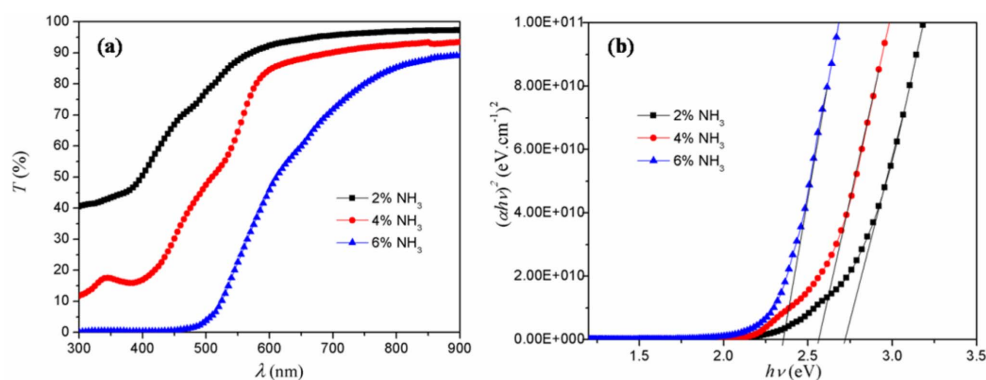


Fig. 2. (a) Transmission (T) spectra of α -Fe₂O₃ film obtained with 2, 4, and 6% doses of NH₃ and (b) plots of $(\alpha hv)^2$ vs hv for these α -Fe₂O₃ films.

using the Tauc relation^[38] $\alpha h\nu = C_1(h\nu - E_g)^n$, where C_1 denotes a constant, h the Planck's constant, and the prefix $n = 0.5$ for a direct optical band gap transition.

The calculated optical direct band gap values were 2.72, 2.56, and 2.35 eV, respectively, for 2%, 4%, and 6% doses of NH_3 , and these results indicate that the optical band gap decreases with increase in film thickness. We attribute that the variation in the optical band gap to (i) stress-induced distortion of the optical band gap by film/substrate interactions, (ii) density of dislocation, (iii) quantum size effect, (iv) change in grain boundary barrier height due to change in crystallite size in the polycrystalline film,^[39] and (v) change in lattice symmetry.^[31] In our case, as all the films were prepared under similar synthesis conditions on similar substrates, factors (i) and (ii) may be ignored. The quantum confinement effect is mostly observed in crystallites with sizes less than 6 nm (for $\alpha\text{-Fe}_2\text{O}_3$ crystallites).^[1,29,40] The barrier height depends upon the crystallite size D according to the expression^[41] $E_b = E_{bo} + C(X - fD)^2$, where the original barrier height E_{bo} , constant C , barrier width X , and f are specific to the materials. In our case, the variation in the crystallite size is negligible (~ 14 to 19 nm), and therefore, we speculate that the change in barrier height is also negligible in its contribution to the change in the band gap.

Lattice modification has been reported to affect the electronic energy levels of $\alpha\text{-Fe}_2\text{O}_3$ nanocrystals.^[31] A decrease in the size of $\alpha\text{-Fe}_2\text{O}_3$ nanocrystallites is reported to be equivalent to the application of negative pressure, which is expected to lower the lattice symmetry owing to the anisotropic nature of the $\alpha\text{-Fe}_2\text{O}_3$ lattice with a consequent increase in the axial ratios c/a , as can be observed from the values listed in Table 1.^[31] We note that size-induced lattice modification (c/a) yields distinct electronic (or magnetic) properties of $\alpha\text{-Fe}_2\text{O}_3$ nanocrystals.^[31] An increase in the c/a ratio results in an increase in ionicity and Fe-O bond separation during the anisotropic expansion of smaller size crystallite. The most intense absorption peak of $\alpha\text{-Fe}_2\text{O}_3$ ^[31,42] is given by the expression $E = -10Dq + 10B + 6C - 26B^2/$

$10Dq$, where $10Dq$ denotes the crystal field splitting, and B and C the Racah parameters that describe the neighboring covalency effect in a transition metal system.^[31] The second-order term ($-26B^2/10Dq$) is extremely small compared to the sum of the terms $10B$ and $6C$ according to the estimated ligand field theory parameters.^[31] Since the Racah parameters B and C increase with decrease in nanocrystallite size under low pressure,^[42] the observed blue shift (band gap change) in the absorption peak of the $\alpha\text{-Fe}_2\text{O}_3$ film with reduced crystallite size is likely the consequence of increase in the magnitude of the Racah parameters.

To calculate the refractive index of $\alpha\text{-Fe}_2\text{O}_3$ films, the reflectance was determined by using the expression^[42] $R = 1 - [T \exp(A)]^{1/2}$, where A denotes the absorption of the film. Finally, the refractive index (n) of the films was calculated using the approximation^[43,44] $n = [(1 + R) / (1 - R)] + [((4R) / (1 + R)^2) - (k)^2]^{1/2}$, where k denotes the extinction coefficient related to the absorption coefficient (α) as $k = \alpha\lambda/4\pi$. We observed that at a particular wavelength, the refractive index of the film increases with the NH_3 doses, as shown in Fig. 3. The increase in refractive index with increasing film thickness can be attributed to an increase in the packing density of the film that is concurrent with increase in the film thickness. As the film thickness increases, its porosity decreases,^[45] thereby resulting in increased refractive index of the film. The increased size of the crystallites in the film increases its density due to the reduced crystallite boundaries^[46-48] and consequently, this contributes to increase in the refractive index. The film density was calculated by using the Lorentz-Lorenz relation,^[49] $\rho_f/\rho_b = [(n_f^2 - 1)(n_b^2 + 2)] / [(n_f^2 + 2)(n_b^2 - 1)]$, where ρ_f denotes the density of the $\alpha\text{-Fe}_2\text{O}_3$ film, ρ_b the density of bulk $\alpha\text{-Fe}_2\text{O}_3$, n_f the refractive index of the film, and n_b the refractive index of bulk material ($n_b = 3.003$ at $\lambda = 633$ nm).^[36] For n_f values of 1.31, 1.47, and 2.01 corresponding to films formed with 2%, 4%, and 6% doses of NH_3 , respectively, the calculated relative densities (ρ_f/ρ_b) are listed in Table 1. The results indicate that with increase in the NH_3 dose, the thickness as well as the size of nanocrystallites in the $\alpha\text{-Fe}_2\text{O}_3$ film increases, which results in an increase in the packing density and refractive index of the film.

3.2 Variation in optical properties with PVA concentration

The XRD patterns of the films obtained with various PVA concentration values are shown in Fig. 4(a). Here, the XRD peak intensity decreases with increasing PVA concentration. This decrease in the peak intensity is due to decrease in the crystalline nature of film via the PVA capping effect.^[50] Based on calculations from the XRD data (Table 2), we obtain the crystallite sizes for PVA concentrations of 8, 32, and 80 μM as 26.80, 14.6, and 12.26 nm, respectively. The SEM images in Figs. 4(b) to 4(d) also exhibit a change in the

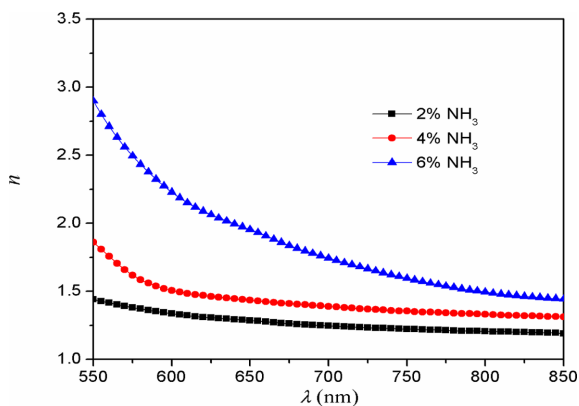


Fig. 3. Refractive index (n) vs wavelength (λ) plots of $\alpha\text{-Fe}_2\text{O}_3$ films obtained with 2%, 4%, and 6% doses of NH_3 .

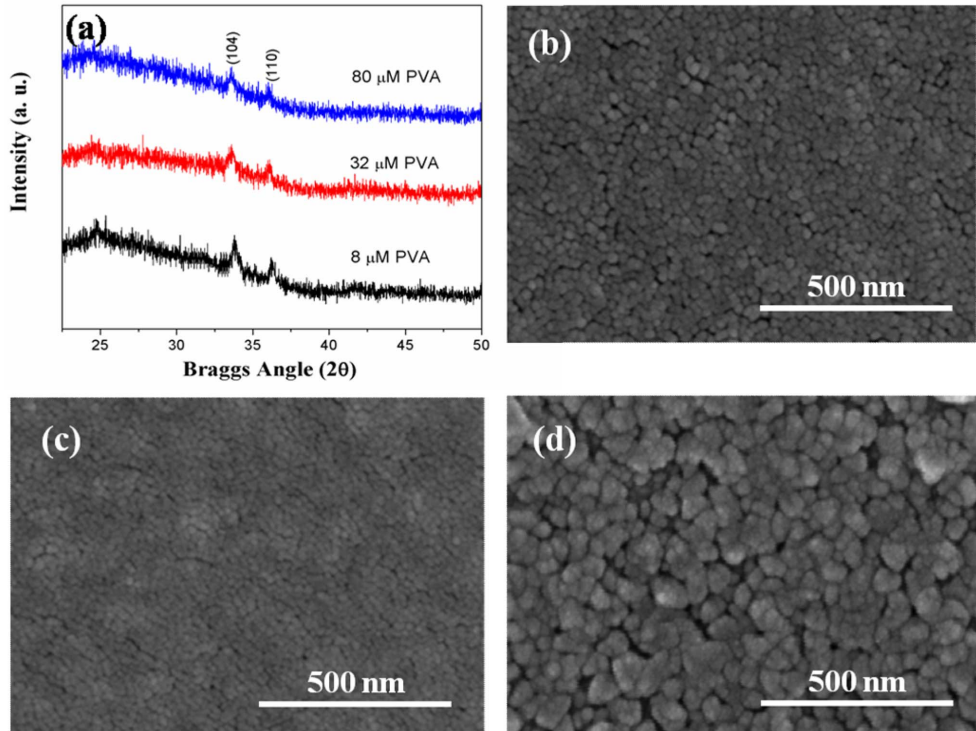


Fig. 4. (a) XRD patterns and (b), (c), and (d) SEM images of α -Fe₂O₃ films formed at 8, 32, and 80 μ M PVA concentrations, respectively.

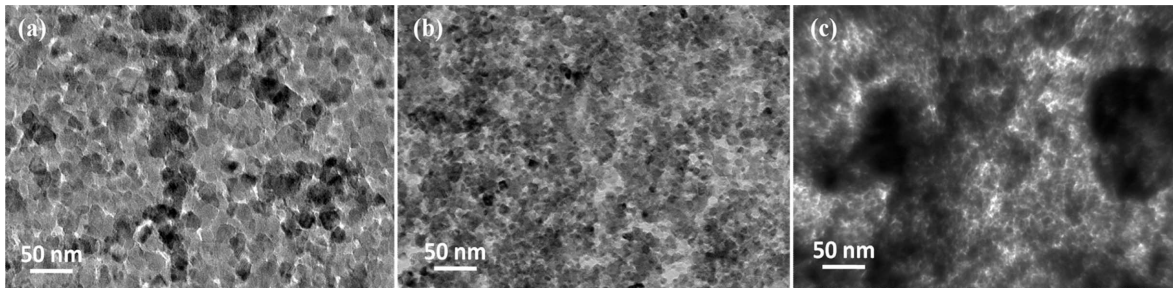


Fig. 5. TEM images of α -Fe₂O₃ films formed at 6% dose of NH₃ with (a) 8, (b) 32, and (c) 80 μ M PVA concentrations.

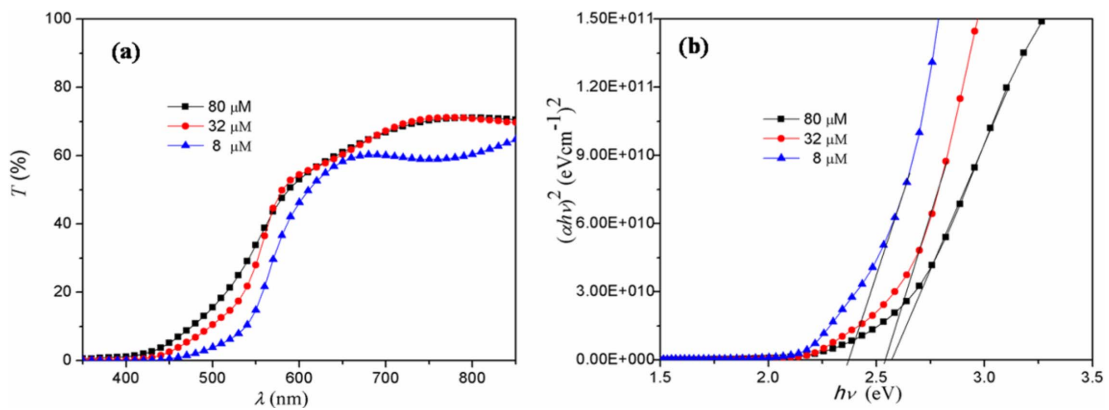


Fig. 6. (a) Transmission (T) spectra of α -Fe₂O₃ films formed at 8, 32, and 80 μ M PVA concentrations, and (b) $(\alpha h\nu)^2$ vs $h\nu$ plot of the films.

Table 2. Values of PVA concentration, NH₃ dosage, average thickness (t), crystallite size (D), lattice parameters ($a = b, c$), optical band gap energy (E_g), refractive index (n), and relative density (ρ_f/ρ_b) for α -Fe₂O₃ films.

PVA concentration	Dose of NH ₃	Thickness (t) (nm)	D (nm) from XRD	$a=b$ (Å)	c (Å)	c/a (Å)	E_g (eV)	n (at 589 nm)	ρ_f/ρ_b
8 μ M	6%	398	26.80 \pm 2.82	4.98	13.55	2.72088	2.37	2.30	0.6857
32 μ M	6%	401	14.60 \pm 1.10	4.99	13.66	2.73747	2.54	2.13	0.6637
80 μ M	6%	396	12.26 \pm 0.87	5.02	13.78	2.74502	2.57	2.09	0.5323

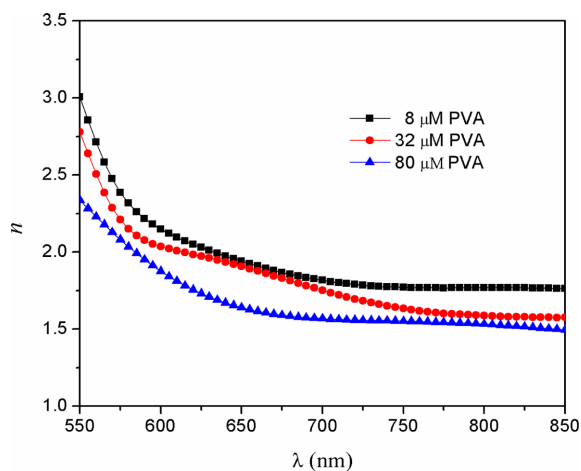


Fig. 7. Refractive index (n) vs wavelength (λ) plots of α -Fe₂O₃ films formed at 8, 32, and 80 μ M PVA concentrations.

morphology of the films with increasing concentration of PVA. Figure 4(d) shows an aggregation of small nanoparticles, which is confirmed by the XRD data and the TEM image in Fig. 5(c). From the TEM image, it is observed that the larger nanoparticles are aggregations of smaller nanoparticles, as reported in a previous study.^[30] Figures 5(a) to 5(c) show TEM images corresponding PVA concentrations of 8, 32, and 80 μ M, respectively. It is clear from the TEM images that small nanocrystallites aggregate with increasing PVA concentration. We conclude that for a particular dose of NH₃, increasing the PVA concentration results in a decrease in the nanocrystallite size (although they are aggregated).

As regards optical properties, the α -Fe₂O₃ films show increased transmission with increasing PVA concentration from 8 to 32 μ M, as shown in Fig. 6(a). The increased transmission with increasing PVA concentration is due to reduction in the crystallinity of the films. The crystallinity of the film increases as the crystallites size increases because the increased crystallite size results in the reduction of nanocrystallites boundaries due to coalition of small crystallites.^[48] However in our case, the situation is opposite; as the concentration of PVA is increased, the crystallite size decreases, and hence, decreased crystallinity leads to increased transmission. On the other hand, we observed that for PVA concentrations ranging from 32 to 80 μ M, the transmission remains unchanged (Fig. 6(a)). This may be

due to increase in transmission being counteracted by increase in light scattering. An increase in light scattering is expected due to increasing roughness caused by the aggregation of small nanocrystallites with increase in PVA concentration. The blue shift in the transmission spectra (Fig. 6(a)) with increasing concentration of PVA indicates an increasing optical band gap in the α -Fe₂O₃ films. Figure 6(b) shows the increase in the optical band gap from 2.37 to 2.54 and then to 2.57 eV corresponding to PVA concentrations of 8, 32, and 80 μ M. In this case as well the band gap variation can be explained on the basis of change in the lattice symmetry in a manner similar to the case of NH₃. The parameters related to change in PVA concentration are listed in Table 2.

Next, we examine the change in the refractive index with increasing PVA concentration. The refractive index decreases with increasing PVA concentration, as shown in Fig. 7, which is again due to variation in the density of the α -Fe₂O₃ films. In this case, the trend is opposite to that observed in case when ammonia dosage is increased, i.e., the density of the film decreases (Table 2) with increasing PVA concentration unlike the case of increasing NH₃ dosage.

The density of the films decreases due to the increasing porosity of α -Fe₂O₃ nanocrystals in the film with increasing concentration of PVA.^[51] The PVA molecules that are flexible penetrate the voids between clusters of α -Fe₂O₃ nano crystallites, and when the films are annealed, the PVA molecule evaporate leaving large voids within the α -Fe₂O₃ nanocrystallites, thereby making them mesoporous. This increase in the porosity (decrease in packing density of α -Fe₂O₃ films) with increasing PVA concentration results in a decrease in the refractive index of the films. From the application point of view, these mesoporous α -Fe₂O₃ nanostructures are highly desirable in many applications such as lithium-ion batteries^[52] gas sensors,^[53] and photochemical^[54] and photoelectrochemical applications.^[54]

In the above mentioned set of experiments, we observed that the synthesis parameters i.e., dosage of NH₃ and PVA concentration, significantly affect the optical properties of α -Fe₂O₃ films. We also observed that annealing temperature is also an important factor for the tuning of the optical properties of the films,^[55] and therefore we examined the combined effect of annealing temperature along with variation in these synthesis parameters in our third set of

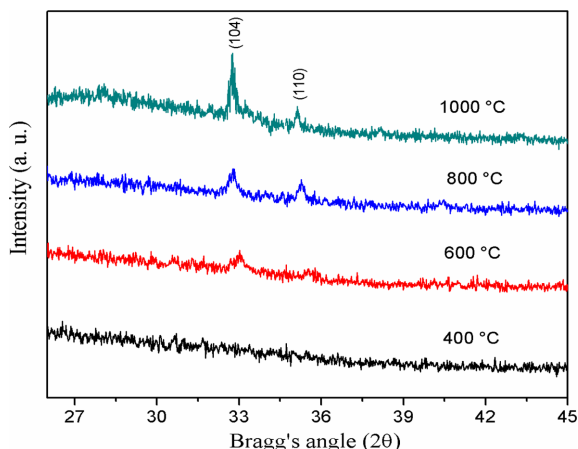


Fig. 8. XRD patterns of α -Fe₂O₃ films annealed at 400°C, 600°C, 800°C, and 1000°C.

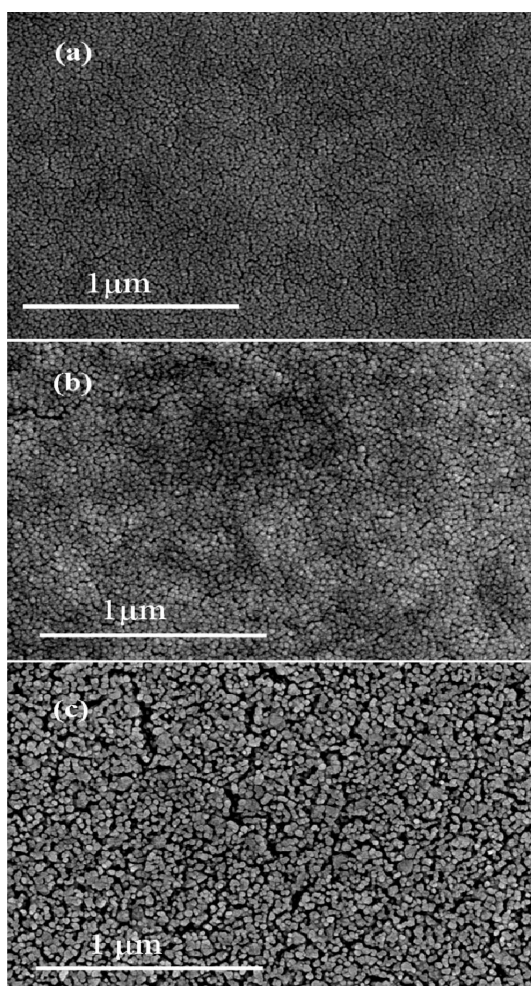


Fig. 9. SEM images of α -Fe₂O₃ films annealed at (a) 600°C, (b) 800°C, and (c) 1000°C.

experiments. As regards this set of experiments, we fixed the NH₃ and PVA concentrations and varied only the annealing

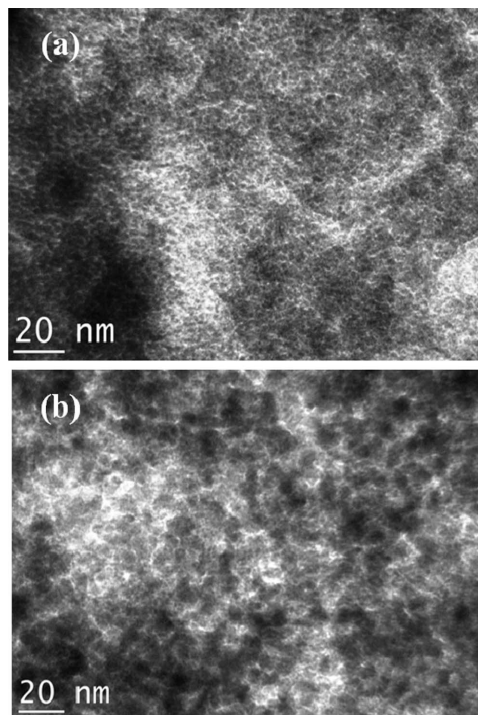


Fig. 10. TEM images of α -Fe₂O₃ films annealed at (a) 200°C, and (b) 400°C.

temperature, as discussed in following section.

3.3 Variation in optical properties with annealing temperature

To study the effect of annealing temperature on the films, we selected an α -Fe₂O₃ film formed at 6% dosage of NH₃ and a PVA concentration of 32 μM. The α -Fe₂O₃ films were annealed in an argon environment at 200°C, 400°C, 600°C, 800°C, and 1000°C. The films were characterized for structural and optical properties as in the previous cases. We observed that unheated films and those annealed at 200°C, and 400°C exhibited no XRD peak. This is probably due to the amorphous nature of the films below 400°C. Figure 8 shows the XRD patterns of films annealed at and above 400°C. Here, only the films annealed above 400°C exhibit crystalline XRD peaks. Further, our calculation from XRD data indicate that for the film annealed at 600°C, the average crystallite size is 24 nm, and this size increased to 31 nm and then to 46 nm for annealing temperatures of 800°C and 1000°C. The variation in the nanocrystallites sizes of the samples annealed at 600°C, 800°C, and 1000°C can also be observed in the SEM images shown in Fig. 9.

Since no XRD peaks were observed for films annealed at 200°C and 400°C, in order to estimate the crystallite sizes in these films, the corresponding TEM images (Fig. 10) were processed by using Image J software package. These samples exhibited nanocrystallite sizes of approximately

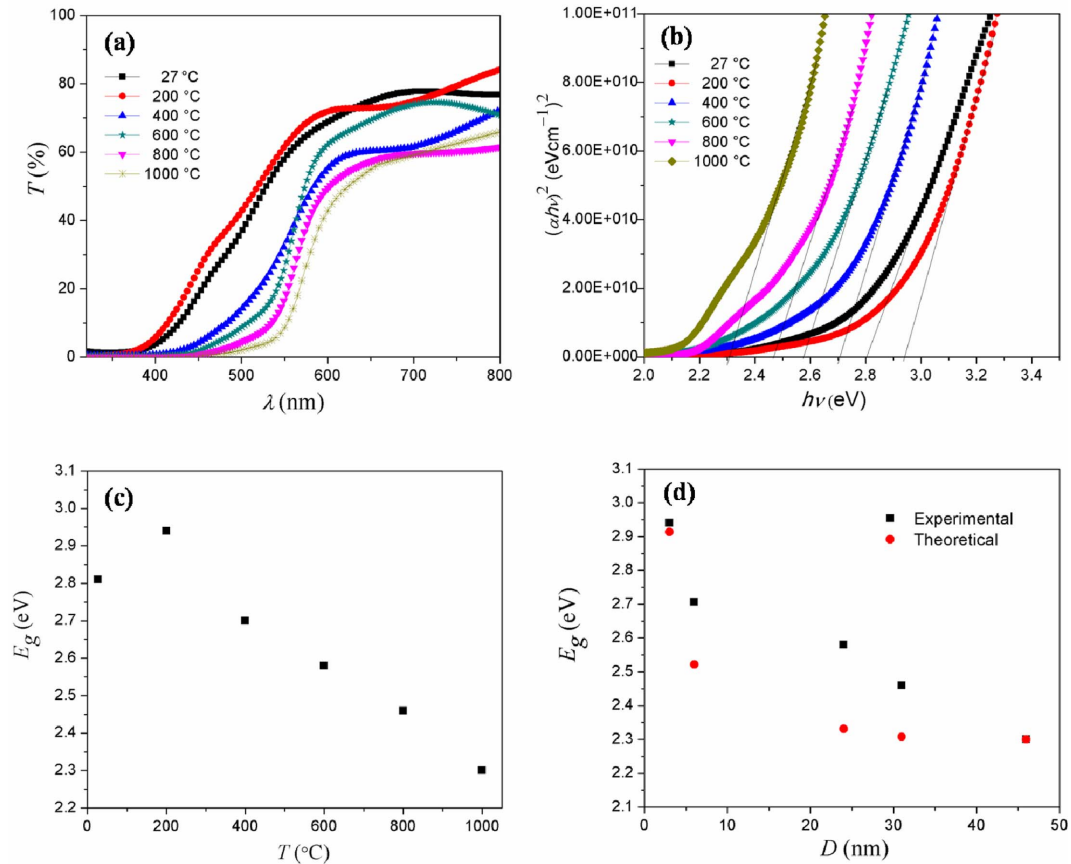


Fig. 11. (a) Transmission spectra of α -Fe₂O₃ films annealed at 200°C, 400°C, 600°C, 800°C, and 1000°C, (b) plots of $(\alpha hn)^2$ vs hn of α -Fe₂O₃ films, (c) variation in optical band gap (E_g) with annealing temperature, and (d) optical band gap vs crystallite size for experimental and theoretical values.

3 nm (200°C) and 6 nm (400°C). The increase in crystallite size with increasing annealing temperature indicates that the crystalline particle size in the film can be varied by varying either the dosage of NH₃, concentration of PVA, or annealing temperature.

Figure 11(a) shows the transmission spectra of these films. With increase in annealing temperature, the transmission decreases and a red shift is observed. The decrease in the transmission of α -Fe₂O₃ films with increased annealing temperature is due to the increasing crystallite size with increasing temperature and increasing roughness caused by the formation of large nanocrystallites that increase scattering.^[56,57] The density of the crystallite boundaries in the film decreases due to the increasing crystallite size (crystallinity) as well as reflection, which enhances the absorption, thereby leading to reduced transmission.^[48,58-60] However, the film in the SEM image (in Fig. 9(c)), appears to be porous when compared with the films shown in Figs. 9(a) and 9(b); nevertheless, the film simultaneously (Fig. 11(a)) exhibits decreased transmission, which indicates that the porosity may exist only at the surface of the film and the overall porosity of the film does not affect the transmittance

as much the crystallinity of the film does.

The optical band gaps for films annealed at 200°C, 400°C, 600°C, 800°C, and 1000°C temperatures are obtained as 2.94, 2.70, 2.58, 2.46, and 2.30 eV, respectively. The maximum band gap of 2.94 eV is obtained for the sample annealed at 200°C, which band gap value decreases with increasing annealing temperature, as shown in Figs. 11(b) and 11(c).

As regards our experiments in varying the annealing temperature, we can classify α -Fe₂O₃ films into two categories: films that exhibit the quantum size effect as they have crystallite sizes less than 5 or 6 nm and those that do not exhibit the quantum size effect as they have relatively larger crystallite sizes (their blue shift is due to only the change in lattice symmetry). The crystallite-size dependence of the optical band gap due to quantum confinement is expressed by the equation^[31] $E_g = E_g^o + n^2 \hbar^2 \pi^2 / 2\mu R^2 - 1.8 e^2 / \epsilon R$, where E_g^o can be assumed as the lowest value of the band gap^[29] obtained in our experiment, R denotes the size of the crystallite, e the electronic charge, ϵ the dielectric constant, and μ the effective electron and hole masses. It is known that smaller crystallites in the film exhibit a larger optical band

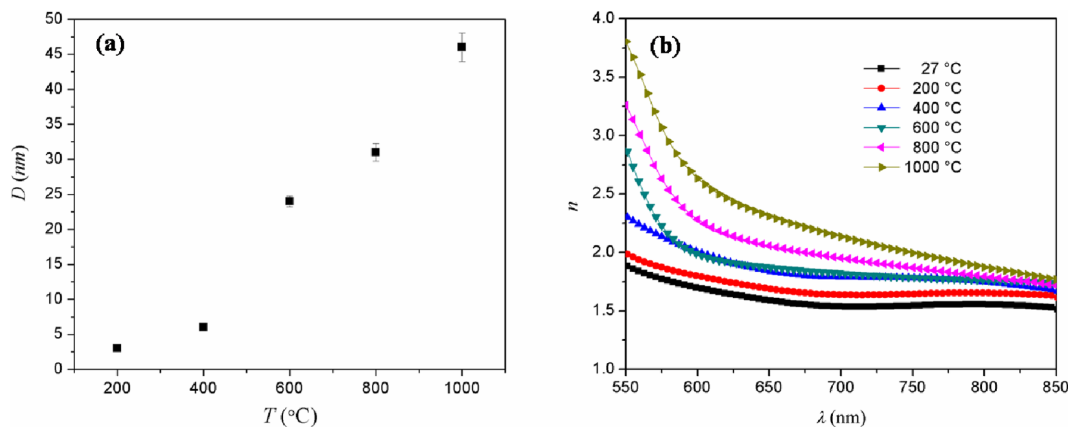


Fig. 12. (a) Plot of crystallite size (D) vs annealing temperature and (b) variation in refractive index (n) as a function of wavelength (λ) of the unheated and annealed α -Fe₂O₃ films at different temperatures.

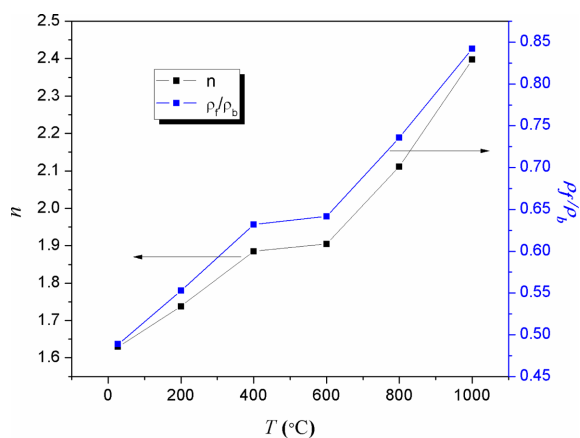


Fig. 13. Variation in refractive index (n) and relative density (ρ_f/ρ_b) with annealing temperature.

gap due to the quantum size effect and therefore, we observed a blue shift in the 3 nm crystallite film (Fig. 11(d)).^[29] In Fig. 11(d), we observe that the experimental value of the band gap for the 3 nm crystallite size coincides with the theoretical value, thereby indicating that this film exhibits the quantum size effect. On increasing the crystalline size above 3 nm by increasing the annealing temperature, a deviation between experimental and theoretical values is observed, as shown in Fig. 11(d). The reason for this observed deviation is speculated to be due to the partial amorphous nature^[61] of α -Fe₂O₃ films along with the change in the lattice symmetry of α -Fe₂O₃ crystallites. The resultant absorption of photons is due to both the amorphous and nanocrystalline phases of α -Fe₂O₃ particles, and hence, the absorption edges in the experimental results exhibit a higher blue shift than the theoretical values.

As regards the second category of α -Fe₂O₃ films, other studies have also reported variations in the band gap with change in the annealing temperature due to change in lattice

symmetry.^[32] In fact, the phase sharing of the octahedral dimer and the electrostatic repulsion of the Fe³⁺ cation are responsible for the trigonal distortion of the octahedron, thereby giving rise to C_{3v}-type symmetry.^[33] With appropriate thermal treatment, the crystallite size of the α -Fe₂O₃ films increases and the structure relaxes to maximize the distance between two iron cations in Fe₂O₉ dimers.^[32] As the annealing temperature is increased, the average crystallite size increases, and hence, the optical band gap decreases (Fig. 11(c)). The variation in the crystallite size in α -Fe₂O₃ films with annealing temperatures is shown in Fig. 12(a). Here, the calculated c/a ratios corresponding to α -Fe₂O₃ films annealed at 600°C, 800°C, and 1000°C are 2.749, 2.739, and 2.732, respectively. As the c/a ratio decreases with increasing annealing temperature, the films exhibit structural relaxation, which leads to decrease in the optical band gap.

Finally, the variation in the refractive index (1.7 to 2.8 at 589 nm) of these films formed with increasing annealing temperature is shown in Fig. 12(b). As expected, the α -Fe₂O₃ films show an increase in the refractive index with annealing temperature.^[62,63] The variation in the refractive index with annealing temperature can be correlated with the packing density of the films as in the previous cases. From Fig. 13, we observe that the films annealed at lower temperatures have lower packing densities than those annealed at higher temperatures.

The lower packing density at lower annealing temperatures is due to the incorporation of oxygen during film growth,^[36] which creates voids on annealing. As the annealing temperature increases, the increase in thermal energy facilitates the coalition of small crystallites, which increases the packing density of α -Fe₂O₃ films due to reduction in the number of voids.^[64,65] In conclusion, we note that our method facilitates greater control over the tuning of the optical properties of α -Fe₂O₃ films by varying either one, two or all three process parameters, i.e., NH₃ dosage, PVA concentration,

and annealing temperature.

4. CONCLUSIONS

We tailored the structural and optical properties of α -Fe₂O₃ films formed on the surface of a precursor solution. In our method the parameter of NH₃ dosage can be used to easily control the thickness of a floating α -Fe₂O₃ film on the surface of a precursor solution, and the PVA concentration in the precursor solution can be used to control the size of nanocrystallites composing the film. Lattice modification due to the change in lattice symmetry with the α -Fe₂O₃ crystallite size is speculated as the reason for the observed shift in the band gap. Further, the refractive index also changes due to change in the packing density of α -Fe₂O₃ films. The post-synthesis annealing temperature can be varied to control the size of the resultant crystalline particles, which can be utilized to further tune the optical band gap and refractive index of α -Fe₂O₃ films. Our method can significantly affect the optical band gap without the use of any dopant, and therefore, the α -Fe₂O₃ films obtained using our method are suitable for hydrogen generation from water via photocatalysis without the application of a bias voltage.

ACKNOWLEDGEMENTS

This study was supported by a Research Grant for Nanotechnology Lab, Jaypee University of Information Technology, Waknaghat, Solan (India) and the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MEST) (No. 2012-0005656).

REFERENCES

1. L. Vayssieres, C. Sathe, S. M. Butorin, D. K. Shuh, J. Nordgren, and J. Guo, *Adv. Mater.* **17**, 2320 (2005).
2. M. Chirita and I. Grozescu, *Chem. Bull. "POLTEHNICA" Univ. (Timisoara)* **54**, 1 (2009).
3. R. Al-Gaashani, S. Radiman, N. Tabet, and A. R. Daud, *J. Alloy. Compd.* **550**, 395 (2013).
4. S. Shen, C. X. Kronawitter, J. Jiang, S. S. Mao, and L. Guo, *Nano Res.* **5**, 327 (2012).
5. H. A. Garcia, R. P. de Melo Jr., A. Azevedo, and C. B. de Araujo, *Appl. Phys. B* **111**, 313 (2013).
6. T. Hayakawa and M. Nogami, *Sci. Technol. Adv. Matter.* **6**, 66 (2005).
7. Y. Matsumoto, *J. Solid State Chem.* **126**, 227 (1996).
8. C. Liu and B. Yang, *J. Mater. Chem.* **19**, 2884 (2009).
9. S. D. Hart, G. R. Maskaly, B. Temelkuan, P. H. Prideaux, J. D. Joannopoulos, and Y. Fink, *Science* **296**, 510 (2002).
10. B. G. Prevo, Y. Hwang, and O. D. Velev, *Chem. Mater.* **17**, 3642 (2005).
11. N. Nishimuura, Y. Shibasaki, M. Ozawa, and Y. Oishi, *J. Photopolym. Sci. Technol.* **25**, 355 (2012).
12. A. Kasikov, J. Aarik, H. Mändar, M. Moppel, M. Pärs, and T. Uustare, *J. Phys. D: Appl. Phys.* **39**, 54 (2006).
13. S. Wu, G. Zhou, and M. Gu, *Opt. Mater.* **29**, 1793 (2007).
14. X. Wei, H. Shi, X. Dong, Y. Lu, and C. Du, *Appl. Phys. Lett.* **97**, 011904 (2010).
15. J. G. Liu and M. Ueda, *J. Mater. Chem.* **19**, 8907 (2009).
16. L. Zhenmin, L. Xiaoyong, W. Hong, M. Dan, X. Chaojian, and W. Dan, *Nanotechnology* **20**, 245603 (2009).
17. A. Duret and M. Gratzel, *J. Phys. Chem. B* **109**, 17184 (2005).
18. N. T. Hahn, H. C. Ye, D. W. Flaherty, A. J. Bard, and C. B. Mullins, *ACS Nano* **4**, 1977 (2010).
19. T. J. LaTempa, X. J. Feng, M. Paulose, and C. A. Grimes, *J. Phys. Chem. C* **113**, 16293 (2009).
20. A. Kay, I. Cesar, and M. Gratzel, *J. Am. Chem. Soc.* **128**, 15174 (2006).
21. K. Woo, H. J. Lee, J. P. Ahn, and Y. S. Park, *Adv. Mater.* **15**, 176 (2003).
22. S. K. Mohapatra, S. E. John, S. Banerjee, and M. Mishra, *Chem. Mater.* **21**, 3048 (2009).
23. P. M. Rao and X. L. Zheng, *Nano Lett.* **9**, 3001 (2009).
24. G. Wang, Y. Ling, D. A. Wheeler, K. E. N. George, K. Horsley, C. Heske, J. Z. Zhang, and Y. Li, *Nano Lett.* **11**, 3503 (2011).
25. D. A. Wheeler, G. Wang, Y. Ling, Y. Li, and J. Z. Zhang, *Energy Environ. Sci.* **5**, 6682 (2012).
26. H. G. Cha, J. Song, H. S. Kim, W. Shin, K. B. Yoon, and Y. S. Kang, *Chem. Commun.* **47**, 2441 (2011).
27. C. Aydin, S. A. Mansour, Z. A. Alahmed, and F. Yakuphanoglu, *J. Sol-Gel Sci. Technol.* **62**, 397 (2012).
28. Z. D. Pozun and G. Henkelman, *J. Chem. Phys.* **134**, 224706 (2011).
29. M. B. Sahana, C. Sudhakar, G. Setzler, A. Dixit, J. S. Thakur, G. Lawes, R. Naik, V. M. Naik, and P. P. Vaishnava, *Appl. Phys. Lett.* **93**, 231909 (2008).
30. P. Kumar, R. K. Singh, N. Rawat, P. B. Barman, S. C. Kattiyal, H. Jang, H. N. Lee, and R. Kumar, *J. Nanopart. Res.* **15**, 1532 (2013).
31. L. Lu, L. Li, X. Wang, and G. Li, *J. Phys. Chem. B* **109**, 17151 (2005).
32. K. Sivula, R. Zboril, F. L. Formal, R. Robert, A. Weidenkaff, J. Tucek, J. Frydrych, and M. Grätzel, *J. Am. Chem. Soc.* **132**, 7436 (2010).
33. N. Pailhé, A. Wattiaux, M. Gaudon, and A. Demourgues, *J. Solid State Chem.* **181**, 2697 (2008).
34. S. Chaleawlerumpon and N. Pimpha, *Mater. Chem. Phys.* **135**, 1 (2012).
35. Z. Xin, S. Xiao-Hui, and Z. Dian-Lin, *Chin. Phys. B* **19**, 086802 (2010).
36. M. F. Al-Kuhaili, M. Saleem, and S. M. A. Durrani, *J. Alloys Compd.* **521**, 178 (2012).

37. A. D. Trolino, E. M. Bauer, G. Scavia, and C. Veroli, *J. Appl. Phys.* **105**, 113109 (2009).
38. G. P. Joshi, N. S. Saxena, R. Mangal, A. Mishra, and T. P. Sharma, *Bull. Mater. Sci.* **26**, 387 (2003).
39. P. Tyagi and A. G. Vedeshwar, *Bull. Mater. Sci.* **24**, 297 (2001).
40. L. Brus, *J. Phys. Chem.* **90**, 2555 (1986).
41. J. C. Slater, *Phys. Rev.* **103**, 1631 (1956).
42. K. L. Bray, *Top. Curr. Chem.* **213**, 1 (2001).
43. E. Güneri and A. Kariper, *J. Alloy. Compd.* **516**, 20 (2012).
44. D. K. Dwivedi, Dayashankar, B. B. Singh, and M. Dubey, *J. Non-Cryst. Solids* **356**, 1563 (2010).
45. S. Sönmezoglu, A. Arslan, T. Serin, and N. Serin, *Phys. Scr.* **84**, 065602 (2011).
46. M. Rajendran, M. G. Krishna, and A. K. Bhattacharya, *Int. J. Mod. Phys.* **15**, 201 (2001).
47. T. Tan, Z. Liu, H. Lu, W. Liu, and H. Tian, *Opt. Mater.* **32**, 432 (2010).
48. M. Oztas, *Chin. Phys. Lett.* **25**, 4090 (2008).
49. M. Harris, H. A. Macleod, S. Ogura, E. Pelletier, and B. Vida, *Thin Solid Films* **57**, 173 (1979).
50. R. G. Shimmin, A. B. Schoch, and P. V. Braun, *Langmuir* **20**, 5613 (2004).
51. K. Kandori, N. Yamamoto, A. Yasukawa, and T. Ishikawa, *Phys. Chem. Chem. Phys.* **4**, 6116 (2002).
52. K. Brezesinski, J. Haetge, J. Wang, S. Mascotto, C. Reitz, A. Rein, S. H. Tolbert, J. Perlich, B. Dunn, and T. Brezesinski, *Small* **7**, 407 (2011).
53. B. Sun, J. Horvat, H. S. Kim, W.-S. Kim, J. Ahn, and G. Wang, *J. Phys. Chem. C* **114**, 18753 (2010).
54. B. Ahmmad, K. Leonard, Md. S. Islam, J. Kurawaki, M. Muruganandham, T. Ohkubo, and Y. Kuroda, *Adv. Powder Technol.* **24**, 160 (2013).
55. S. Benramache and B. Benhaoua, *Superlattice Microst.* **52**, 1062 (2012).
56. A. Zendeenam, M. Shirazi, S. Daulatshah, and M. Sadat, *Armenian J. Phys.* **3**, 305 (2010).
57. J. H. Lee, W. C. Song, J. S. Yi, K. J. Yang, W. D. Han, and J. Hwang, *Thin Solid Films* **431-432**, 349 (2003).
58. K. C. Preetha, K. V. Murali, A. J. Ragina, K. Deepa, A. C. Dhanya, and T. L. Remadevi, *IOP Conf. Series: Materials Science and Engineering* **43** 012009 (2013).
59. L. Irimpan, V. P. N. Nampoori, P. Radhakrishnan, B. Krishnan, and A. Deepthy, *J. Appl. Phys.* **103**, 033105 (2008).
60. P. Taneja and Pushan Ayyub, *Phys. Rev. B* **65**, 245412 (2002).
61. S. T. Tan, B. J. Chen, X. W. Sun, W. J. Fan, H. S. Kwok, X. H. Zhang, and S. J. Chua, *J. Appl. Phys.* **98**, 013505 (2005).
62. K. Mörl, U. Röpke, B. Knappe, J. Lehmann, R. Perthel, and H. Schröder, *Thin Solid Films* **60**, 49 (1979).
63. A. A. Akl, *Appl. Surf. Sci.* **256**, 7496 (2010).
64. I. W. Chen and X. H. Wang, *Nature* **404**, 168 (2002).
65. Z. He and J. Ma, *J. Phys. D: Appl. Phys.* **35**, 2217 (2002).