INFONET, GIST
Journal Club

Robust Face Recognition via Sparse Representation

Authors:	J. Wright, Allen, Y. Yang, A. Ganesh, S. Shankar, and Yi Ma
Publication:	IEEE Trans. On Pattern Analysis and Machine Intelligence, Feb.2009
Speaker:	Woongbi Lee

Short summary: In this paper, sparse signal representation is investigated for recognizing human faces from frontal views with varying expression and illumination, as well as occlusion and disguise. Based on a sparse representation computed by l^1-minimization, this face recognition problem is cast as a general classification among multiple linear regression models. Even with severe occlusion and corruption, their algorithms show high performance classification of high dimensional data.

I. Introduction

In this paper, the discriminative nature of sparse representation for classification is exploited. Instead of using the generic dictionaries, the test sample is represented in an overcomplete dictionary whose base elements are the training samples themselves. If sufficient training samples are available from each class, it will be possible to represent the test samples as a linear combination of just those training samples from the same class. This representation is naturally sparse, involving only a small fraction of the overall training database. In many problems of interest, it is actually the sparsest linear representation of the test sample in terms of this dictionary and can be recovered efficiently via l^1-minimization.

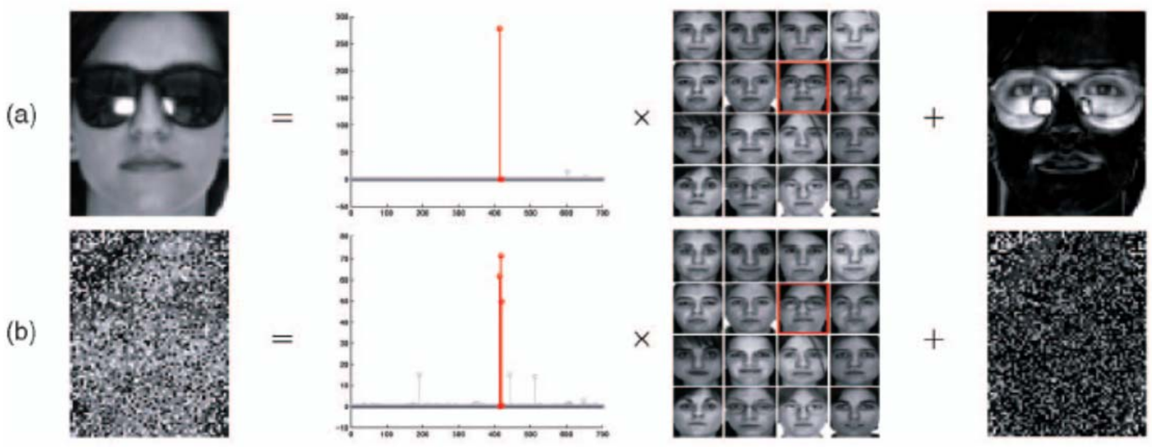


Fig. 1 Simple Example of face recognition: (a) occluded (b) corrupted

There are 700 training images of 100 individuals (7 each). A test image (left) is represented as a sparse linear combination of all the training images plus sparse errors due to occlusion or corruption. Red coefficients correspond to training images of the correct individual.

The theory of sparse representation and compressed sensing yields new insights into two crucial issues in automatic face recognition: the role of feature extraction and the difficulty due to occlusion.

The role of feature extraction: Which low-dimensional features of an object image are the most relevant or informative for classification is a central issue in face recognition. As conventional methods, there exist Eigenfaces (PCA), Fisherfaces (LDA), Laplacianfaces, and a host of variants. The theory of compressed sensing implies that the precise choice of feature space is no longer critical: Even random features contain enough information to recover the sparse representation and hence correctly classify any test image. What is critical is that the dimension of the feature space is sufficiently large and that the sparse representation is correctly computed.

Robustness to occlusion: Occlusion brings about significant troubles in face recognition. This is due to the unpredictable nature of the error occurred by occlusion. Typically, this error corrupts only a small part of the image pixels and therefore is sparse in the standard basis given by individual pixels. The sparse representation of an occluded test image naturally separates the component of the test image arising due to occlusion from the component arising from the identity of the test subject.

II. Classification Based on Sparse Representation

In this section, they introduce classification using sparse representation and show that the sparse representation can be computed by l^1-minimization and can be used for classifying and validating any given test sample.

Object recognition aims to use labeled training samples from k distinct object classes to correctly determine the class to which a new test sample belongs. n_i training samples from the i-th class form a matrix $A_i \doteq \left[v_{i,1}, v_{i,2}, \cdots, v_{i,n_i}\right] \in \mathbb{R}^{m \times n_i}$ whose columns are the training face images of the i-th subject. The image is represented by $w \times h$ gray scale with the vector $v \in \mathbb{R}^m$ ($m = wh$).

A. Test Sample as a Sparse Linear Combination of Training Samples

For the structure of the A_i for recognition, one particularly simple and effective approach models the samples from a single class as lying on a linear subspace. Given sufficient training samples of the i-th object class, $A_i = \left[v_{i,1}, v_{i,2}, \cdots, v_{i,n_i}\right] \in \mathbb{R}^{m \times n_i}$, any new test sample $y \in \mathbb{R}^m$ from the same class will approximately lie in the linear span of the training samples associated with object i:

$$y = \alpha_{i,1} v_{i,1} + \alpha_{i,2} v_{i,2} + \cdots + \alpha_{i,n_i} v_{i,n_i}$$

for some scalars, $\alpha_{i,j} \in \mathbb{R}$, $j = 1, 2, \ldots, n_i$.

A new matrix **A** is defined for entire training set as the concatenation of the n training samples of all k object classes:

$$A_i \doteq [A_1, A_2, \cdots, A_k] = \left[v_{i,1}, v_{i,2}, \cdots, v_{i,n_k}\right]$$

Then, the linear representation of y can be written as

$$y = Ax_0 \in \mathbb{R}^m$$

where $x_0 = \left[0, \cdots, 0, \alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n_i}, 0, \cdots, 0\right]^T \in \mathbb{R}^n$ is a coefficient vector whose entries are zero except those associated with the i-th class.

Obviously, if $m > n$, the system of equations $y = Ax$ is over-determined, and the correct x_0 can usually be found as its unique solution. In robust face recognition, the system $y = Ax$ is typically under-determined, and so, its solution is not unique. Conventionally, this difficulty is resolved by choosing the minimum l^2-norm solution.

$$\left(l^2\right): \hat{x}_2 = \arg\min \|x\|_2 \quad \text{subject to} \quad Ax = y$$

While this optimization problem can be easily solved by the pseudo-inverse of A, the solution $\hat{x}_2$ is not especially informative for recognizing the test sample y. To resolve this difficulty, they instead exploit the following optimization problem:

$$\left(l^0\right): \hat{x}_0 = \arg\min \|x\|_0 \quad \text{subject to} \quad Ax = y$$

where $\|\cdot\|_0$ denotes the l^0-norm, which counts the number of nonzero entries in a vector. However, the problem of finding the sparsest solution of an underdetermined system of linear equations is NP-hard and difficult even to approximate.

B. Sparse Solution via l^1-Minimization

If the solution x_0 sought is sparse enough, the solution of the l^0-minimization problem is equal to the solution to the following l^1-minimization problem.

$$\left(l^1\right): \hat{x}_1 = \arg\min \|x\|_1 \quad \text{subject to} \quad Ax = y$$

This problem can be solved in polynomial time by standard linear programming methods.

1) Dealing with Small Dense Noise

Since real data are noisy,

$$y = Ax_0 + z$$

where $z \in \mathbb{R}^m$ is a noise term with bounded energy $\|z\|_2 < \varepsilon$. The sparse solution x_0 can be approximately recovered by solving the following stable l^1-minimization problem:

$$\left(l_s^1\right): \hat{x}_1 = \arg\min \|x\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 \le \varepsilon$$

This convex optimization problem can be efficiently solved via second-order cone programming.

C. Classification Based on Sparse Representation

Given a test sample y from one of the classes in the training set, the sparse representation $\hat{x}_1$ is computed. Noise and modeling error may lead to small nonzero entries associated with multiple object classes. Based on the global sparse representation, one can design many possible classifiers to resolve this. By using linear structure, we instead classify *y* based on how well the coefficients associated with all training samples of each object reproduce y. For each class *i*, let $\delta_i: \mathbb{R}^n \to \mathbb{R}^n$ be the characteristic function that selects the coefficients associated with the *i*-th class. For $x \in \mathbb{R}^n$, $\delta_i(x) \in \mathbb{R}^n$ is a new vector whose only nonzero entries are the entries in *x* that are associated with class *i*. Using only the coefficients associated with the *i*-th class, one can approximate the given test sample y as $\hat{y}_i = A\delta_i(\hat{x}_1)$. We then classify y based on the these approximations by assigning it to the object class that minimizes the residual between y and $\hat{y}_i$:

$$\min_i r_i(y) \doteq \left\| y - A\delta_i(\hat{x}_1) \right\|_2$$

The recognition procedure is summarized in Algorithm 1. Implementation is done by minimizing the l^1-norm via a primal-dual algorithm for linear programming.

```
Algorithm 1: Sparse Representation-based Classification (SRC)
1: **Input**: a matrix of training samples
   $A = [A_1, A_2, \cdots, A_k] \in \mathbb{R}^{m \times n}$ for $k$ classes, a test sample $y \in \mathbb{R}^m$, (and an optional error tolerance $\varepsilon > 0$.)
2: Normalize the columns of A to have unit $l^2$-norm.
3: Solve the $l^1$-minimization problem:
   $\hat{x}_1 = \arg\min_x \|x\|_1$ subject to $Ax = y$.
   (Or alternatively, solve $\hat{x}_1 = \arg\min_x \|x\|_1$ subject to $\|Ax - y\|_2 \le \varepsilon$.)
4: Compute the residuals $r_i(y) \doteq \left\| y - A\delta_i(\hat{x}_1) \right\|_2$ for $i = 1, \cdots, k$.
5: **Output**: identity $(y) = \arg\min_i r_i(y)$.
```

Example 1. Original image: 192 x 168, downsampled: 12 x 10, size of matrix *A* is 120 x 1207, 38 classes

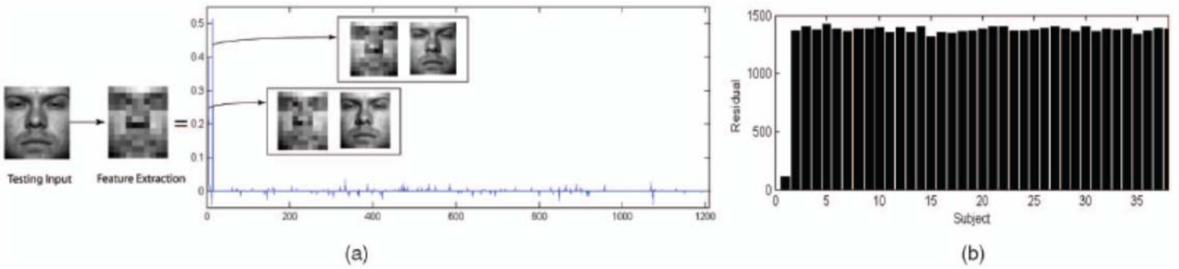


Fig. 2 A valid test image: (a) coefficients (b) residuals

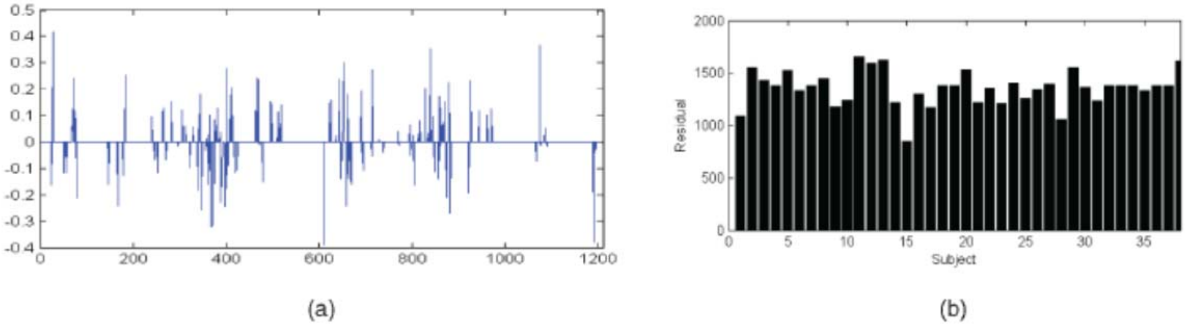


Fig. 3 Non-sparsity of the l^2-minimizer: (a) coefficients (b) residuals

D. Validation Based on Sparse Representation

Before classifying a given test sample, test for validity of the sample is necessary.

Example 2. Randomly selecting an irrelevant image from Google and downsample it to 12 x 10 in the example 1

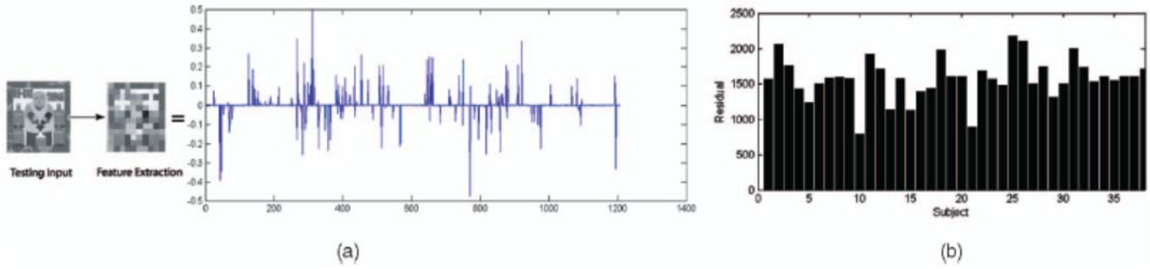


Fig. 4 Example of an invalid test: (a) coefficients (b) residuals

Definition 1 (sparsity concentration index (SCI))

$$SCI\left(x\right) \doteq \frac{k \cdot \max_i \left\|\delta_i\left(x\right)\right\|_1 / \left\|x\right\|_1 - 1}{k-1} \in \left[0,1\right]$$

If $SCI\left(\hat{x}\right) = 1$, the test image is represented using only images from a single object, and if $SCI\left(\hat{x}\right) = 0$, the sparse coefficients are spread evenly over all classes. So, we can choose a threshold $\tau \in \left(0,1\right)$ and do the validity test.

If $SCI\left(\hat{x}\right) \geq \tau$, then the test image is valid.

III. Two Fundamental Issues in Face Recognition

A. The Role of Feature Extraction

One benefit of feature extraction, which carries over to the proposed sparse representation framework, is reduced data dimension and computational cost. Conventionally, on class of methods extract holistic face features such as Eignefaces, Fisherfaces, and Laplacianfaces. Another class of methods tries to extract meaningful partial facial feature such as patches around eyes or nose.

Since most feature transformations involve only linear operations, the projection from the image space to the feature space can be represented as a matrix $R \in \mathbb{R}^{d \times m}$ with $d \ll m$.

$$\tilde{y} \doteq Ry = RAx_0 \in \mathbb{R}^d$$

The system of equations $\tilde{y} = RAx \in \mathbb{R}^d$ is underdetermined in the unknown $x \in \mathbb{R}^n$. By solving the following l^1-minimization problem,

$$\left(l_r^1\right) : \hat{x}_1 = \arg\min \left\|x\right\|_1 \quad \text{subject to} \quad \left\|RAx - \tilde{y}\right\|_2 \leq \varepsilon$$

for a given error tolerance $\varepsilon > 0$.

For the sparse representation approach to recognition, it is important how the choice of the feature extraction R affects the ability of the l^1-minimization to recover the correct sparse solution x_0. There is remarkable analysis: if the solution x_0 is sparse enough, then with overwhelming probability, it can be correctly recovered via l^1-minimization from any sufficiently large number d of linear measurements $\tilde{y} = RAx_0$. In other words, if x_0 has $t << n$ nonzeros, then with overwhelming probability, $d\left(\geq 2t\log\left(n/d\right)\right)$ random linear measurements are sufficient for l^1-minimization to recover the correct sparse solution x_0. Random features can be regarded as a less-structured counterpart to classical face features such as Eignefaces or Fisherfaces. The linear projection generated by a Gaussian random matrix is called Randomfaces.

Definition 2. Randomfaces.

- a transform matrix $R \in \mathbb{R}^{d\times m}$ whose entries are independently sampled from a zero mean normal distribution, and each row is normalized to unit length.

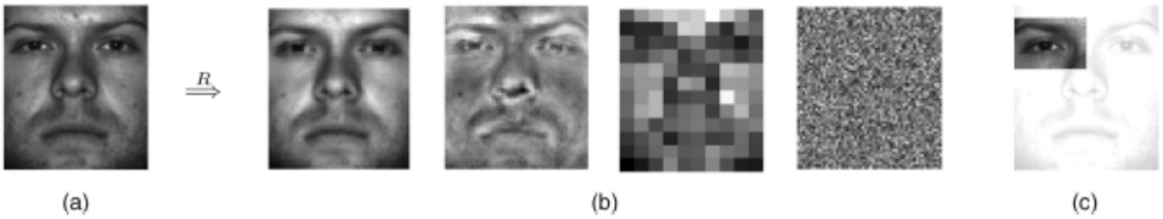


Fig. 5 Examples of feature extraction (a) Original face image (b) 120D representations in terms of four different features: Eigenfaces, Laplacianfaces, downsampled (12 x 10 pixel) image, and random projection (c) The eye is a popular choice of feature for face recognition. In this case, the feature matrix R is simply a binary mask.

B. Robustness to Occlusion or Corruption

In many practical face recognition scenarios, the test image y could be partially corrupted or occluded. The linear model can be modified as

$$y = y_0 + e_0 = Ax_0 + e_0$$

where $e_0 \in \mathbb{R}^m$ is a vector of errors – a fraction, ρ_r of its entries are nonzero. The nonzero entries of e_0 model which pixels in y are corrupted or occluded.

Let us assume that the corrupted pixels are a relatively small portion of the image. Then, the above equation can be written as

$$y = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x_0 \\ e_0 \end{bmatrix} \doteq Bw_0$$

where $B = \begin{bmatrix} A & I \end{bmatrix} \in \mathbb{R}^{m\times(n+m)}$, so the system $y = Bw$ is always underdetermined and does not have a unique solution for *w*. From the analysis of sparsity of x_0 and e_0, the correct generating w_0 has at most $n_i + \rho m$ nonzeros. We want to recover w_0 as the sparsest solution to the system $y = Bw_0$.

$$\left(l_e^1\right): \widehat{w}_1 = \arg\min \|w\|_1 \text{ subject to } Bw = y$$

Once the sparse solution $\widehat{w_1} = \begin{bmatrix} \hat{x}_1 & \hat{e}_1 \end{bmatrix}$ is computed, setting $y_r \doteq y - \hat{e}_1$ recovers a clean image of the subject with occlusion or corruption compensated for. To identify the subject, we slightly modify the residual $r_i\left(y\right)$, computing it against the recovered image y_r

$$r_i\left(y\right) = \left\|y_r - A\delta_i\left(\hat{x}_1\right)\right\|_2 = \left\|y - \hat{e}_1 - A\delta_i\left(\hat{x}_1\right)\right\|_2$$

IV. Experimental Verification

A. *Feature Extraction and Classification Methods*

SRC algorithm using several conventional holistic face features, namely, Eigenfaces, Laplacianfaces, and Fisherfaces, and compare their performance with two unconventional features: randomfaces and downsampled images. They compare the SRC algorithm with three classical algorithms, namely, NN, and NS, discussed in the previous section, as well as linear SVM (support vector machine).

Solving $\left(l_r^1\right): \hat{x}_1 = \arg\min \|x\|_1$ subject to $\left\|RAx - \tilde{y}\right\|_2 \le \varepsilon$ with the error tolerance $\varepsilon = 0.05$.

1) *Extended Yale B Database*

The Extended Yale B database consists of 2,414 frontal-face images of 38 individuals. The copped and normalized 192 x 168 face images were captured under various laboratory-controlled lighting conditions. For each subject, they randomly select half of the images for training (about

32 images per subject) and the other half for testing. They compute the recognition rates with the feature space dimensions 30, 56, 120, and 504, whose numbers corresponding to downsampling ratios of 1/32, 1/24, 1/16, and 1/8, respectively.

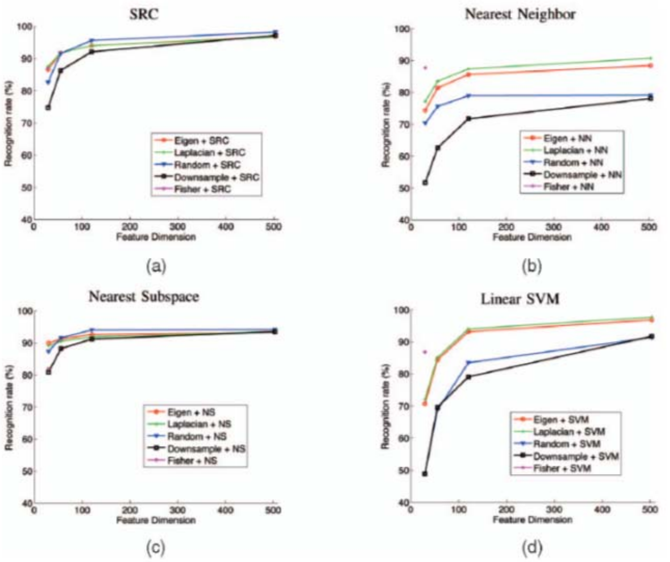


Fig. 6 Recognition rates on Extended Yale B database, for various feature transformation and classifiers. (a) SRC (b) NN (c) NS (d) SVM

2) *AR Database*

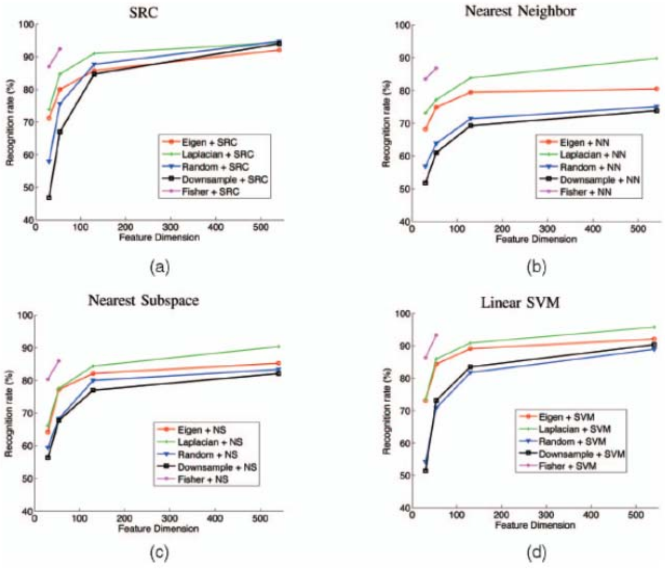
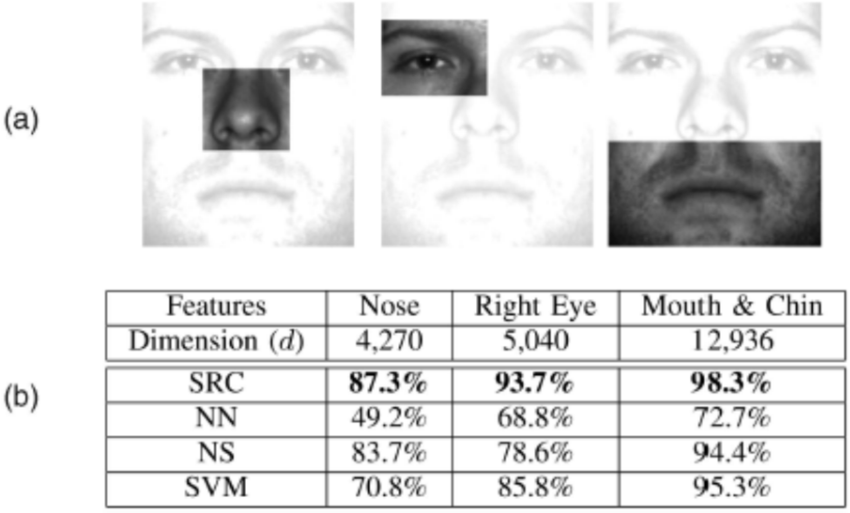


Fig. 7 Recognition rates on AR database, for various feature transformation and classifiers. (a) SRC (b) NN (c) NS (d) SVM

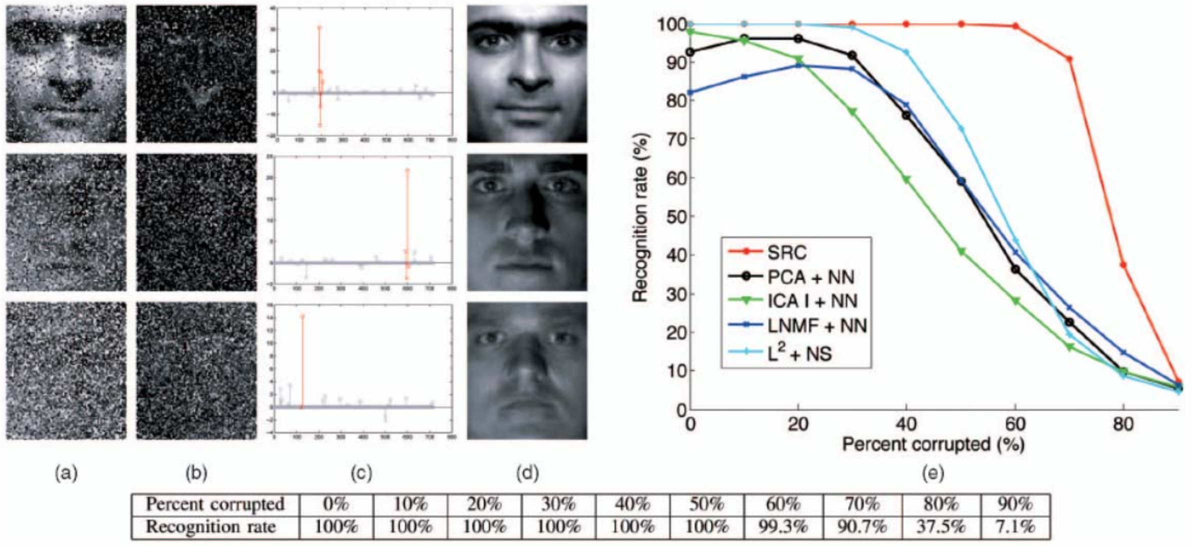
B. Partial Face Features



Features	Nose	Right Eye	Mouth & Chin
Dimension (d)	4,270	5,040	12,936
SRC	**87.3%**	**93.7%**	**98.3%**
NN	49.2%	68.8%	72.7%
NS	83.7%	78.6%	94.4%
SVM	70.8%	85.8%	95.3%

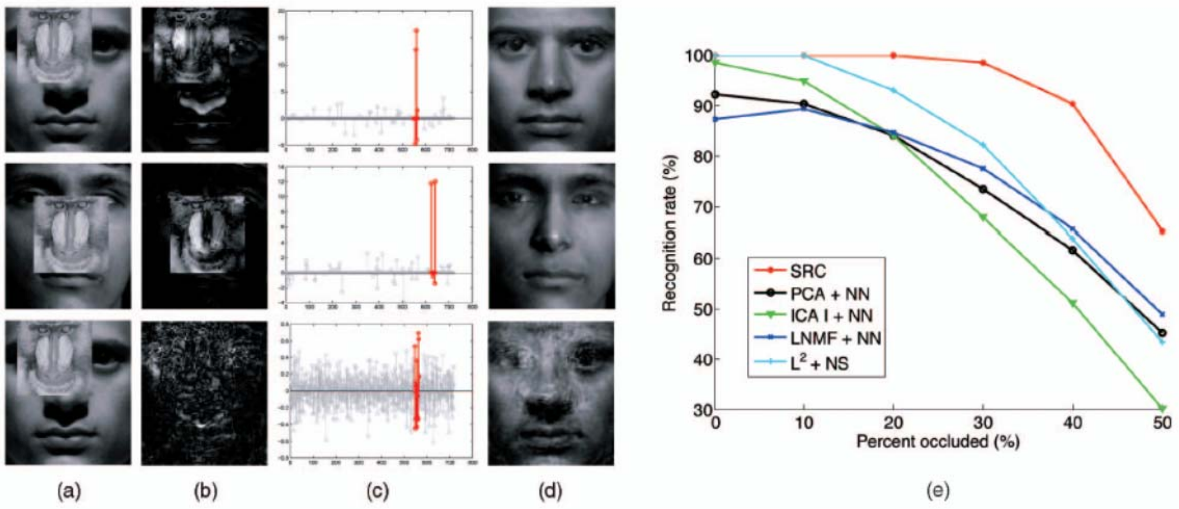
Fig. 8 Recognition with partial face features (a) example features, (b) recognition rates of SRC, NN, NS, and SVM on the Extended Yale B database.

C. Recognition Despite Random Pixel Corruption



Percent corrupted	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Recognition rate	100%	100%	100%	100%	100%	100%	99.3%	90.7%	37.5%	7.1%

D. *Recognition Despite Random Block Occlusion*



Percent occluded	0%	10%	20%	30%	40%	50%
Recognition rate	100%	100%	99.8%	98.5%	90.3%	65.3%

V. Conclusions

In this paper, exploiting sparsity can be used for the high-performance classification of high-dimensional data such as face images. The number of features is more important than the choice of features. With occluded and corrupted images, their classification algorithm shows still high-performance.

VI. DISCUSSION

After meeting, please write discussion in the meeting and update your presentation file.

Appendix

Reference

[1]
[2]
[3]