Journal of Electronic Imaging

JElectronicImaging.org

Progressive sparse representation-based classification using local discrete cosine transform evaluation for image recognition

Xiaoning Song
Zhen-Hua Feng
Guosheng Hu
Xibei Yang
Jingyu Yang
Yunsong Qi

SPIE. IS&T imaging.org

Progressive sparse representation-based classification using local discrete cosine transform evaluation for image recognition

Xiaoning Song,[a,b,*] Zhen-Hua Feng,[b] Guosheng Hu,[b] Xibei Yang,[c] Jingyu Yang,[d] and Yunsong Qi[c]
[a]Jiangnan University, Department of computer science, School of Internet of Things Engineering, Wuxi 214122, China
[b]University of Surrey, Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, Guildford GU2 7XH, United Kingdom
[c]Jiangsu University of Science and Technology, Department of computer science, School of Computer Science and Engineering, Zhenjiang 212003, China
[d]Nanjing University of Science and Technology, Department of computer science, School of Computer Science and Engineering, Zhenjiang 212003, China

Abstract. This paper proposes a progressive sparse representation-based classification algorithm using local discrete cosine transform (DCT) evaluation to perform face recognition. Specifically, the sum of the contributions of all training samples of each subject is first taken as the contribution of this subject, then the redundant subject with the smallest contribution to the test sample is iteratively eliminated. Second, the progressive method aims at representing the test sample as a linear combination of all the remaining training samples, by which the representation capability of each training sample is exploited to determine the optimal "nearest neighbors" for the test sample. Third, the transformed DCT evaluation is constructed to measure the similarity between the test sample and each local training sample using cosine distance metrics in the DCT domain. The final goal of the proposed method is to determine an optimal weighted sum of nearest neighbors that are obtained under the local correlative degree evaluation, which is approximately equal to the test sample, and we can use this weighted linear combination to perform robust classification. Experimental results conducted on the ORL database of faces (created by the Olivetti Research Laboratory in Cambridge), the FERET face database (managed by the Defense Advanced Research Projects Agency and the National Institute of Standards and Technology), AR face database (created by Aleix Martinez and Robert Benavente in the Computer Vision Center at U.A. B), and USPS handwritten digit database (gathered at the Center of Excellence in Document Analysis and Recognition at SUNY Buffalo) demonstrate the effectiveness of the proposed method. [DOI: 10.1117/1.JEI.24.5.053010]

Keywords: sparse representation-based classification; local discrete cosine transform evaluation; progressive learning; image recognition.

Paper 15277 received Apr. 11, 2015; accepted for publication Aug. 18, 2015; published online Sep. 21, 2015.

1 Introduction

The great success of sparse representation in image processing triggers a great deal of research and practices on representation-based pattern classification. Recently, a sparse representation-based classification (SRC) method[1,2] has been successfully proposed by Wright et al., which led to promising results in face recognition. SRC aims to represent a test sample using an untraditional dictionary that consists of all training samples across all subjects. Then, it evaluates the contribution to represent the test sample from each class and exploits the optimal evaluation result to classify the test sample. The role of the SRC is that the performance is measured in terms of sparsity of the representation and fidelity to the original signals. In fact, the traditional dictionary learning rules were not assumed to have any particular semantic meaning because they are typically chosen from standard bases such as Fourier, Wavelet, Curvelet, and Gabor, or even random matrices.[3,4] Fortunately, we can reach a conclusion that the robust representation has naturally discriminative properties on the basis of the literatures:[1,2] among all the subsets of base vectors, it selects the subset that well reconstructs the input signal and rejects all the other possible but inferior representations. For example, some alternative methods have been proposed to reconstruct the dimensional structure of data, and can robustly extract better discriminant features. Specifically, Zhang et al.[5] proposed the tensor linear Laplacian discrimination (TLLD) method for nonlinear feature extraction from tensor data, which could be viewed as an extension of both LDA and LLD[6] in directions of nonlinearity and tensor representation. Meanwhile, to handle the high dimensional face-shape regression problem, a hierarchical pose regression approach has been proposed by Zhang et al.[7] to hierarchically estimate the head rotation, face components, and facial landmarks. Through empirical studies, Sun et al.[8] also discovered three properties of deep neural activations critical for the high performance: sparsity, selectiveness, and robustness. Therefore, the capability of robust representation to uncover discriminative information depends on the effective learning model, which exploits important structures of training samples.

Meanwhile, the local untraditional dictionary learning methods were also developed to enhance representation-based

*Address all correspondence to: Xiaoning Song, E-mail: xnsong@aliyun.com

1017-9909/2015/$25.00 © 2015 SPIE and IS&T

pattern classification. Previous studies have shown that the local learning methods have been focused on by lots of researchers in recent years.[9–16] For example, Yang et al. in Ref. 17 introduced the similarity and distinctiveness of features into collaborative representation-based classification (CRC),[18] and present a more general model. Xu et al.[19,20] have proposed selection strategies to seek best valued training samples for a new SRC model, which had clear rationales and achieved high face-recognition accuracies. Especially, the method proposed in Ref. 19 is somewhat associated with the idea of CRC, but the former usually leads to higher accuracy due to the use of the reasonable selection strategy. Representation-based classification has also been extended for bimodal biometrics.[21,22] Yang et al.[23,24] proposed Kernel CRC and CRC projections method. Liu et al.[25] evaluated the reconstruction error of the test sample to improve the accuracy of CRC. For the abovementioned studies, we refer to all of the algorithms that exploit only a subset of the training samples rather than all of them to classify each test sample as "local" methods. It can also be viewed as an evaluation method that introduces "local" training samples to represent and classify the query samples. Specifically, some heuristic strategies are generally presented to determine the optimized "local" training samples for the test sample. We then use these remaining "local" training samples to perform classification. Therefore, we can conclude that the "local" methods practically convert the original classification problem into a simpler one that contains scale-reduced subjects. In fact, the rationale of these "local" methods is described as follows: it has been empirically proven and widely admitted that if the test sample is provided with high correlation to the training samples from a subject, it should be great reasonable to classify the test sample into this subject.

In addition, discrete trigonometric transform such as DCT has been widely used in image processing for transform-based coding. The main reason for employing DCT to transform image-feature space is that features can be extracted from images in a compressed format. Here, compression means that the signal enclosed by a limited number of DCT coefficients can be restored. Meanwhile, the cosine similarity measure based on Bhattacharya's distance is defined as the inner product of these two vectors divided by the product of their lengths, which is a classical measure method. The related work is proposed in Refs. 26–29. Furthermore, local image descriptors based on interesting regions have proven to be very successful in pattern-recognition applications. Song and Li[30] proposed local polar DCT features (LPDFs), which extract and rearrange the selected two-dimensional (2-D) DCT features in a designed polar geometric structure.

The present study proposes a progressive sparse representation-based classification algorithm (P-SRC) using a local correlative degree evaluation to perform face recognition. The contributions of our work are threefold.

- The contribution from each subject for representing a test sample is calculated by the sum of the contributions of all the training samples of this subject, then the redundant subject with the smallest score could be eliminated from the training set. This procedure is iteratively implemented for the remaining subjects till the predefined termination conditions have been met.
- The transformed DCT evaluation is constructed to measure the similarity between the test sample and each local training sample by using cosine distance metrics in the DCT domain. Then the representation capability of each local training sample is exploited to determine the optimal "nearest neighbors" for representing the test sample.
- The final goal of the proposed method is to generate an optimal weighted sum of L nearest neighbors that is obtained under the local correlative degree evaluation, which is approximately equal to the test sample and we can use this weighted linear combination to perform classification. It is noted that the mechanism of the progressive dictionary learning in this method not only achieves a high accuracy but also can be clearly interpreted.

The rest of the paper is organized as follows: Sec. 2 describes a typical global training samples representation-based algorithm. The proposed method is presented in Sec. 3. Experimental results are reported in Sec. 4 and Sec. 5 concludes the paper.

2 Outline of Global Training Samples Representation

As the typical sparse representation-based classification algorithm, a different classification rule for the test sample can be developed, in which the linear combination (i.e., coefficient) of training samples is sequentially handled. This leads to the global training samples representation method.[19] This section introduces the global method that exploits all the training samples to represent and classify the test sample.

Suppose that there are n training samples, respectively, denoted by n column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The global method assumes that test sample $\mathbf{y}$ can be approximately represented by a linear combination of all the training samples. In other words, the following equation is approximately satisfied:

$$\mathbf{y} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i. \tag{1}$$

Equation (1) can be rewritten as

$$\mathbf{y} = \mathbf{X}\alpha, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. If $\mathbf{X}^T\mathbf{X}$ is nonsingular, the least-squares solution of Eq. (2) is:

$$\alpha = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3}$$

if $\mathbf{X}^T\mathbf{X}$ is (or is nearly) singular, we solve α using

$$\alpha = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{4}$$

where λ is a small positive constant and $\mathbf{I}$ is the identity matrix. Once we obtain α using Eqs. (3) or (4), we refer to $\|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|$ as the deviation of the linear combination $\mathbf{X}\boldsymbol{\alpha}$ from test sample $\mathbf{y}$.

According to Eq. (1), we know that the contribution of the i'th training sample is $\alpha_i\mathbf{x}_i$; thus, we can calculate the sum of the contribution of the training samples from one

class. For instance, if all the training samples from the k'th class are $\mathbf{x}_s, \ldots, \mathbf{x}_t$, then the contribution of the k'th class is $\mathbf{g}_k = \alpha_s \mathbf{x}_s + \cdots + \alpha_t \mathbf{x}_t$. The deviation of g_k from $\mathbf{y}$ can be calculated by $D_k = \|\mathbf{y} - \mathbf{g}_k\|^2$, and the smallest deviation D_k means that $\mathbf{g}_k$ has the combination of the most competitive within-class training samples to represent the test sample.

3 Progressive Sparse Representation-Based Classification

The proposed framework can be divided into three stages, namely, redundant subject elimination stage, local correlative degree evaluation stage, and the classification stage. Figure 1 shows the schematic diagram of the proposed P-SRC framework.

3.1 *Motivation of the Present Work*

There exist two previous works that we have undertaken on how to design traditional or untraditional dictionaries to better fit the sparse model by either selecting one from a prespecified set of linear transforms or adapting the dictionary to a set of training signals.

Specifically, in our previous work,[31] we develop an iterative class elimination algorithm to represent a test sample as a linear combination of the relatively competitive training samples; here, this training set can be viewed as an untraditional dictionary. The contribution from each subject in presenting the test sample is calculated by the sum of contribution of all the training samples of this subject, then the redundant subject with the smallest score to this test sample would be eliminated from the training sets. This procedure is iteratively implemented for the remaining subjects after the predefined termination conditions have been met, and the final remaining training samples are used to generate a representation of the test sample and to classify it. Meanwhile, a different update rule for the dictionary can be proposed, in which the atoms (i.e., columns) in the dictionary are sequentially handled. This leads to the K-means singular value decomposition (K-SVD) algorithm, as developed by Aharon et al.[32] In another previous study,[33] considering that the existing K-SVD algorithm is employed to dwell on the concept of a binary class assignment, which means that the multiclass samples are definitely assigned to the given classes. Thus, the method proposed in our study provides a parameterized fuzzy adaptive way to adapting the traditional dictionaries, which is called parameterized fuzzy K-SVD. Actually, it is worth stressing that the proposed fuzzy K-SVD algorithm has been one of the effective algorithms due to its capacity for optimization update rule for the traditional dictionary.

In contrast, the key idea of the present work is to accomplish face recognition by interpreting a P-SRC algorithm under a local correlative degree evaluation. The proposed method has the following characteristics: the remaining samples obtained by the iterative elimination stage of redundant subjects still have different representation abilities in representing the test sample. This motivates us to consider whether we can construct an evaluation model to find the more important training samples that hold high correlations to the test sample. Now, we highlight the favorable properties of P-SRC and main contributions of this work as follows. First, when the initial updated training set is obtained by iterative elimination stage of redundant subjects, we construct the transformed DCT evaluation to measure the similarity between the test sample and each remaining training sample by using cosine distance metrics in the DCT domain. Second, we collect the former L training samples that have large similarity values in the DCT domain; the remaining training samples should be discarded. Third, we determine an optimal weighted sum of L nearest neighbors, which is approximately equal to the test sample, and we can use this weighted linear combination to perform classification.

3.2 *Stage of Redundant Subject Elimination*

This section details the first stage of the proposed P-SRC. This stage aims at representing a test sample as a linear combination of the most competitive training samples, and the update of the coefficient columns is implemented by integrating an elimination mechanism for redundant subjects. Thus, one redundant subject with the smallest score to this test sample would be iteratively eliminated from the training sets. In fact, if there are many indistinctive subjects for test sample's classification, this step only strengthens the most competitive subjects.

Suppose that there are C classes and n training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Motivated by collaborative representation-based

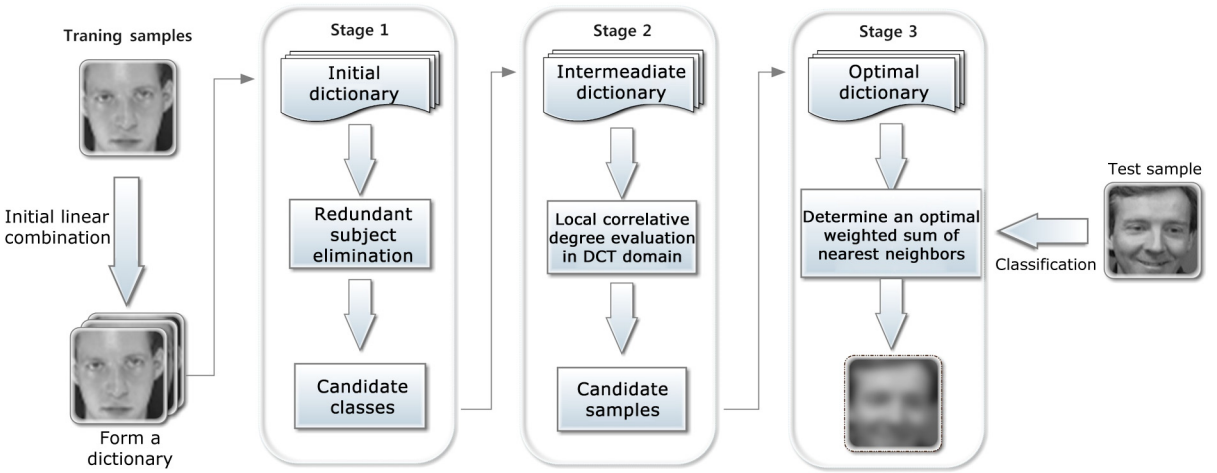

Fig. 1 Schematic diagram of the proposed progressive sparse representation-based classification (P-SRC) algorithm.

classification in Ref. 18, we exploit a linear combination of k'th subject to represent the test sample. If $\mathbf{y}$ belongs to the k'th class, it should be represented as below:

$$\mathbf{y} = \overline{\mathbf{X}}_k \boldsymbol{\beta}_k, \quad k = 1,2,\ldots,C, \tag{5}$$

where $\overline{\mathbf{X}}_k = [\overset{(1)}{\mathbf{w}}_k, \overset{(2)}{\mathbf{w}}_k, \ldots, \overset{(M)}{\mathbf{w}}_k]$, $\overset{(m)}{\mathbf{w}}_k$ is the m'th training sample in k'th class. When $\overline{\mathbf{X}}_k$ is a nonsingular square matrix, $\boldsymbol{\beta}_k$ can be solved by using $\boldsymbol{\beta}_k = \overline{\mathbf{X}}_k^{-1}\mathbf{y}$; otherwise, it can be solved by means of $\boldsymbol{\beta}_k = (\overline{\mathbf{X}}_k^T\overline{\mathbf{X}}_k + \mu\mathbf{I})^{-1}\overline{\mathbf{X}}_k^T\mathbf{y}$, where μ is a small positive constant and $\mathbf{I}$ is the identity matrix. The recursive subject elimination strategy proposed in this stage aims to remove the redundant subjects that receive a small error in the linear representation equation. The subset of training samples from a redundant subject will be entirely removed after each iteration; meanwhile, the linear representation equation is recomputed. Therefore, the theoretical description of the first step of P-SRC is described as follows.

This step essentially builds an empirical risk minimizer (ERM) φ, on the basis of training data $(\overline{\mathbf{X}}_k, \mathbf{y})$, $k = 1,2,\ldots,C$, for which both the signal $\overline{\mathbf{X}}_k$ and the true label $\mathbf{y}$ are known. Evidently, we expect φ to have good generalization performance, meaning that we want the true error rate of φ as low as possible. In fact, φ can be well estimated under the condition that the empirical risk is minimized.[34,35] According to the ERM theory, we conclude that the set of training samples from the k'th class $(k = 1,2,\ldots,C)$ makes a respective contribution to representing the test sample, and this contribution can be evaluated by reconstruction error between $\overline{\mathbf{X}}_k\boldsymbol{\beta}_k$ and $\mathbf{y}$, i.e., $r_k = \|\mathbf{y} - \overline{\mathbf{X}}_k\boldsymbol{\beta}_k\|_2^2$. The r_k can also be viewed as a discrimination measurement between the test sample and the k'th class. It shows that the redundant subject corresponding to the smallest score can be eliminated from the training set, and the procedure is iteratively implemented for the remaining subjects after the predefined termination conditions have been met. Therefore, the first step of P-SRC is treated as an iterative method that alternates between sparse representation and a process of updating the training classes to better fit the test sample.

3.3 *Stage of Local Discrete Cosine Transform Evaluation*

The second stage of P-SRC aims at designing a correlation evaluation for the remaining samples that are obtained from the first stage of our method. Actually, the samples obtained in the first stage still have different representation abilities in representing the test sample. We intend to introduce an evaluation model to find the important training samples that hold high degrees of correlation to the test sample. Hereafter, we can use these competitive training samples to accurately represent the test sample and perform the final classification. In this study, the correlation degree evaluation in DCT domain is introduced first.

According to the cosine distance metrics in DCT domain, it is the cosine value of the angle between two vectors, e.g., $\mathbf{g}_i$ denotes the training samples of the remaining subjects, $\mathbf{y}$ denotes the test sample, as follows:

$$\cos(\mathbf{g}_i, \mathbf{y}) = \frac{\mathbf{g}_i \cdot \mathbf{y}}{\|\mathbf{g}_i\|\|\mathbf{y}\|}, \tag{6}$$

where $\cdot$ indicates the dot-product of the vectors, and $\|\cdot\|$ indicates the length of the vector. For vectors with non-negative elements, the resulting similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

Equation (6) can be viewed as a measurement of the cosine similarity in the DCT domain between $\mathbf{g}_i$ and $\mathbf{y}$. A large value of this measurement means that $\mathbf{g}_i$ is similar to the test sample $\mathbf{y}$, thus we collect the former L training samples that have large similarity values and the remaining training ones should be discarded. Assume that $H = \{h_1,\ldots,h_L\}$ is a set of some numbers, standing for the set of labels of L competitive training samples. The label is defined as follows: if a useful training sample is derived from the k'th class $(k = 1,\ldots,C)$, the category of this training sample is labeled as k. H should be one subset of the set $\{1,\ldots,C\}$, i.e., $H \subset \{1,\ldots,C\}$. On the contrary, if a discarded training sample is from the k'th class $(k' = 1,\ldots,C)$ then the category k' must not be an element of H. Consequently, the test sample could not be finally classified into the k'th class.

3.4 *Classification Stage*

The third stage of the P-SRC is to represent the test sample as a linear combination of the determined L nearest neighbors and uses the representation result to classify the test sample. This phase assumes that the following equation is approximately satisfied:

$$\mathbf{y} = \gamma_1\mathbf{g}_1 + \cdots + \gamma_L\mathbf{g}_L, \tag{7}$$

where $\mathbf{g}_i (i = 1,\ldots,L)$ are the identified L nearest neighbors and $\gamma_i (i = 1,\ldots,L)$ are the coefficients. Here, Eq. (7) is rewritten as

$$\mathbf{y} = \mathbf{G}\gamma, \tag{8}$$

where $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_L]^T$, $G = [\mathbf{g}_1,\ldots,\mathbf{g}_L]$. If $\mathbf{G}$ is a nonsingular square matrix, we can solve $\boldsymbol{\gamma}$ by using $\boldsymbol{\gamma} = \mathbf{G}^{-1}\mathbf{y}$; otherwise, we can solve it by using $\boldsymbol{\gamma} = (\mathbf{G}^T\mathbf{G} + \mu\mathbf{I})^{-1}\mathbf{G}^T\mathbf{y}$, where μ is a small positive constant and $\mathbf{I}$ is the identity matrix.

Since the nearest neighbors might be from different classes, we calculate the sum of the contribution to represent the test sample of the neighbors from each class and exploit the sum to classify the test sample. More specifically, if all the neighbors from the k'th $(k = 1,\ldots,C)$ class are $\mathbf{g}_s,\ldots,\mathbf{g}_t$, then the sum of the contribution to representing the test sample of the k'th class is described as follows:

$$\mathbf{y}_k = \gamma_s\mathbf{g}_s + \ldots + \gamma_t\mathbf{g}_t. \tag{9}$$

Thus, the residual formula of $\mathbf{y}_k$ from $\mathbf{y}$ is presented as below:

$$D_k = \|\mathbf{y} - \mathbf{y}_k\|_2^2. \tag{10}$$

Obviously, the above formula allows the residual between the test sample and each $\mathbf{y}_k$ to be evaluated in a fair way by simultaneously exploiting $\|\mathbf{y} - \mathbf{y}_k\|_2^2$. Finally, a smaller deviation D_k means a greater contribution to representing the test sample, and we classify $\mathbf{y}$ into the class that produces the smallest deviation.

3.5 *Detailed Progressive Sparse Representation-Based Classification Algorithm*

The pipeline of the proposed P-SRC algorithm is detailed as follows:

P-SRC Task:

The update of the training dictionary should be implemented to better represent the test data y, by approximating the solution to the classification problem.

Step 1. **Initialization**:

Suppose that we have C subjects, and let $\overline{\mathbf{X}} = [\overline{\mathbf{X}}_1, \ldots, \overline{\mathbf{X}}_C]$.

Main procedure:

Step 2. **Stage of redundant subject elimination**:

1. Code $\mathbf{y}$ over $\overline{\mathbf{X}}$ by $\boldsymbol{\beta} = \mathbf{P}\mathbf{y}$, obtain coefficient $\boldsymbol{\beta}$, where, $\mathbf{P} = (\overline{\mathbf{X}}^T\overline{\mathbf{X}} + \mu\mathbf{I})^{-1}\overline{\mathbf{X}}^T$, μ is a small positive constant and $\mathbf{I}$ is the identify matrix.
2. Design the following procedure to update the columns of the training dictionary matrix $\overline{\mathbf{X}}$ and obtain $\overline{\mathbf{X}}^j$. Repeat for $j = 1,2,\ldots,\eta$, compute the regularized residuals

$$r_k = \|\mathbf{y} - \overline{\mathbf{X}}_k\boldsymbol{\beta}_k\|_2^2/\|\boldsymbol{\beta}_k\|_2^2, \quad k = 1,\ldots,C.$$

3. Discard the redundant subject corresponding to the smallest score from the training set as $\overline{\mathbf{X}}^j = \{\overline{\mathbf{X}} - \overline{\mathbf{X}}_k\}$, that is, the set of training samples from k'th subject is entirely eliminated.
4. If the predefined termination condition η has been met, go to *Step3*. Otherwise, go to *Step2* for another iteration. Here, η stands for the number of iterations of redundant subject elimination.

Step 3. **Stage of local DCT evaluation**:

1. Employ DCT to transform image feature space, the features can be extracted from images in the compressed format.
2. Use the transformed DCT evaluation to measure the similarity between the test sample and each remaining training sample by using Eq. (6).
3. Exploit $\cos(\mathbf{g}_i, \mathbf{y})$ to identify L training samples that have greatest contributions, representing the test sample by linear combination of $[\mathbf{g}_1, \ldots, \mathbf{g}_L]$.

Step 4. **Stage of classification**:

1. If all the nearest neighbors from the r'th $(r \in C)$ subject are $\mathbf{g}_s, \ldots, \mathbf{g}_t$, where $\{\mathbf{g}_s, \ldots, \mathbf{g}_t\} \subset \{\mathbf{g}_1, \ldots, \mathbf{g}_L\}$, then calculate the sum of the contributions to reconstruct test sample $\mathbf{y}$

$$\mathbf{y}_r = \gamma_s\mathbf{g}_s + \cdots + \gamma_t\mathbf{g}_t.$$

2. The residual of $\mathbf{y}_r$ from $\mathbf{y}$ is calculated by using

$$D_r = \|\mathbf{y} - \mathbf{y}_r\|_2^2.$$

3. Output the identity of $\mathbf{y}$ as

$$\text{Identify}(\mathbf{y}) = \arg\min_r\{D_r\}.$$

3.6 *Discussions*

This section provides an extensive discussion regarding potential advantages of the proposed P-SRC method. The superiority of the proposed method stems from two aspects as follows.

First, considering that the training samples from different subjects might have inherent mutual relationship easily leads to ignoring the relationship between the different subjects when we design a strategy that the redundant subjects are entirely eliminated only once. As a result, the problem of multicollinearity that causes the weights to become extremely large or small will be raised because of numerical ill-conditioning. In contrast, in the present study, the redundant subject corresponding to the smallest score could be eliminated from the training sets with each iteration. In the experiment, the range of elimination percentage is designed as [10%, 90%] with 10% intervals; then the optimal value of η can be empirically assigned. Therefore, this step of P-SRC is viewed as an iterative method that alternates between a sparse representation and a process of updating the training classes to better fit the test sample. It is able to greatly reduce the inverse influence on the classification when a part of training samples from different subjects are very dissimilar to the test sample.

Second, we measure the similarity between the test sample and each local training sample by using cosine distance metrics in the DCT domain. Then the representation capability of each local training sample is exploited to determine the optimal "nearest neighbors" for representing the test sample by a linear formulation. The goal of the proposed method is to determine an optimal weighted sum of nearest neighbors that are obtained under the local correlation evaluation; we then use this weighted linear combination to perform robust classification. Therefore, the proposed method can also be treated as an evaluation method that introduces "local" training samples to represent and classify the query samples, i.e., it belongs to the local untraditional dictionary learning methods. It is worth noting that the local polar DCT features proposed by Song and Li[30] are extracted and used for feature matching between two images; Song and Li first quantize the preprocessed local patch in the gray-level domain. Then the 2-D DCT features in the frequency domain are extracted and a subset of the reordered coefficients selected as compact descriptor. In contrast, we design a criterion of correlation degree in DCT domain to evaluate the training samples. The evaluation aims to find the useful training samples that hold high correlation to represent the test sample. Therefore, this is the key difference between this previous study and the proposed method.

4 Experimental Results

This section conducts the experiments of P-SRC on various image classification tasks, including face recognition and handwritten digit recognition.

4.1 *Face Recognition*

We conducted a number of experiments on the ORL,[36] FERET,[37] and AR[38] databases. The face images of these three databases were obtained under the conditions of varying pose, facial expression, or lighting. Occluded face images are also included in the AR face database.

Meanwhile, a biometric system can be regarded as a pattern-recognition system, where a feature set is first extracted from the original samples and then is compared with the stored template set to make a decision on the identity of an individual. In biometric verification mode,[39] the decision is whether a person is "who he/she claims to be." In biometric identification mode, the decision is "whose biometric data is this?" More specifically, face verification aims to determine whether two given faces refer to the same person. The identification task is inherently more difficult than verification, since the input face image must be compared with, and matched to, each face in the enrolment database. The test face is then identified as belonging to the face class that shows the highest similarity.

There are two main respective evaluation plots for face verification and identification: the receiver operating characteristic (ROC) curve[40] and the recognition rate (PR) curve. The ROC curve examines the relation between the true-positive rate and the false-positive rate, while the PR curve extracts the relation between the number of samples (or classes) and identification precision. Concretely, the ROC curve describes the performance of a verification or diagnostic rule. This curve is generated by plotting the fraction of true positives out of the positives (true-positive rate) versus the fraction of false positives out of the negatives (false-positive rate), at various threshold settings. In the two-class verification case (for example, face and nonface), the true positive means the portion of face images to be detected by the system, while the false positive means the portion of nonface images to be detected as faces. Thus, the ROC curve must be used in the experiments of face verification. In turn, the PR curve is usually employed in the experiments of face identification. Therefore, the present study focuses on a progressive SRC for face recognition (identification), which is the reason why we utilize PR curve to evaluate the performance of the experiments.

The ORL contains a set of faces taken between April 1992 and April 1994 at the Olivetti Research Laboratory in Cambridge, UK. It contains 40 distinct persons with 10 images per subject. The images were taken at different time instances, with varying lighting conditions, facial expressions, and facial details. All persons are in the up-right, frontal position, with tolerance for some side movement. Each image was normalized and presented by a 46×56 pixel array, whose gray levels ranged between 0 and 255. Some sample images from the ORL database are shown in Fig. 2.

The FERET face image database is a result of the FERET program, which was sponsored by the U.S. Department of Defense through the DARPA program. It has become a standard database for the evaluation of state-of-the-art face recognition techniques. The proposed algorithm was evaluated on a subset of FERET database, which includes 1400 images of 200 individuals with seven different images of each individual. Some sample images from the FERET database are shown in Fig. 3. For the AR face database, we used 3120 gray images from 120 subjects with each subject providing 26 images. Some sample images from the AR database are shown in Fig. 4.

We resized each face image of the AR database to a 40×50. The face images of the FERET databases were also resized using the same algorithm. In the experiments on the AR database, which is randomly partitioned into a training set and a test set with no overlap between the two. The partition of the database into training and testing sets, which call for four images per individual, randomly

Fig. 2 Part of images from ORL face image database.

Fig. 3 Part of images from FERET face database.

Fig. 4 Part of images from AR face database.

chosen for training, and the remaining images for test. Thus, a training set of 480 images and a test set with 2640 images are created. To make full use of the available data and to more accurately evaluate the generalization power of algorithms, the figures of merit are success rates averaged over 10 runs, with each run being performed on such random partitions in the AR database. Moreover, the error margin for both methods (mean and standard deviations) is provided in the following experiment. The proposed method exploits 200 finally remaining training samples to represent and classify the test sample on the AR database.

In the experiments on the ORL database, five samples per class were used as training samples and the others were used as test samples. The 40 finally remaining training samples are used to represent and classify the test sample. For the ORL database, four sets of training samples and test samples were generated. The first set of training samples consists of the first, second, third, fourth, and fifth samples of each subject. The second, third, and fourth training sets are composed of the first, second, third, fourth, and sixth samples of each subject; the first, second, third, fourth, and seventh samples of each subject; the first, second, third, fourth, and eighth samples of each subject, respectively. For each set of training samples, the set of test samples consists of the samples that were not used as training samples. In the experiments on the FERET database, we chose the former four images per individual for training and the remaining images for testing. The proposed method exploited 200 finally remaining training samples to represent and classify the test sample.

Table 1 indicates the classification error rates of principle component analysis (PCA),[41] linear discriminant analysis (LDA),[42] two-phase test sample sparse representation (TPTSR),[19] coarse-to-fine face recognition (CFFR),[21] sparse representation-based classification (SRC),[1] extended sparse representation-based classification (ESRC),[43] collaborative representation-based classification (CRC),[18] sparse discriminant analysis with $l_{2,1}$-minimizaiton ($l_{2,1}$-SDA)[44] and the proposed method on the ORL face image database. As shown in Table 1, it is therefore reasonable to believe that the proposed method is the most effective one in the presence of different number of training samples per individual.

Table 2 presents the recognition accuracies of PCA,[41] LDA,[42] SRC,[1] TPTSR,[19] CFFR,[21] and the proposed method, and the error margin for both methods (mean and standard deviations) is given in Table 2. For all methods, the average CPU time consumed for training and testing is also given in Table 2. Likewise, the experimental results indicate that the proposed method is the most effective one among the facial feature extraction approaches.

The comparison of classification error rates between the proposed algorithm and PCA,[41] LDA,[42] SRC,[1] TPTSR,[19] CFFR,[21] and the proposed methods used on the FERET database are also summarized in Table 3. As shown in Table 3, the proposed algorithm performs better than the others.

As described in the above experiments, Tables 1 to 3 show the respective classification results of different methods.

Table 1 The classification error rate (%) of each method varies with number of training samples per individual on the ORL face image database.

	Number of training samples	
Method	5	6
PCA(50)	7.2	5.2
PCA(100)	7.9	6.0
PCA(150)	7.8	6.1
LDA(39)	4.8	3.7
TPTSR	4.4	3.3
CFFR	3.7	2.8
SRC	4.5	3.6
CRC	4.2	3.3
ESRC	4.3	3.3
$l_{2,1}$-SDA	4.8	3.8
Proposed method	**2.9**	**1.4**

Note: The bold values denote the best experimental results compared with other methods.

Note: PCA(50), PCA(100), and PCA(150) indicate that PCA used 50, 100, and 150 transform axes for feature extraction, respectively. LDA(39) means that the LDA used 39 transform axes for feature extraction.

Table 2 Comparison results (%) between the different algorithms on the AR face image database.

Results/methods	PCA(100)	LDA(119)	SRC	TPTSR	CCFR	Proposed method
Accuracy	51.27 ± 3.63	55.12 ± 3.29	64.29 ± 2.88	66.48 ± 2.64	65.78 ± 2.52	$\mathbf{68.25 \pm 2.40}$
CPU time (s)	38.7	40.8	92.6	95.7	97.1	122.8

Note: The bold value denote the best experimental results compared with other methods.
Note: PCA(100) indicates that PCA used 100 transform axes for feature extraction. LDA(119) means that the LDA used 119 transform axes for feature extraction.

Table 3 Comparison results (%) between the different algorithms on FERET face image database.

PCA(100)	LDA(119)	SRC	TPTSR	CFFR	Proposed method
56.8	61.3	54.6	64.7	67.1	**69.2**

Note: The bold value denote the best experimental results compared with other methods.
Note: PCA(100) indicates that PCA used 100 transform axes for feature extraction. LDA(119) means that the LDA used 119 transform axes for feature extraction.

These figures clearly demonstrate that the proposed method is the most effective one among the traditional facial feature extraction approaches. However, it is worth stressing that the proposed method needs more CPU time for the whole process (re-estimation processes) because it costs more computation by producing the sparseness in a supervised way, which alternates between a sparse representation and a process of updating the untraditional training dictionary to better fit the test data.

Moreover, a sensitivity (robust) face-recognition system should deal well with the case where the samples of the same class (subject) are very different due to some factors such as illumination, variable poses, variable expressions, and disguises, especially performed in cases with insufficient samples. Therefore, if we observe the problem from the performance perspective of facial recognition, the high recognition rate of the proposed method could also be regarded as a sensitivity (robust) capability. Therefore, we design an experiment to verify the sensitivity of our proposed method as follows. The AR database is divided into three subsets, which include ordinary subset (14 samples per subject), sunglasses disguise subset (6 samples per subject), and scarf disguise subset (6 samples per subject). We choose all the images from sunglasses disguise subset as the training samples, and the test images belong to the ordinary subset without sunglasses and scarf. Thus, in this experiment, the difference between the test and training samples is very large due to the disguise. Actually, these two types of the samples are derived from the different signal domains. Table 4 indicates the recognition rates of the proposed method versus elimination percentage of redundant samples in different signal domains.

Meanwhile, Fig. 5 shows an example of the first 36 remaining images (all images from sunglasses disguise subset and 720 in total) ordered by their contribution degrees. It shows that the first few remaining samples with higher contribution degrees are all from the same class of the test sample, which leads to correct classification result.

Similarly, we made another experiment on the FERET to show the selection of the best features for all postures of a person by P-SRC. In this experiment, the former two images per individual were chosen for training, and the rest were used for test. Thus, a training set of 400 images and a test set with 1000 images are generated. Figure 6 illustrates an example of the first 20 remaining images (400 in total) ordered by their contribution degrees. It shows that the first remaining samples with highest contribution degree are from the same class (within-class) of the test sample, which also leads to correct classification result.

As described in Sec. 3.5, the contribution can be evaluated by the residual between $\mathbf{y}_r$ and $\mathbf{y}$, i.e., $D_r = \|\mathbf{y} - \mathbf{y}_r\|_2^2$. The D_r can also be viewed as a discrimination measurement between the test sample and the i'th class. Here, we made an experiment on the ORL to show the discrimination of images of different persons by P-SRC. In this experiment, the former four images per individual were chosen for training and the rest were used for test. Thus, a training set of 160 images and a test set with 240 images are generated. For instance, we choose the fifth image of the first subject as the test sample. Table 5 indicates the discrimination measurement (range from 0 to 1) between this test sample and each subject. From Table 5, we can find that the smallest measurement result 0.42 occurs in the first subject (with same class to the test sample), and it should be noted that the value 1 that appears in Table 5 demonstrates the corresponding subjects (classes) have been completely discarded from the training sets.

4.2 *Handwritten Digit Recognition*

In this experiment, the handwritten digit recognition on the widely used USPS database,[45,46] which has 7291 training

Table 4 Recognition rates of progressive sparse representation-based classification (P-SRC) versus elimination percentage of redundant samples in different signal domains.

Elimination rate	90%	80%	70%	60%	50%	40%	30%	20%	10%
Recognition rate	**0.935**	0.923	0.913	0.907	0.911	0.910	0.914	0.915	0.914

Note: The bold value denote the best experimental results compared with other methods.

Fig. 5 An example of the first 36 remaining images (720 in total) ordered by their contribution degrees. The test image is the first sample of first subject. In this experiment, we selected all images from the sunglasses disguise subset (six images per person) as the training samples. It shows that the first few remaining samples with higher contribution degrees are all from the same class (within-class) of the test sample, which leads to correct classification result.

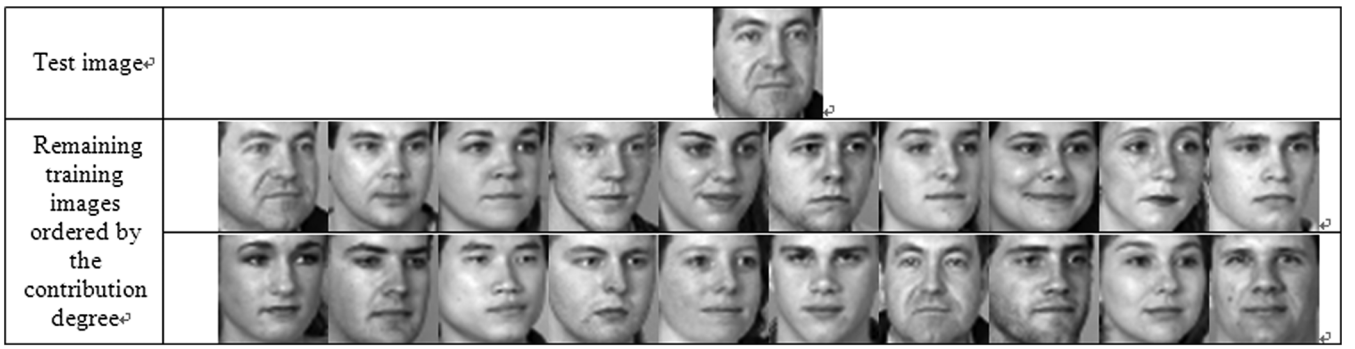

Fig. 6 An example of the first 20 remaining images (400 in total) ordered by their contribution degrees. The test image is third sample from the first subject. In this experiment, the former two images per individual were chosen for training. It shows that the first remaining samples with highest contribution degree are from the same class (within-class) of the test sample, which also leads to correct classification result.

and 2007 test images, is performed. We artificially augmented the training set by shifting the digit images by 1 pixel in every direction. The proposed method is compared with COPAR, JDL, and the handwritten digit recognition methods reported in Refs. 47–49. These methods include the state-of-the-art reconstructive DL methods with linear and bilinear classifier models, which is denoted by REC-L and REC-BL in Ref. 48, the state-of-the-art supervised DL methods with generative training and discriminative training, which is denoted by SDL-G and SDL-D in Ref. 48, the state-of-the-art methods of sparse representation for signal classification, which is denoted by SRSC in Ref. 47 and DLSI in Ref. 49. In addition, the results of some problem-specific methods (i.e., the standard Euclidean KNN and SVM with a Gaussian kernel) reported in Ref. 49 are also listed. Here, the number of atoms in each subdictionary of P-SRC is set to 200. Figure 7 illustrates the part of learned dictionary atoms of digits 5 and 6 and Table 6 indicates the recognition error rates of the proposed P-SRC and the other methods. We can find that the P-SRC outperforms all the competing methods. Meanwhile, it should be noted that the SVM classifier performs classification with a one-versus-all strategy. In comparison, P-SRC interprets a progressive dictionary update rule under a local DCT evaluation, and its classifier in design mode can also be clearly interpreted.

Table 5 The discrimination measurement (range from 0 to 1) between the test sample and each subject on ORL.

Subject	1	2	3	4	5	6	7	8	9	10
Measurement	**0.42**	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subject	11	12	13	14	15	16	17	18	19	20
Measurement	1.00	1.00	0.96	1.00	1.00	0.83	1.00	1.01	1.00	1.00
Subject	21	22	23	24	25	26	27	28	29	30
Measurement	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subject	31	32	33	34	35	36	37	38	39	40
Measurement	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95

Note: The bold value denote the best experimental results compared with other methods.

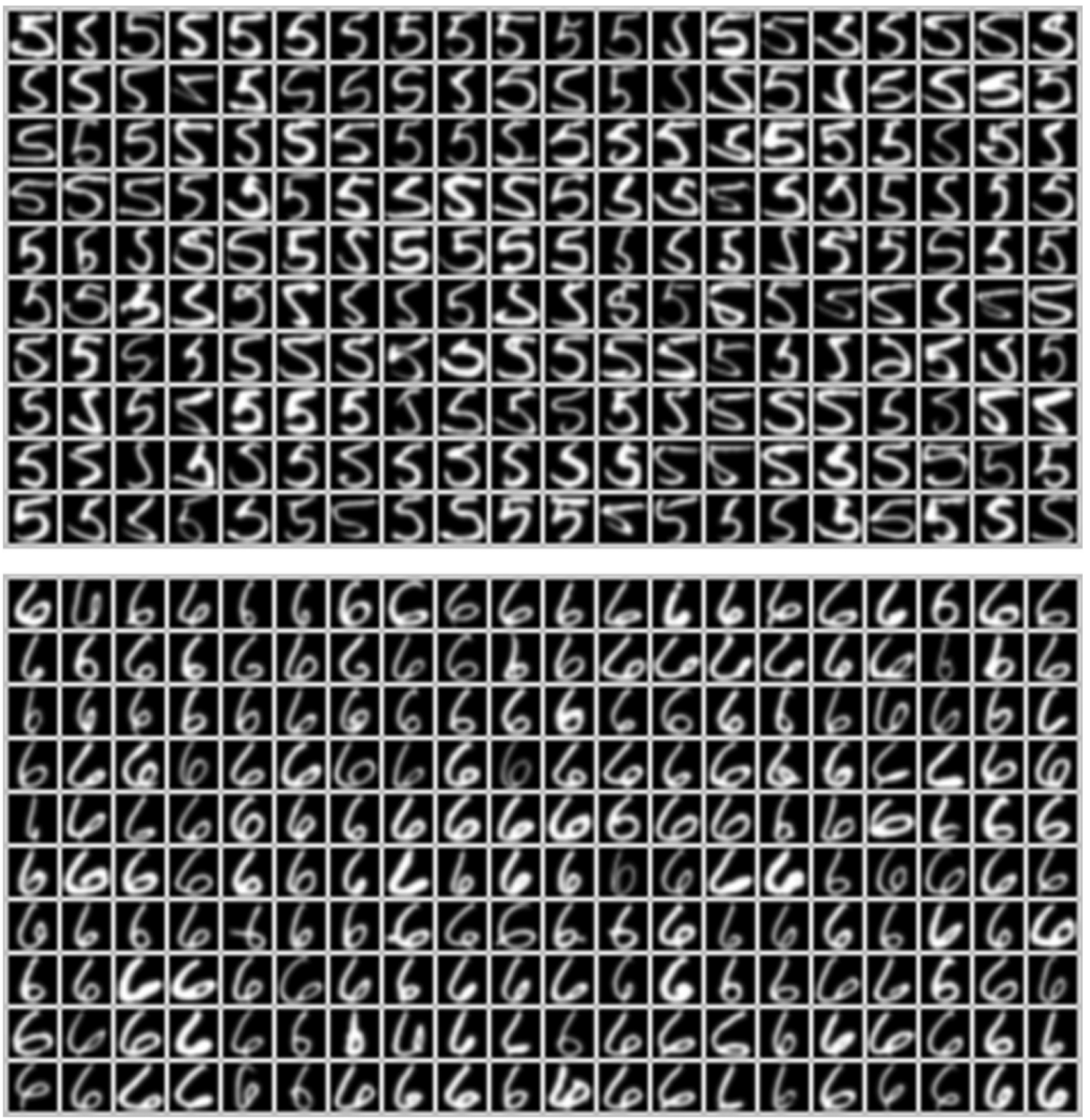

Fig. 7 The part of learned dictionary atoms of digits 5 and 6 by P-SRC.

Table 6 Handwritten digit error rates (%) of competing methods on the USPS dataset.

Method	Error rate
SRC	6.05
REC-L(BL)	6.83 (4.38)
SDL-G(D)	6.67 (3.54)
DLSI	3.98
KNN	5.20
SVM	4.20
COPAR	3.61
JDL	6.08
Proposed method	**2.85**

Note: The bold value denote the best experimental results compared with other methods.

5 Conclusion

This paper developed a progressive sparse representation-based classification (P-SRC) algorithm. The P-SRC aims to exploit an optimal representation of training samples from the classes with major relevant contributions, by which the representation ability of each training sample is exploited to determine some optimal "nearest neighbors" for the test sample. It is noted that the transformed DCT evaluation measures the similarity between the test sample and each local training sample by using cosine distance metrics in the DCT domain. Future work is required to enable such a trend; among the many possible untraditional dictionary research directions, we should notice two: (1) exploration of the connection between the chosen untraditional dictionary update rule in the SRC and the method used later in the application, and (2) a study of the effect of introducing weights to the untraditional training dictionary, allowing them to get varying fuzzy degrees of popularity.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive advice. This work was supported by the National Science Foundation of China (Grant Nos. 61373055 and 61471182), the Natural Science Foundation of Jiangsu Province (Grant Nos. BK2012700 and BK20130473), the Foundation of Artificial Intelligence Key Laboratory of Sichuan Province (Grant No. 2012RZY02), the Open Project Program of the State Key Laboratory of CAD&CG of Zhejiang University (Grant No. A1418), the Foundation of Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University.

References

1. J. Wright et al., "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(2), 210–227 (2009).
2. J. Wright et al., "Sparse representation for computer vision and pattern recognition," *Proc. IEEE* **98**(6), 1031–1044 (2010).
3. E. Cande`s and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?," *IEEE Trans. Inf. Theory* **52**(12), 5406–5425 (2006).
4. E. Cande`s, "Compressive sampling," in *Proc. Int. Congress of Mathematicians*, Vol. 3, pp. 1433–1452, Madrid, Spain (2006).
5. W. Zhang, Z. C. Lin, and X.O. Tang, "Tensor linear Laplacian discrimination (TLLD) for feature extraction," *Pattern Recognit.* **42**, 1941–1948 (2009).
6. D. Zhao et al., "Linear Laplacian discrimination for feature extraction," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–7 (2007).
7. Z. P. Zhang et al., "Hierarchical facial landmark localization via cascaded random binary patterns," *Pattern Recognit.* **48**, 1277–1288 (2015).
8. Y. Sun, X. G. Wang, and X. O. Tang, "Deeply learned face representations are sparse, selective, and robust," (2014) arXiv preprint arXiv:1412.1265.
9. X. Gao et al., "Local face sketch synthesis learning," *Neurocomputing* **71**(10–12), 1921–1930 (2008).
10. C. Deng et al., "Invariant image watermarking based on local feature regions," in *Proc. Intl. Conf. on Cyberworlds 2008 (CW2008)*, pp. 6–10 (2008).
11. T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *Proc. 10th European Conf. on Computer Vision (ECCV)*, pp. 725–738 (2008).
12. W. Bian and D. Tao, "Biased discriminant Euclidean embedding for content-based Image retrieval," *IEEE Trans. Image Process.* **19**(2), 545–554 (2010).
13. T. Zhang et al., "Local coordinates alignment (LCA): a novel manifold learning approach," *Int. J. Pattern Recognit. Artif. Intell.* **22**(4), 667–690 (2008).
14. V. Vural et al., "Using local dependencies within batches to improve large margin classifiers," *J. Mach. Learn. Res.* **10**, 183–206 (2009).
15. Y. Xu, G. Feng, and Y. Zhao, "One improvement to two-dimensional locality preserving projection method for use with face recognition," *Neurocomputing* **73**, 245–249 (2009).
16. Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Networks* **22**(7), 1119–1132 (2011).
17. M. Yang et al., "Relaxed collaborative representation for pattern classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR, pp. 2224–2231 (2012).
18. L. Zhang et al., "Collaborative representation based classification for face recognition," arXiv preprint arXiv:1204.2358.
19. Y. Xu et al., "A two-phase test sample sparse representation method for use with face recognition," *IEEE Trans. Circuits Syst. Video Technol.* **21**(9), 1255–1262 (2011).
20. Y. Xu, W. M. Zuo, and Z. Z. Fan, "Supervised sparse representation method with a heuristic strategy and face recognition experiments," *Neurocomputing* **79**, 125–131 (2012).
21. Y. Xu et al., "Using the idea of the sparse representation to perform coarse-to-fine face recognition," *Inf. Sci.* **238**, 138–148 (2013).
22. Y. Xu et al., "A sparse representation method of bimodal biometrics and palmprint recognition experiments," *Neurocomputing* **103**, 164–171 (2013).
23. W. K. Yang et al., "Image classification using kernel collaborative representation with regularized least square," *Appl. Math. Comput.* **222**, 13–28 (2013).
24. W. Yang, Z. Wang, and C. Sun, "A collaborative representation based projections method for feature extraction," *Pattern Recognit.* **48**(1), 20–27 (2015).
25. Z. H. Liu et al., "A novel classification method for palmprint recognition based on reconstruction error and normalized distance," *Appl. Intell.* **39**, 307–314 (2013).
26. H. Khorrami and M. Moavenian, "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification," *Expert Syst. Appl.* **37**, 5751–5757 (2010).
27. J. J. Wu et al., "Cosine interesting pattern discovery," *Inf. Sci.* **184**, 176–195 (2012).
28. J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Math. Comput. Modell.* **53**, 91–97 (2011).
29. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, New York (1983).
30. T. C Song and H. L. Li, "Local polar DCT features for image description," *IEEE Signal Process. Lett.* **20**(1), 59–62 (2013).
31. X. N. Song et al., "A new sparse representation-based classification algorithm using iterative class elimination," *Neural Comput. Appl.* **24**, 1627–1637 (2014).
32. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006).
33. X. N. Song et al., "A parameterized fuzzy adaptive K-SVD approach for the multi-classes study of pursuit algorithms," *Neurocomputing* **123**, 131–139 (2014).
34. V. Vapnik, *Statistical Learning Theory*, Wiley, New York (1998).

35. T. Mary-Huard, S. Robin, and J.J. Daudin, "A penalized criterion for variable selection in classification," *J. Multivar. Anal.* **98**, 695–705 (2007).
36. T. Heap and F. Samaria, "Real-time hand tracking and gesture recognition using smart snakes," in *Proc. Interface to Human and Virtual Worlds*, pp. 50, Montpellier, France (1995).
37. P. J. Phillips et al., "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(10), 1090–1104 (2000).
38. A. M. Martinez, "The AR face database," *CVC Tech. Rep.*, 24 (1998).
39. L. L. Shen, L. Bai, and M. Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification," *Image Vis. Comput.* **25**(5), 553–563 (2007).
40. C. Gigliarano, S. Figini, and P. Muliere, "Making classifier performance comparisons when ROC curves intersect," *Comput. Stat. Data Anal.* **77**, 300–312 (2014).
41. J. Yang et al., "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(1), 131–137 (2004).
42. X. N. Song et al., "A complete fuzzy discriminant analysis approach for face recognition," *Appl. Soft Comput.* **10**, 208–214 (2010).
43. W. H. Deng, J. N. Hu, and J. Guo, "Extended SRC: undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(9), 1864–1870 (2012).
44. X. Shi et al., "Face recognition by sparse discriminant analysis via joint L2, 1-norm minimization," *Pattern Recognit.* **47**(7), 2447–2453 (2014).
45. J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(5), 550–554 (1994).
46. M. Yang et al., "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.* **109**(3), 209–232 (2014).
47. K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Neural Information and Processing Systems*, pp. 609–616 (2006).
48. J. Mairal et al., "Supervised dictionary learning," in *Proc. Neural Information and Processing Systems*, pp. 1033–1040 (2009).
49. I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3501–3508, IEEE (2010).

Xiaoning Song received his PhD in pattern recognition and intelligence system at Nanjing University of Science and Technology, China, in 2010. He was a visiting researcher in the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, United Kingdom from 2014 to 2015. Presently, he is an associate professor in Jiangnan University, Wuxi, China. His current research interests include pattern recognition, sparse representation, image recognition, and fuzzy systems.

Zhen-Hua Feng is currently a research fellow at the Centre for Vision, Speech and Signal Processing (CVSSP), faculty of engineering and physical sciences, University of Surrey, United Kingdom. He has published more than 10 papers in his fields of research. His current research interests include pattern recognition, sparse representation, automatic face alignment algorithms, including active shape models, active appearance models, and their extensions.

Guosheng Hu received his PhD in pattern recognition and intelligence system at Centre for Vision, Speech and Signal Processing (CVSSP), faculty of engineering and physical sciences, University of Surrey, United Kingdom, in 2015. His current research interests include pattern recognition, sparse representation, and three-dimensional morphable model.

Xibei Yang received his PhD in pattern recognition and intelligence system at Nanjing University of Science and Technology, China, in 2010. Presently, he is an associate professor in the School of Computer Science and Engineering, Jiangsu University of Science and Technology, China. His research interests include knowledge discovery, granular computing, and rough set theory.

Jingyu Yang is currently a distinguished professor in the Department of Computer Science at Nanjing University of Science and Technology. He is the author of more than 400 scientific papers in computer vision, pattern recognition, and artificial intelligence. He has won more than 20 provincial and national awards. His current research interests include pattern recognition, robot vision, image processing, data fusion, and artificial intelligence.

Yunsong Qi received his PhD in pattern recognition and intelligence system at Nanjing University of Science and Technology, China, in 2011. Currently, he is a professor in the School of Computer Science and Engineering, Jiangsu University of Science and Technology, China. His research interests include data mining and machine learning.