Machine Vision and Applications (2018) 29:991–1007
https://doi.org/10.1007/s00138-018-0941-z

SPECIAL ISSUE PAPER

Extended sparse representation-based classification method for face recognition

Yali Peng[1] · Lingjun Li[2] · Shigang Liu[3] · Jun Li[4] · Xili Wang[3]

Received: 1 October 2017 / Revised: 28 April 2018 / Accepted: 6 May 2018 / Published online: 25 May 2018

Abstract
In sparse representation algorithms, a test sample can be sufficiently represented by exploiting only the training samples from the same class. However, due to variations of facial expressions, illuminations and poses, the other classes also have different degrees of influence on the linear representation of the test sample. Therefore, in order to represent a test sample more accurately, we propose a new sparse representation-based classification method which can strengthen the discriminative property of different classes and obtain a better representation coefficient vector. In our method, we introduce a weighted matrix, which can make small deviations correspond to higher weights and large deviations correspond to lower weights. Meanwhile, we improve the constraint term of representation coefficients, which can enhance the distinctiveness of different classes and make a better positive contribution to classification. In addition, motivated by the work of ProCRC algorithm, we take into account the deviation between the linear combination of all training samples and of each class. Thereby, the discriminative representation of the test sample is further guaranteed. Experimental results on the ORL, FERET, Extended-YaleB and AR databases show that the proposed method has better classification performance than other methods.

Keywords Sparse representation · Weighted matrix · Discriminative property · Representation coefficient

1 Introduction

Compared to other biometrics [1–3], face recognition [4–6] has the following characteristics: nonmandatory, noncontact, concurrency. In addition, face recognition is easy to operate and accords with human visual character. The keys to the success of face recognition system are a cutting-edge core algorithm, practical recognition rate and recognition speed. Face recognition system integrates many technologies such as artificial intelligence [7–9], machine recognition [10–12], machine learning [13–15], model theory [16–18], expert system [19–21], video processing [22–24] and, moreover, combines the theory and implementation of intermediate value processing. The realization of its core technology shows the transformation from weak artificial intelligence to strong artificial intelligence.

Numerous face recognition algorithms have emerged. Among them, the typical algorithms include principal component analysis (PCA) [25–27], linear discriminant analysis (LDA) [28–30] and locality preserving projections (LPP) [31–33]. Besides, Roweis et al. [34] proposed an unsupervised learning algorithm, namely locally linear embedding (LLE). Its main principle is that if there is a set of data with nested manifolds, the order of the data points between the nested space and the local field within the low-dimensional space should be maintained. Belkin et al. [35] proposed a nonlinear dimension reduction algorithm, which constructs a representation of data sampled from a low-dimensional manifold embedded in a higher-dimensional space, called as Laplacian eigenmap. Another classical face recognition algorithm called local binary patterns (LBP) has been proposed by Ahonen et al. [36]. This algorithm considers both shape and texture information to represent face images and takes the extraction of local features as identification standards. LBP is insensitive to illumination variation.

✉ Shigang Liu
shgliu@snnu.edu.cn

1 Key Laboratory of Modern Teaching Technology, Ministry of Education, Xi'an 710062, China

2 Engineering Laboratory of Teaching Information Technology of Shaanxi Province, Xi'an 710119, China

3 School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

4 School of Automation, Southeast University, Nanjing 210096, China

Wright et al. [37] proposed a very important approach which applies sparse representation for robust face recognition. Specifically, a test sample is sparsely coded by an over-complete dictionary whose base elements are training samples. Then, the test sample is assigned to a certain class which yields the least coding error. This algorithm, namely sparse representation-based classification (SRC), achieves a great success in face recognition. Naseem et al. [38] proposed linear regression classification (LRC) algorithm which can address the problem of illumination invariant in face recognition. It assumes that a test sample can be represented by a linear combination of the training samples of each class, respectively. Xu et al. [39] proposed a two-phase test sample representation (TPTSR) algorithm. Although TPTSR also is based on the idea of sparse representation, unlike SRC, its first phase removes a great number of training samples that are dissimilar to the test sample and takes the class labels of the remaining training samples as candidate classes for the test sample. This will be very helpful for the second phase to more accurately represent the test sample. Zhang et al. [40] raised questions about the pivotal role of l_1-norm sparsity on improving face recognition performance and presented a collaborative representation classifier (CRC).

All of the above-mentioned algorithms have neglected several key problems. First, in the conventional error loss term, all the errors are treated equally, but the error difference between different classes is ignored, and these differences may be useful for classification. Second, for the constraint term of the representation coefficients, most algorithms tend to focus only on the use of norms, such as the l_1-norm, l_2-norm, while ignoring the possibility that this item can be improved to enhance the discriminability between different classes. Therefore, in this paper, firstly, we introduce weighted matrix into the error term between test samples and its reconstructed values and further enhance the differences between different classes. Secondly, we improve the constraint term of the representation coefficients, while maintaining sparsity, and enhance the discriminant of the algorithm. Finally, in order to stabilize the target solution, we added one item to the objective function. Specifically, the deviations between the linear combination of all training samples and of each class are taken into account. Experiments show that the proposed method can achieve a very satisfactory accuracy for face recognition. Our method has the following advantages. (1) The weighted matrix is introduced into the objective function. Smaller deviations are assigned greater weight; larger deviations are assigned lesser weight. (2) The improvement of representation coefficients constraint term enhances the discrimination of different classes. (3) The weighted matrix is not fixed. Instead, the weighted matrix and representation coefficients are updated iteratively, until convergence. Hence, the continuous updating of weights reflects the flexibility of the proposed method.

The remainder of this paper is organized as follows. Section 2 presents the proposed method. Section 3 describes the underlying rationale of the proposed algorithm. Section 4 shows experimental results. Section 5 provides the conclusion of the paper.

2 Related work

Because our proposed algorithm is based on the sparse representation algorithm, we studied a large number of related improved sparse representation algorithms, and our algorithm is also inspired by these algorithms. Therefore, this section mainly introduces these algorithms and gives our own understanding about these algorithms.

Deng et al. [41] proposed an extended sparse representation-based classifier (ESRC) algorithm, which can extend SRC to applications where there is a single training sample (or very few training samples) per class. In ESRC, there is a basis matrix that represents the universal intraclass variant bases. These variations usually caused by exaggerated expressions, occlusions or unbalanced lighting changes; we can obtain these variant bases by subtracting the natural image from other images of the same class. Finding a sparse representation of the test image in terms of the training set and the intraclass variant bases, the object function of ESRC is

$$\begin{bmatrix} \hat{x}_1 \\ \hat{\beta}_1 \end{bmatrix} = \arg\min \left\| \begin{bmatrix} x \\ \beta \end{bmatrix} \right\|_1, \quad \text{s.t.} \left\| [\mathbf{A}, \mathbf{D}_I] \begin{bmatrix} x \\ \beta \end{bmatrix} - y \right\|_2 \leqslant \varepsilon, \tag{1}$$

where $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_k] \in \mathbf{R}^{d \times n}$ stands for a training samples matrix, and k denotes the number of classes. $\mathbf{D}_I \in \mathbf{R}^{d \times p}$ is a matrix of intraclass variant bases. $y \in \mathbf{R}^d$ is a test sample, and $\varepsilon > 0$ is an optimal error tolerance. $x, \hat{x} \in \mathbf{R}^n$ and $\beta, \hat{\beta} \in \mathbf{R}^p$. Then computing the residuals

$$r_i(y) = \left\| y - [A, D_I] \begin{bmatrix} \delta_i(\hat{x}_1) \\ \hat{\beta}_1 \end{bmatrix} \right\|_2. \tag{2}$$

Finally, the label of the test sample is $Identity(y) = \arg\min_i r_i(y)$. The ESRC algorithm can achieve higher performance with a smaller number of bases.

Tang et al. [42] designed an weighted group sparse representation classification (WGSRC). Each class usually plays a different role in representing a test sample. WGSRC assigns each class weights according to the similarity between a test sample and training samples of each class. For representing a test sample, the training samples from the neighbors and the highly relevant classes of it are taken into account. In the WGSRC algorithm, more structure information and discriminative information are considered. In WGSRC, there is a $l_{2,1}$-norm regularization function

$$\beta'^{*} = \min_{\beta} \left\| y - \mathbf{X}\mathbf{S}^{-1}\beta \right\|_2^2 + \lambda \sum_{i=1}^{c} \|\beta_i\|_2, \tag{3}$$

where $\mathbf{X}$ is training samples matrix, and y is a test sample. The matrices $\mathbf{S} = \text{diag}([s_1, s_2, \ldots, s_c])$, and $s_i = \left[s_{i1}, s_{i2}, \ldots, s_{in_i}\right]^{\mathrm{T}}$, $s_{jk} = w_j d_{jk}$, $d_{jk} = \exp\left(\frac{\|y - x_{ik}\|_2}{\sigma_2}\right)$ $(i = 1, 2, \ldots, c;\ j = 1, 2, \ldots, n_i)$ are used to assess the relative importance of training samples per class for representing a test sample. According to Eq. (3), compute the sparse coefficient $\beta^* = \mathbf{S}^{-1}\beta'^{*}$. Then, the label of a test sample is Identity $(y) = \arg\min_i \left\| y - \mathbf{X}_i \beta_i^* \right\|_2$.

Timofte et al. [43] improved collaborative representations with regularized least squares and proposed a weighted collaborative representation classifier (WCRC). Because the training samples are not equally discriminative in classification, the training samples or features are weighted. Wu et al. [44] proposed a learned collaborative representation-based classifier (LCRC) and attempted to explain why such choice of weights works and how to optimize those weights in WCRC. Xu et al. [45] proposed a new discriminative sparse representation method for robust face recognition via l_2 regularization. The objective function of this method is defined as

$$\min_{\mathbf{B}} \|y - \mathbf{X}\mathbf{B}\|^2 + \gamma \sum_{i=1}^{L} \sum_{j=1}^{L} \left\| \mathbf{X}_i \mathbf{B}_i + \mathbf{X}_j \mathbf{B}_j \right\|^2, \tag{4}$$

where $\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_L]$ is representation coefficient, $B_i = [b_{m(i-1)+1}, \ldots, b_{mi}]^{\mathrm{T}}$ $(i = 1, 2, \ldots, L)$. γ is a positive constant. Representation coefficient $\mathbf{B}$ is calculated by using $\mathbf{B} = \left((1 + 2\gamma)\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\gamma L\mathbf{M}\right)^{-1}\mathbf{X}^{\mathrm{T}}y$, where $\mathbf{M} = \begin{bmatrix} X_1^{\mathrm{T}}X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_L^{\mathrm{T}}X_L \end{bmatrix}$. According to $p = \arg\min_i \|\mathbf{X}_i\mathbf{B}_i - y\|_2^2$, test sample y is classified to the p-th class.

Cai et al. [46] analyzed the algorithm theory of CRC and proposed a probabilistic collaborative representation-based classifier (ProCRC). The object function is

$$(\hat{\alpha}) = \arg\min_{\alpha} \left\{ \|y - \mathbf{X}\alpha\|_1 + \lambda \|\alpha\|_2^2 + \frac{\gamma}{K} \sum_{k=1}^{K} \|\mathbf{X}\alpha - \mathbf{X}_k\alpha_k\|_2^2 \right\}. \tag{5}$$

Solving Eq. (5), $\hat{\alpha}$ can be obtained, i.e., $\hat{\alpha} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{\gamma}{K}\sum_{k=1}^{K}\left(\bar{\mathbf{X}}'_k\right)^{\mathrm{T}}\bar{\mathbf{X}}'_k + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}y$, where $\bar{\mathbf{X}}'_k = \mathbf{X} - \mathbf{X}'_k$, and $\mathbf{X}'_k = [0, \ldots, \mathbf{X}_k, \ldots, 0]$. According to $j = \arg\min_k \left\|\mathbf{X}_k\delta_k\left(\hat{\alpha}\right) - y\right\|_2^2$, test sample y is classified to the j-th class.

3 Proposed method

For convenience of the following description of algorithms, here we normalize mathematical notations and expressions. Assume that there are c known classes. Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_c]$ be a set of d-dimensional training samples from c classes, where $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_c]$ is the training samples of class i, i.e., $\mathbf{X}_i = \left[x_{(i-1)n+1}, x_{(i-1)n+2}, \ldots, x_{in}\right]$, $x_{(i-1)n+1}, x_{(i-1)n+2}, \ldots, x_{in}$ are column vectors. The total number of training samples of each class is n. Then the total number of training samples for all classes is N, i.e., $N = nc$. Column vector y denotes a test sample.

3.1 Sparse representation-based classification

Sparse representation is referred to as compressed sensing theory and has been widely applied to various areas of signal processing. As a special kind of signal, face images have sparse characteristic in many cases. Therefore, the introduction of sparse representation theory to face recognition has become a research hot spot. In all sparse representation algorithms, on the premise of an over-complete input dictionary, these algorithms select a small amount of atoms in the dictionary to represent a signal y and enable representation coefficients vector to achieve sparse. It should be noted that the basic elements of the dictionary are called as atoms. In face recognition, face images are regarded as the basic elements of the dictionary (i.e., atoms), and the signal y refers to a test sample. In other words, an arbitrary test sample y can be represented by a linear combination of all the training samples, i.e.,

$$\begin{aligned} y &= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \cdots + \mathbf{X}_c\beta_c \\ &= x_1 b_1 + \cdots + x_{(i-1)n+1} b_{(i-1)n+1} + \cdots + x_{in} b_{in} + \cdots + x_N b_N \\ &= \mathbf{X}\beta, \end{aligned} \tag{6}$$

where $\beta = \left[b_1, \ldots, b_{(i-1)n+1}, \ldots, b_{in}, \ldots, b_N\right] \in \mathbf{R}^N$ is representation coefficients vector. Representation coefficients vector β can be obtained by solving the following formula,

$$\min_{\beta} \|y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_p, \tag{7}$$

where λ is a constant. Its role is to balance the contribution of reconstruction error and representation coefficients vector, meanwhile, to make the least square solution stable. After getting the optimal coefficients vector, the test sample can be, respectively, reconstructed by exploiting the training samples of each class, i.e., $y = \mathbf{X}_i\hat{\beta}_i$, where $\hat{\beta}_i$ is the representation coefficients vector associated with class

i ($i = 1, 2, \ldots, c$). Finally, the test sample y can be classified into class k with minimum reconstruction error, i.e.,

$$k = \arg\min_i \left\| y - \mathbf{X}_i \hat{\beta}_i \right\|_2. \tag{8}$$

Moreover, it needs to be pointed out that different p norm constraints represent different algorithms in Eq. (7). When the value of p is 1, the corresponding algorithm is sparse representation-based classification (SRC) algorithm. Initially, in order to seek the sparsest solution of $y = \mathbf{X}\beta$, l_0-norm is used to regularize representation coefficient vector β. l_0-norm can count the number of nonzero elements in coefficients vector. However, according to the literature [47], solving the sparsest solution of an linear optimization equations is NP-hard. Fortunately, it has been certified that the solution acquired by exploiting l_0-norm minimization is equal to the solution of l_1-norm minimization [48–50]. If $p = 2$, i.e., $\min_\beta \|y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$, this is classical collaborative representation-based classification (CRC) algorithm which makes l_2-norm as the constraint of representation coefficients vector. The coefficients vector of CRC is not inclined to absolute zero and does not have sparse theoretically. In other words, the coefficients vector obtained using the l_2-norm regularization are not as sparse as those obtained using the l_1-norm regularization. But it is not necessary to use the strong l_1-norm. By using the much weaker l_2-norm regularization, we can have similar classification results but with lower complexity. Moreover, the effectiveness of CRC algorithm is also reflected in the final classification process. The reconstruction error of each class $e_i = \left\| y - \mathbf{X}_i \hat{\beta}_i \right\|_2^2$ can be used for classification in SRC. In fact, $e_i = \left\| y - \mathbf{X}_i \hat{\beta}_i \right\|_2^2$ can be readily derived that

$$e_i = \left\| y - \mathbf{X}_i \hat{\beta}_i \right\|_2^2 = \|y - \hat{y}\|_2^2 + \left\| \hat{y} - \mathbf{X}_i \hat{\beta}_i \right\|_2^2, \tag{9}$$

where $\hat{y}$ is the reconstruction approximation of the test sample y.

$e_i^* = \left\| \hat{y} - \mathbf{X}_i \hat{\beta}_i \right\|_2^2$ plays a major role in classification. According to the geometric interpretation of this formula $e_i^* = \left\| \hat{y} - \mathbf{X}_i \hat{\beta}_i \right\|_2^2$ in the literature [40], if y belongs to the i-th class, and collaborative representation is applied to the linear representation of the test sample y. The angle between $\hat{y}$ and $\mathbf{X}_i \hat{\beta}_i$ will be small, and the angle between $\mathbf{X}_i \hat{\beta}_i$ and $\sum_{j \neq i} \mathbf{X}_j \hat{\beta}_j$ will be big. Such a double detection makes $e_i = \left\| y - \mathbf{X}_i \hat{\beta}_i \right\|_2^2$ used in classification more reliable. In addition, the sparsity of $\left\| \hat{\beta}_i \right\|_2$ can contribute some discriminant information for classification. Hence, the final classification rule of CRC is $k = \arg\min_i \frac{\left\| y - \mathbf{X}_i \hat{\beta}_i \right\|_2}{\left\| \hat{\beta}_i \right\|_2}$, and then the test sample y is classified to the k-th class.

3.2 Description of the proposed method

In fact, image features (e.g., pixels) have different contributions to image classification. Furthermore, due to variations of facial expressions, illuminations and poses, the pixels in the same region of the identical human face vary widely. These variations can be called generalized noises. In conventional sparse representation-based classification algorithms, the first term of the objective function is the deviation between a real sample (i.e., test sample) and a sparse linear combination of training samples. This deviation can be viewed as the noises mentioned above. The existence of noise means that it is impossible to precisely express the test sample as a sparse linear combination of training samples. The deviation can be coarsely explained as a sum of difference between the test sample and each class. However, in a conventional sparse representation algorithm, difference between the test sample and each class is treated equally, which weakens the distinctiveness of different classes. It is useful for image classification methods to enhance the difference of different classes based on a prior knowledge. Thus, we adopt a weighted least square algorithm. In addition, we also optimize the constraint term of the coefficients vector. Moreover, motivated by ProCRC, we take the deviation between the linear combination of all training samples and of each class into account. The aim of our method is to strengthen the discriminant property of different classes and obtain an optimal representation coefficients vector. The objective function of the proposed method is defined as

$$\begin{aligned} \min_\beta \frac{1}{2}(y - \mathbf{X}\beta)^{\mathrm{T}} \mathbf{W} (y - \mathbf{X}\beta) \\ + \gamma \sum_{i=1}^{c} \sum_{j=1}^{c} \beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j + \lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i \beta_i\|_2^2, \end{aligned} \tag{10}$$

where γ and λ are positive constants and are used to balance the three terms in the objective function of the proposed method. $\mathbf{W}$ is a weighted matrix. y is a test sample. It is also a column vector. β denotes coefficients vector. The above objective function is a convex function. Therefore, we can exploit the derivative of the function to get the extremum of the function. Because the convex function can avoid falling into local extremum, we can obtain an optimal solution. There are two unknown variables in the objective function, namely coefficient vector β and weighted matrix $\mathbf{W}$. For improving the computing efficiency, the weighted matrix is set as

$$\mathbf{W}(i,i) = 1\big/|\mathbf{X}(i,:)\,\beta - y_i|, \tag{11}$$

where y_i stands for the i-th row of y. $\mathbf{X}(i,:)$ denotes the i-th row of $\mathbf{X}$. In this way, it allows that small deviations are given greater weights, and large deviations are given smaller weights. Because small deviations mean that the corresponding elements have the key information which is useful for the classification, giving large weights allows key information to be strengthened. Likewise, the information which is useless to the classification will be artificially weakened. Hence we only need to calculate the derivative with respect to β. (Please refer to "Appendix 1" for the proof.) So the derivative over β of $\frac{1}{2}(y-\mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(y-\mathbf{X}\beta) + \gamma\sum_{i=1}^{c}\sum_{j=1}^{c}\beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j + \lambda\sum_{i=1}^{c}\|\mathbf{X}\beta-\mathbf{X}_i\beta_i\|_2^2$ is

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\left(\frac{1}{2}(y-\mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(y-\mathbf{X}\beta)+\gamma\sum_{i=1}^{c}\sum_{j=1}^{c}\beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j\right.\\
&\qquad\left.+\lambda\sum_{i=1}^{c}\|\mathbf{X}\beta-\mathbf{X}_i\beta_i\|_2^2\right)\\
&=-\mathbf{X}^{\mathrm{T}}\mathbf{W}(y-\mathbf{X}\beta)+2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta\\
&\quad-2\gamma\mathbf{M}\beta+2\lambda\left[\sum_{i=1}^{c}(\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i\right]\beta n.
\end{aligned}\tag{12}$$

Based on the characteristic of convex function and the extremum of one variable function, we can get the optimal value of β is

$$\hat{\beta}=\left[\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}+2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X}-2\gamma\mathbf{M}+2\lambda\sum_{i=1}^{c}(\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}y. \tag{13}$$

By summarizing the above analysis and description, we present the main steps of the proposed method in Algorithm 1.

Algorithm 1

1: Normalize the columns of $\mathbf{X}$ to have unit l_2-norm.
2: $\mathbf{W}$ (i.e., $\mathbf{W}(i,i)=1\big/|\mathbf{X}(i,:)\,\beta-y_i|$) and $\hat{\beta}$ are updated iteratively, until converge.
3: Compute the residuals $r_i(y) = \left\|y - X_i\hat{\beta}_i\right\|_2$, where $\hat{\beta}_i$ is the coefficient vector associated with class i.
4: Output the identity of y as $\mathrm{Identity}(y) = \arg\min_i r_i(y)$.

4 Theoretical analysis of the proposed method

In order to make our method easier to understand, this section analyzes the rationales and advantages of the proposed method.

4.1 Rationales of the proposed method

The purpose of face recognition is to identify the class of the test sample by exploiting the identification algorithm. Thus, it is necessary to use the algorithm to perform training before recognition. In this process, a large number of training samples are needed, and these training samples are from different classes. In general, the contribution of each training sample in classification has difference. In traditional sparse representation algorithms, the residuals between a test sample and the linear representation of the training samples of each class are different. It may lead to the existence of heteroscedasticity in the algorithmic model. The so-called heteroscedasticity means that the dispersion degree of a test sample y around the regression line $y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \cdots + \mathbf{X}_c\beta_c$ varies with samples. It is pointed out that if the algorithmic model is proved to be heteroscedastic, developing a new algorithmic model is necessary. The weighted least squares algorithm is one of the most frequently used algorithms. In the least square algorithm, each residual is treated equally. But based on prior knowledge, it is expected that great contributions correspond to higher weights, and less contributions have lower weights. It is the reason why we introduce a weighted matrix in our objective function. In our method, the determination of the weighted matrix is adaptive to (dependent on) the test sample. Moreover, the samples are depicted by pixels whose values are different from each other. The pixel features in the face region are more helpful in recognition than those on the background or edge regions. Hence, for simplifying the calculation, we define the weighted matrix as $\mathbf{W}(i,i) = 1\big/|\mathbf{X}(i,:)\,\beta - y_i|$. The size of $\mathbf{W}$ is $d \times d$, where d is the dimension of a sample.

Wright et al. mentioned that SRC inherently possesses discrimination. In other words, the sparse representation algorithm can select the class whose the linear representation is closest to the real test sample, and can eliminate the candidate classes which cannot compactly represent the test sample. However, it is not enough to solely rely on the natural discrimination of the sparse representation algorithm. So, further enhancing the differences of different classes is very useful. Our method can strengthen the discrimination capabilities of all classes, which are beneficial to obtain discriminative sparse code β and interclass difference, also beneficial to better classify the test sample. In our objective function, the term $\gamma\sum_{i=1}^{c}\sum_{j=1}^{c}\beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j$ plays an important role on enhancing discrimination capabilities for all classes.

Specifically, $\gamma \sum_{i=1}^{c} \sum_{j=1}^{c} \beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j$ can be rewritten as $\gamma \sum_{i=1}^{c} \sum_{j=1}^{c} \beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j = \gamma \sum_{i=1}^{c} \sum_{j=1}^{c} (\mathbf{X}_i \beta_i)^{\mathrm{T}} \mathbf{X}_j \beta_j$. Further, $(\mathbf{X}_i \beta_i)^{\mathrm{T}} \mathbf{X}_j \beta_j$ can also be expressed as $(\mathbf{X}_i \beta_i)^{\mathrm{T}} \mathbf{X}_j \beta_j = \|\mathbf{X}_i \beta_i\|_2 \left\|\mathbf{X}_j \beta_j\right\|_2 \cos\theta$, where θ is the angle between $\mathbf{X}_i \beta_i$ and $\mathbf{X}_j \beta_j$. Therefore, minimizing $\beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j$ is equivalent to the minimization of $\|\mathbf{X}_i \beta_i\|_2 \left\|\mathbf{X}_j \beta_j\right\|_2 \cos\theta$, i.e., $\min\left(\beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j\right) = \min\left(\|\mathbf{X}_i \beta_i\|_2 \left\|\mathbf{X}_j \beta_j\right\|_2 \cos\theta\right)$. According to the properties of the cosine function, the smaller the cosine function $\cos\theta$ is, the greater the angle θ is. This can reduce the correlation of the linear representations of the test sample from different classes. Thus, there is a maximum difference between the representation results of the i-th class and the j-th class, which can enhance discrimination capabilities for representation results of different classes.

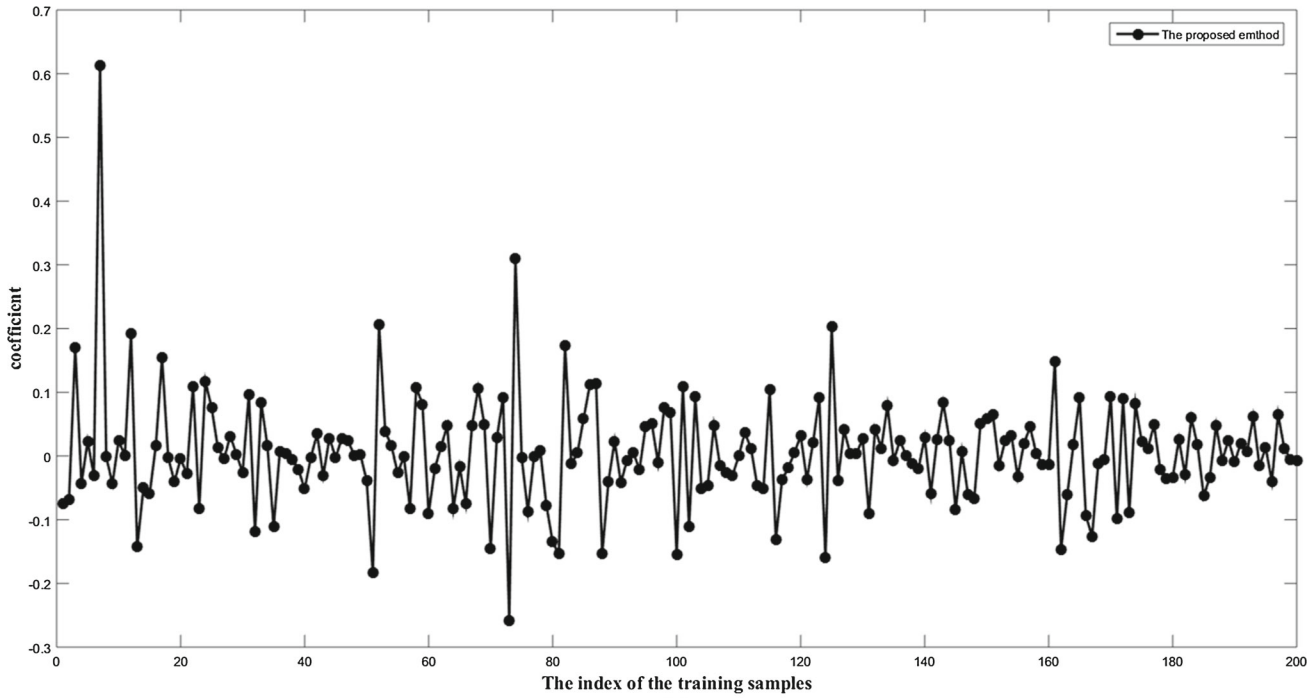
The convex function is a kind of widely used special function. An important property of the convex function is that an extreme small value of the convex function is also a minimum value, and the local minimum value is the global minimum value. The objective function constructed in this paper enables complex problems to be readily solved. Before exploiting the properties of convex function to solve the minimum value, it is necessary to prove that our objective function is a convex function. Fortunately, our objective function satisfies the condition, as we proved in "Appendix 2."

4.2 Advantages of the proposed method

Our method uses the ideas of discriminative and weighted matrix, which are also exploited by other algorithms such as PCRC [51], weighted group sparse representation (WGSR) and ProCRC. Even so, our proposed method still has its own uniqueness. Firstly, PCRC and WGSR put emphasis on applying weighted matrix into sparse constrain term, but ignore reconstruction error term. By comparison, our proposed method is more elaborate. Besides, PCRC just regards the distance between each training sample and the test sample as weight value, while our proposed method not only improves the constrain term of the coefficients vector, but also introduces the weighted matrix into the reconstruction error term to ensure each residual can be treated differently. Moreover, the subsequent experimental results show that our method is more efficient than the PCRC. Secondly, our method enhances the discrimination between different classes by enlarging the angle between the reconstructed vectors of every two classes. We have described the concrete implementation principle in detail in Sect. 4.1. Thirdly, three terms of the our objective function (i.e., $\lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i \beta_i\|_2^2$) are the same with that of ProCRC ($(\hat{\alpha}) = \arg\min_\alpha \|y - \mathbf{X}\alpha\|_1 + \lambda \|\alpha\|_2^2 + \eta \sum_{k=1}^{C} \|\mathbf{X}\alpha - \mathbf{X}_k \alpha_k\|_2^2$). They both aim to further guarantee to obtain a stable coefficients vector. The difference between our proposed method and ProCRC is the first two terms. Especially on the constraint term of coefficients vector (i.e., $\|\alpha\|_2^2$), by improving this term, we achieve the goal of enhancing the discrimination between different classes.

The representation coefficients β produced by sparse representation algorithm can reflect the importance of each training sample for expressing the test sample. Training samples from the same class as the test sample will contribute greatly. On the contrary, training samples from other classes only make a small contribution. If maximizing this contribution difference, it is helpful for sparse representation algorithm to obtain better recognition result. The proposed method is committed to expand the contribution difference of different classes, make different classes more discriminative and enhance the discriminability of the representation coefficients β. Figures 1, 2, 3 and 4 present the representation coefficients and the residuals between the approximate linear representation of the test sample generated from each class and the test sample, respectively. Here we adopt our method, PCRC and CRC to make comparison. From Figs. 1a and 3a, we can observe that when using our method, the maximum coefficient is about 0.63, the closest coefficient value to it is 0.3, and the difference between them is 0.33. In contrast, the maximum coefficient value obtained using CRC is less than 0.15, the closest coefficient value to it is 0.06, and the difference between them is about 0.09. Similarly, from Figs. 2a and 4a, the maximum coefficient value obtained using our method is about 2.3, the closest coefficient value to it is about 0.5, and their difference is about 1.8. While the maximum coefficient value obtained using PCRC is about 0.24, the closest coefficient value to it is about 0.14, and their difference is about 0.1. Hence, we can see that our method can assign higher weights to the class with great contribution and assign lower weights into the classes with less contribution and thereby widen the difference between them. Then, we can see that the minimum residual corresponding to the class is the true class of the test sample from Figs. 1b, 2b, 3b and 4b. Moreover, for our method, the residuals between the test sample and each class tend to be stable except for the true class of the test sample. However, the residuals between the test sample and each class fluctuate greatly when CRC and PCRC are used. This phenomenon illustrates that our method weakens the effects of other classes on the test sample, and reinforces the difference between the correct class and the other classes.

The relationship between the reconstructed test sample (i.e., $\mathbf{X}\beta$) and each class (i.e., $\mathbf{X}_i \beta_i$, $i = 1, \ldots, c$) should be considered. Because the connection between things is often multifaceted, a change in the dependent variable (i.e., $\mathbf{X}\beta$) may be affected by several other independent variables (i.e., $\mathbf{X}_i \beta_i$, $i = 1, \ldots, c$). Therefore $\mathbf{X}\beta = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \cdots + \mathbf{X}_c\beta_c$ can be regarded as a regression model. $\mathbf{X}\beta - \mathbf{X}_i\beta_i$ denotes the deviation between the reconstructed test sample and the linear representation of each class. Minimizing



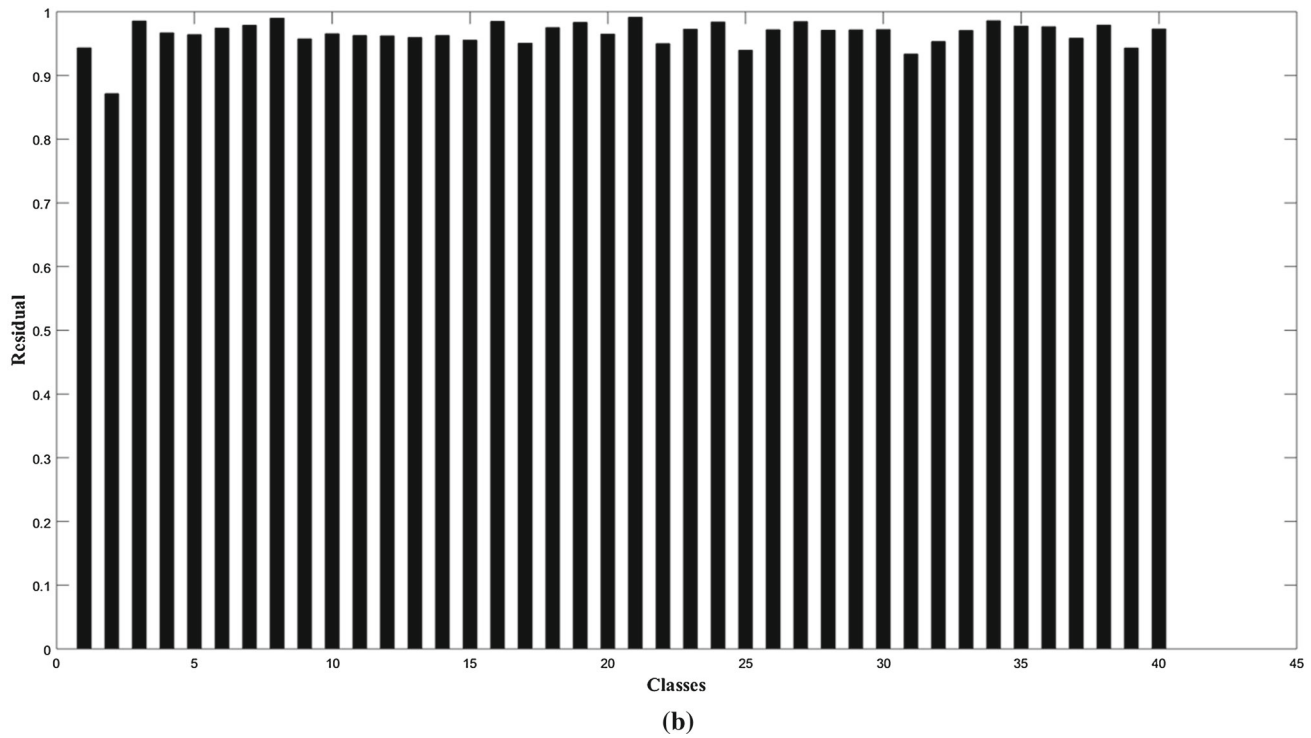


Fig. 1 **a** Representation coefficients on the first test sample from the second class obtained using the proposed method on the ORL face database. The first five face images of each subject are used for training samples, and the rest images are used for test samples. **b** The residuals between the approximate linear representation of the test sample generated from each class and the test sample

these deviations explains a phenomenon that Figs. 1b and 2b show less fluctuation in comparison with Figs. 3b and 4b. Our method can not only enhance the discriminative information in the representation coefficients β, but also weaken the influence of each class on the test sample.

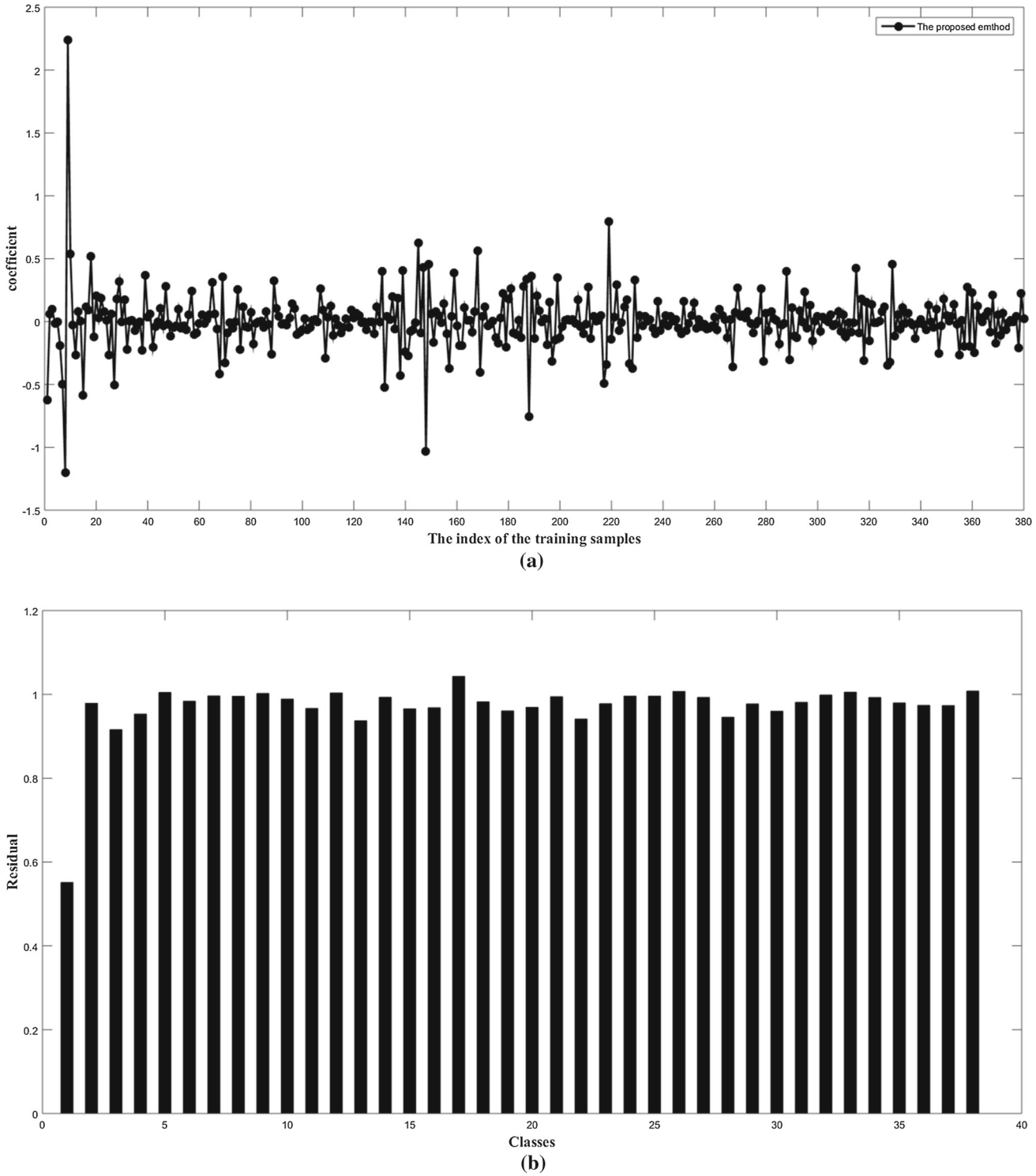


Fig. 2 **a** Representation coefficients on the sixth test sample from the first class obtained using the proposed method on the Extended-YaleB face database. The first ten face images of each subject are used for training samples, and the rest images are used for test samples. **b** The residuals between the approximate linear representation of the test sample generated from each class and the test sample

As for computational complexity, we assume that there are R test samples in all. The dimension of each sample is d. The main computational load of CRC is $\rho = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}y$. The computational complexity of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is $O\left(dN^2\right)$. Let $\mathbf{H}_1 = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}$ and $\mathbf{H}_1$ is a N by N matrix; hence, the computational complexity of $(\mathbf{H}_1)^{-1}$ is

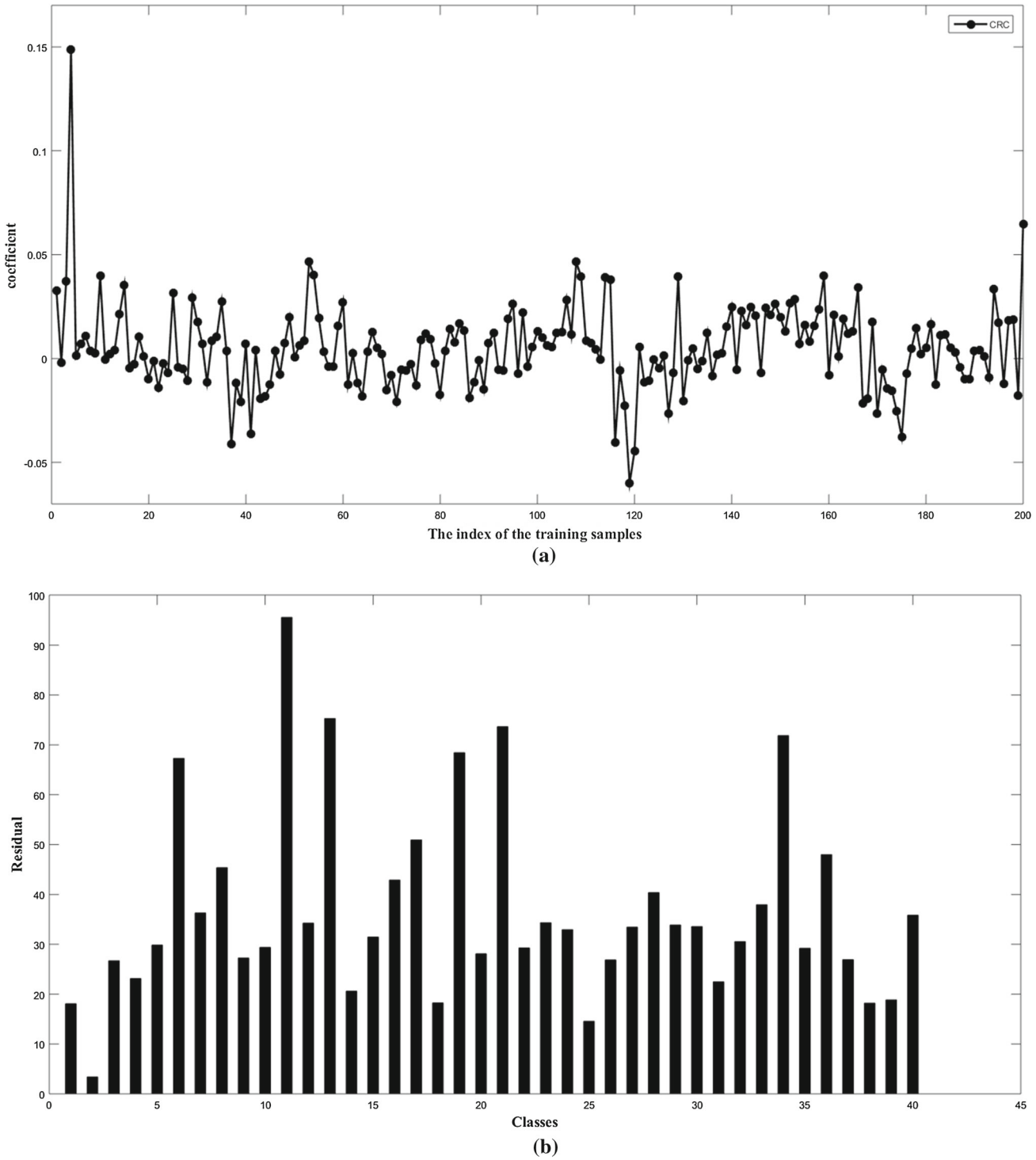


Fig. 3 **a** Representation coefficients on the first test sample from the second class obtained using CRC on the ORL face database. The first five face images of each subject are used for training samples, and the rest images are used for test samples. **b** The residuals between the approximate linear representation of the test sample generated from each class and the test sample

$O\left(N^3\right)$; the computational complexity of $\mathbf{H}_2 = (\mathbf{H}_1)^{-1}\mathbf{X}^{\mathrm{T}}$ is $O\left(dN^2\right)$. Then, CRC has a computational complexity of $O\left(dN^2 + N^3 + dNR\right)$. Similarly, the computational complexity of PCRC is $O\left(dN^2 + N^3 + N^2 + Nd^2 + dNR\right)$. Next, we analyze the computational complexity of our proposed method. Due to the existence of iterative operation,

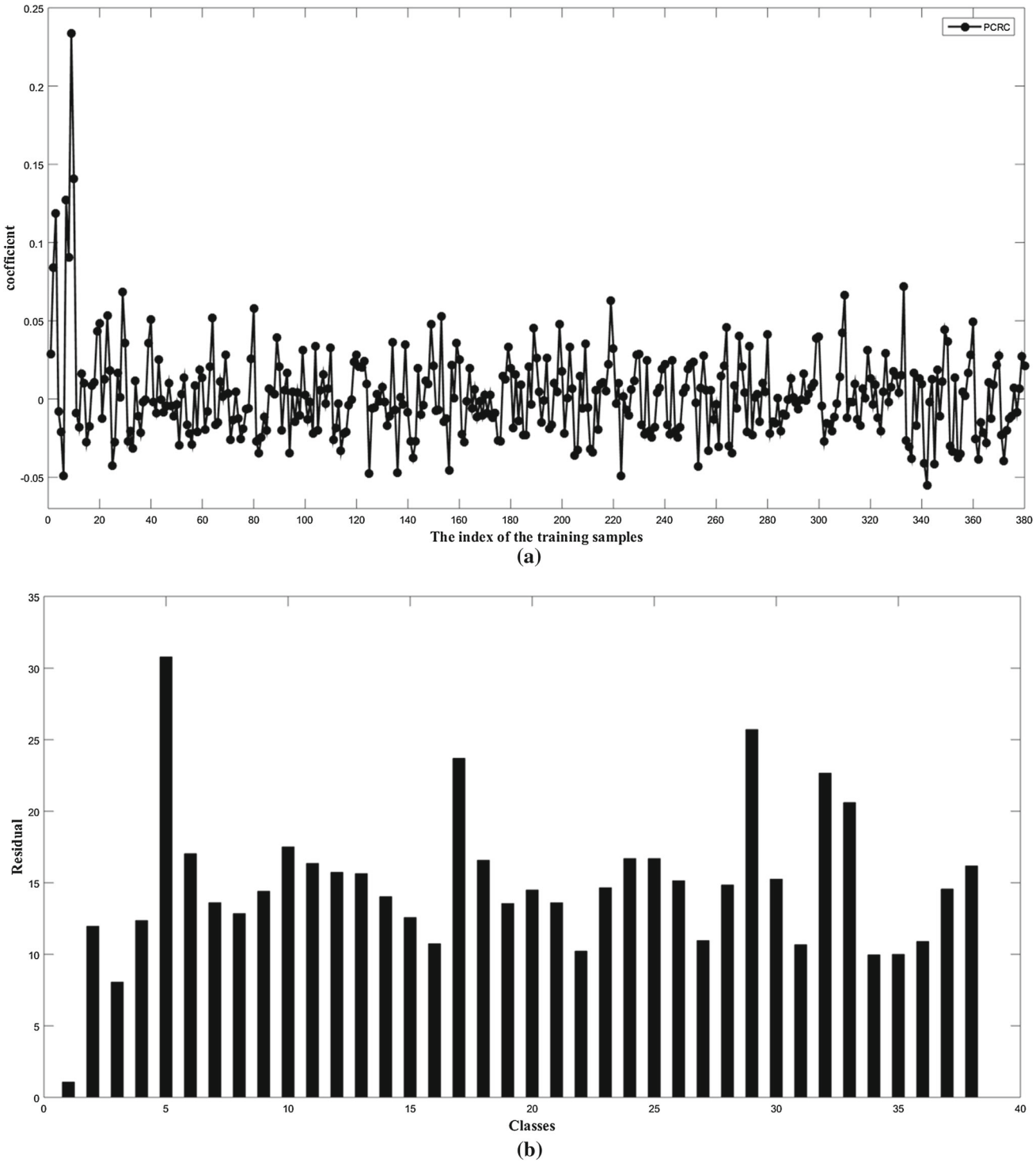


Fig. 4 **a** Representation coefficients on the sixth test sample from the first class obtained using PCRC on the Extended-YaleB face database. The first five face images of each subject are used for training samples, and the rest images are used for test samples. **b** The residuals between the approximate linear representation of the test sample generated from each class and the test sample

we assume that the number of iteration is T. Our proposed method mainly calculates vector $\hat{\beta} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + 2\gamma \mathbf{X}^{\mathrm{T}}\mathbf{X} - 2\gamma\mathbf{M} + 2\lambda \sum_{i=1}^{c} (\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i\right]^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{W}y$.

Let $\mathbf{P} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + 2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X} - 2\gamma\mathbf{M} + 2\lambda \sum_{i=1}^{c} (\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i$. After we calculate $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, we can directly obtain $\mathbf{M}$ and $\sum_{i=1}^{c} (\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i$, and no extra computational complexity is

Fig. 5 Some face images in the ORL database

Table 1 The comparative recognition rates of the ORL database with the number of training samples per class increases

Training samples	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)
Proposed	73.89	87.81	90.00	94.58	95.00	96.88
CRC	71.94	84.06	86.79	91.67	90.50	92.50
KRBM	68.06	80.94	84.29	90.00	88.00	91.87
INNC	72.22	81.87	83.57	89.17	89.50	91.25
CFKNNC	72.22	81.56	82.86	89.17	88.50	92.50
NFRBC	73.33	87.50	91.07	90.42	91.00	93.13
SRC(L1LS)	66.67	82.81	85.71	90.00	88.50	87.50
MI-SRC	73.06	83.13	88.21	87.92	89.50	90.63
LRC	67.50	79.37	81.79	86.25	88.00	94.37
PCRC	67.78	81.56	83.57	86.67	89.00	91.87

Table 2 The comparative recognition rates of the FERET database with the number of training samples per class increases

Training samples	1 (%)	2 (%)	3 (%)	4 (%)
Proposed	50.67	66.80	58.37	63.00
CRC	42.50	57.60	48.38	57.83
KRBM	37.58	47.80	37.12	41.33
INNC	43.50	58.30	50.50	57.33
CFKNNC	49.33	63.40	57.13	60.17
NFRBC	45.50	63.60	58.05	62.33
SRC(L1LS)	32.25	61.60	26.62	32.00
PCRC	50.62	59.00	43.75	57.17

needed. But $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}$ needs extra computational complexity, i.e., $O\left(dN^2 + Nd^2\right)$. Because $\mathbf{P}$ is an N by N matrix, the computational complexity of $(\mathbf{P})^{-1}$ is $O\left(N^3\right)$. The computational complexity of $\mathbf{P}_1 = \mathbf{P}^{-1}\mathbf{X}^{\mathrm{T}}$ is $O\left(dN^2\right)$, the computational complexity of $\mathbf{P}_2 = \mathbf{P}_1\mathbf{W}$ is $O\left(Nd^2\right)$, and the computational complexity of $\mathbf{P}_3 = \mathbf{P}_2 y$ is $O\left(Nd\right)$. So, in summary, the computational complexity of our proposed method is $O\left(TdN^2 + TNd^2 + TN^3 + TdN^2 + TNd^2 + TdNR\right)$.

5 Experiments

In order to verity the effectiveness of the proposed method, we conduct experiments on several common face databases, including ORL, FERET, Extended-YaleB and AR face databases. Meanwhile, several excellent algorithms are used in the experiments for comparison with our method. These algorithms include CRC, INNC [52], CFKNNC [53], NFRBC [54], KRBM [55], SRC, LRC, PCRC and MI-SRC [56].

5.1 Experiment on the ORL database

We adopt the Olivetti Research Laboratory (ORL) face database [57] in this experiment. This database contains a series of face images taken between April 1992 and April 1994 at the laboratory. There are 400 grayscale images taken from 40 objects, and these objects come from different ages, genders and races. Each object provides ten different images. The size of each image is 92 × 112 pixels, with 256 gray levels per pixel. The background of image is black. Facial expressions and details have made some changes, such as smiling or not smiling, open or closed eyes, glasses or no glasses. Facial pose is also varied, and its depth of rotation and revolution of plane can reach 20 degrees. Illumination condition has different changes. This database is currently one of the most widely used standard databases, which contains a larger number of comparative results. The first one, two, three, four and five face images of each object are used for training samples, and the rest images are treated as test samples in our experiment. Each image is resized to 32 × 32 pixels. Figure 5 shows some face images in this database. Experimental results are shown in Table 1.

From Table 1, with the increase of the number of training samples, the classification accuracy of our method and other methods also increases. It also shows the importance of adequate training samples for image classification. On the whole, our method is superior to other methods. When the number of the training samples per class is six, the recognition rates of most methods are over 90%. The accuracy of our method has reached 96.88% and is higher 2.51% than that

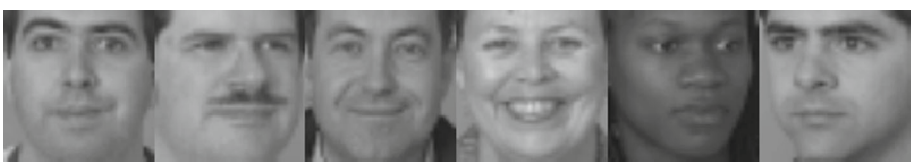


Fig. 6 Some face images in the FERET database

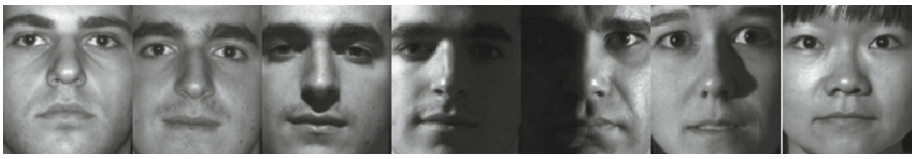


Fig. 7 Some face images in the Extended-YaleB database

of LRC. Moreover, the remarkable features of this database are that facial expressions and poses change variously. So, to some extent, our method is insensitive to the variations of facial expressions and poses.

5.2 Experiment on the FERET database

In order to promote the research and application of face recognition algorithm, Counterdrug Technology Transfer Program (CTTP) launched a Face Recognition Technology (FERET) engineering, which includes a generic face database and general test standard [58]. Each object provides a number of images, including different facial expressions, lighting, poses, gender and ages. Most of them are westerners, and there are 14,051 grayscale images. This database is also one of the most widely used face database and helpful for the research of face recognition. In this experiment, we use 1400 images taken from 200 objects, and each object provides seven different face images. Each image is resized to 40×40 pixels. Then we use the first one, two, three, four and five face images of each object as training samples and take the remaining images as test samples. Several face images in the FERET face database are shown in Fig. 6. Table 2 shows the classification accuracy obtained using different methods.

From Table 2, it is obvious that the proposed method can achieve lower the recognition error rates compared to other methods. For example, we used the first two face images of each subject as training samples and the rest face images as test samples. The classification accuracy rate of our method can outperform the CRC, KRBM, INNC, CFKNNC, NFRBC, SRC and PCRC algorithms by a margin of 9.20, 19.00, 8.50, 3.40, 3.20, 5.20 and 7.80%, respectively.

5.3 Experiment on the Extended-YaleB database

The database used in this experiment includes face images from YaleB and Extended-YaleB face databases under different illumination conditions [59]. This database consists of 10 objects from YaleB database and 28 objects from Extended-YaleB database, and each object has 64 face images captured under different lighting condition and poses. The first five, ten, fifteen, twenty, twenty-five and thirty face images of each object are treated as training samples, and the rest face images are used as test samples, respectively. Each image is resized to 32×32 pixels. Figure 7 shows some face images in the Extended-YaleB database. The comparative recognition rates obtained using different methods are shown in Table 3.

Table 3 The comparative recognition rates of the Extended-YaleB database with the number of training samples per class increases

Training samples	5 (%)	10 (%)	15 (%)	20 (%)	25 (%)
Proposed	52.81	75.05	77.39	81.34	88.26
CRC	45.36	63.50	70.03	73.44	75.71
KRBM	45.54	54.87	55.42	55.74	57.76
INNC	45.14	59.02	62.62	65.37	62.55
CFKNNC	43.04	53.17	59.24	63.34	65.59
NFRBC	65.48	65.79	65.74	69.14	71.66
SRC(L1LS)	51.43	68.42	70.52	72.24	78.54
MI-SRC	42.42	71.20	72.07	77.57	85.12
LRC	45.95	70.89	74.24	73.58	79.10
PCRC	53.61	74.32	76.37	79.67	84.62

From Table 3, we can see that the classification accuracy of our method increases as the number of training samples per class increases. When the number of training samples per class is 25, our method achieves a recognition rate of 88.26%, which is higher 3.14% than MI-SRC. In addition, although our method is lower 12.67% than NFRBC when the number of training samples per class is five, with the increase of training samples, the classification accuracy of our method quickly surpassed that of NFRBC, and the rising range of our method is significantly higher than that of NFRBC. The notable feature of this database is the obvious changes in illumination, so these experimental results illustrate that our method is insensitive to variations of illuminations to some extent.

5.4 Experiment on the AR database

The AR face database is established by the Barcelona computer vision center in Spain, which contains 3288 face images taken from 116 objects [60]. In the acquisition environment, the parameters of camera, illumination conditions and camera distance are strictly controlled. Moreover, image feature frontal view faces with different facial expressions, illumination conditions and occlusions (sunglasses and scarf). Images in the database are divided into two time stages; each stage has thirteen pictures. Facial expressions and illumination have seven variations, and facial occlusion uses three sunglasses and three scarves. We use a subset of the AR

Fig. 8 Some face images in the AR database

Table 4 The comparative recognition rates of the AR database with the number of training samples per class increases

Training samples	2 (%)	4 (%)	5 (%)	6 (%)
Proposed	71.74	70.15	71.31	71.04
CRC	68.33	67.54	70.99	70.83
KRBM	64.97	65.00	66.43	68.50
INNC	69.62	68.45	69.17	69.08
CFKNNC	65.76	63.52	63.33	64.42
NFRBC	68.13	68.43	70.99	71.04
SRC(L1LS)	55.14	63.99	70.07	64.93
MI-SRC	65.66	53.30	58.25	57.92
LRC	59.00	59.96	59.84	68.37
PCRC	68.30	68.75	70.63	70.58

database. This subset contains 3120 face images taken from 120 objects, with each object providing 26 face images. Each image is resized to 50×40 pixels. The first two, four, five, six, eight and ten face images are regarded as training samples, and the rest images are used as test samples. Figure 8 shows several face images in the AR database. Experimental results are shown in Table 4. From these results, our method outperforms the other competing algorithms.

6 Conclusion

A new effective sparse representation-based classification method is proposed for face recognition in this paper. In traditional sparse representation algorithms, the residuals between the test sample and the linear representation obtained using the training samples of each class are treated equally. But based on prior knowledge, each residual is different, which is the specific embodiment of the existence of differences between classes. These differences can make the classification task easier. Therefore, each residual mentioned above should be treated differently, which can enhance the distinctiveness of different classes. So we introduce a weighted matrix in sparse representation method. It can make small deviations correspond to higher weights, and large deviations correspond to lower weights. The constraint term of representation coefficients is improved, and the deviation between the linear representation of all training samples and of each class is taken into account. Then we exploit the obtained optimal representation coefficients to perform classification. According to experimental results on the ORL, FERET, Extended-YaleB and AR databases, our method has good adaptive capability for variety of external factors, such as illumination change, facial expression variations, poses variety and occlusion interference.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Nos. 61672333, 61402274, 61703096, 41471280), China Postdoctoral Science Foundation (No. 2017M611655), the Program of Key Science and Technology Innovation Team in Shaanxi Province (No. 2014KTC-18), the Key Science and Technology Program of Shaanxi Province (No. 2016GY-081), the National Natural Science Foundation of Jiangsu Province (No. BK20170691), the Fundamental Research Funds for the Central Universities (Nos. GK201803059, GK201803088), Interdisciplinary Incubation Project of Learning Science of Shaanxi Normal University.

Appendix 1: The derivative over β of $\frac{1}{2}(y - X\beta)^{T}W(y - X\beta) + \gamma \sum_{i=1}^{c} \sum_{j=1}^{c} \beta_i^{T} X_i^{T} X_j \beta_j + \lambda \sum_{i=1}^{c} \|X\beta - X_i\beta_i\|_2^2$

First, $\frac{d}{d\beta}\left(\frac{1}{2}(y - \mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta)\right) = -\mathbf{X}^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta)$.

Next, letting $f(\beta) = \gamma \sum_{i=1}^{c} \sum_{j=1}^{c} \beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j$, we can calculate the partial derivatives $\frac{\partial f}{\partial \beta_k}$. Then $\frac{df}{d\beta}$ can be obtained by using all $\frac{\partial f}{\partial \beta_k}$ $k = 1, \ldots, c$. Based on mathematical experience,

$$\begin{aligned}\beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j &= (\mathbf{X}_i \beta_i)^{\mathrm{T}} \mathbf{X}_j \beta_j \\ &= \frac{1}{2}\left(\|\mathbf{X}_i\beta_i + \mathbf{X}_j\beta_j\|_2^2 - \|\mathbf{X}_i\beta_i\|_2^2 - \|\mathbf{X}_j\beta_j\|_2^2\right).\end{aligned}$$

So $f(\beta)$ can be rewritten as

$$\begin{aligned}f(\beta) &= \gamma \sum_{i=1}^{c} \sum_{j=1}^{c} \beta_i^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j \beta_j \\ &= \frac{\gamma}{2}\Bigg[\sum_{\substack{i=1,\ldots,c \\ i\neq k}} \left(\|\mathbf{X}_i\beta_i + \mathbf{X}_k\beta_k\|_2^2 - \|\mathbf{X}_i\beta_i\|_2^2 - \|\mathbf{X}_k\beta_k\|_2^2\right) \\ &\quad + \sum_{\substack{j=1,\ldots,c \\ j\neq k}} \left(\|\mathbf{X}_k\beta_k + \mathbf{X}_j\beta_j\|_2^2 - \|\mathbf{X}_k\beta_k\|_2^2 - \|\mathbf{X}_j\beta_j\|_2^2\right) \\ &\quad + \sum_{\substack{i=1,\ldots,c \\ i\neq k}} \sum_{\substack{j=1,\ldots,c \\ j\neq k}} \left(\|\mathbf{X}_i\beta_i + \mathbf{X}_j\beta_j\|_2^2 - \|\mathbf{X}_i\beta_i\|_2^2 - \|\mathbf{X}_j\beta_j\|_2^2\right)\Bigg]\end{aligned}$$

$$= \gamma \sum_{\substack{i=1,\ldots,c \\ i\neq k}} \left(\|\mathbf{X}_i\beta_i + \mathbf{X}_k\beta_k\|_2^2 - \|\mathbf{X}_i\beta_i\|_2^2 - \|\mathbf{X}_k\beta_k\|_2^2 \right) + \frac{\gamma}{2} \sum_{\substack{i=1,\ldots,c \\ i\neq k}} \sum_{\substack{j=1,\ldots,c \\ j\neq k}} \left(\|\mathbf{X}_i\beta_i + \mathbf{X}_j\beta_j\|_2^2 - \|\mathbf{X}_i\beta_i\|_2^2 - \|\mathbf{X}_j\beta_j\|_2^2 \right).$$

The calculation procedure of $\frac{\partial f}{\partial \beta_k}$ is as follows,

$$\begin{aligned}
\frac{\partial f}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k}\left(\gamma \sum_{i=1}^{c}\sum_{j=1}^{c} \beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j \right) \\
&= \frac{\partial}{\partial \beta_k}\left(\gamma \sum_{\substack{i=1,\ldots,c \\ i\neq k}} \left(\|\mathbf{X}_i\beta_i + \mathbf{X}_k\beta_k\|_2^2 - \|\mathbf{X}_i\beta_i\|_2^2 - \|\mathbf{X}_k\beta_k\|_2^2 \right) \right) \\
&= \gamma \sum_{\substack{i=1,\ldots,c \\ i\neq k}} \left(2\mathbf{X}_k^{\mathrm{T}}\left(\mathbf{X}_i\beta_i + \mathbf{X}_k\beta_k\right) - 2\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k\beta_k \right) \\
&= \gamma \sum_{\substack{i=1,\ldots,c \\ i\neq k}} \left(2\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_i\beta_i \right) = 2\gamma \left[\left(\sum_{i=1,\ldots,c} \mathbf{X}_k^{\mathrm{T}}\mathbf{X}_i\beta_i \right) - \mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k\beta_k \right] \\
&= 2\gamma\mathbf{X}_k^{\mathrm{T}}\mathbf{X}\beta - 2\gamma\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k\beta_k.
\end{aligned}$$

Thus, the derivative over β of $f(\beta)$ is $\frac{df}{d\beta} = \begin{bmatrix} \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_c} \end{bmatrix} = \begin{bmatrix} 2\gamma\mathbf{X}_1^{\mathrm{T}}\mathbf{X}\beta - 2\gamma\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\beta_1 \\ \vdots \\ 2\gamma\mathbf{X}_k^{\mathrm{T}}\mathbf{X}\beta - 2\gamma\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k\beta_k \end{bmatrix} = 2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta - 2\gamma\mathbf{M}\beta$,

where $\mathbf{M} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c \end{pmatrix}$.

As for $\frac{\partial}{\partial \beta}\left(\lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2\right)$, we need to analyze $\sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2$ and deduce the deformation formula of $\sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2$ for convenience of calculation. Due to $\mathbf{X}\beta = [\mathbf{X}_1, \ldots, \mathbf{X}_c]\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_c \end{bmatrix} = \mathbf{X}_1\beta_1 + \cdots + \mathbf{X}_c\beta_c$, we have $\mathbf{X}\beta - \mathbf{X}_i\beta_i = \mathbf{X}_1\beta_1 + \cdots + \mathbf{X}_{i-1}\beta_{i-1} + \mathbf{X}_{i+1}\beta_{i+1} + \cdots + \mathbf{X}_c\beta_c$. Letting $\mathbf{S}_i = [0, \ldots, \mathbf{X}_i, \ldots, 0]$ and $\mathbf{Z}_i = \mathbf{X} - \mathbf{S}_i = \left[\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, 0, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_c\right]$, we can obtain the deformation formula of $\mathbf{X}\beta - \mathbf{X}_i\beta_i$, i.e., $\mathbf{X}\beta - \mathbf{X}_i\beta_i = \mathbf{Z}_i\beta = \mathbf{X}_1\beta_1 + \cdots + \mathbf{X}_{i-1}\beta_{i-1} + \mathbf{X}_{i+1}\beta_{i+1} + \cdots + \mathbf{X}_c\beta_c$. Therefore, the derivative over β of $\lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2$ is

$$\begin{aligned}
&\frac{\partial}{\partial \beta}\left(\lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2 \right) \\
&= \frac{\partial}{\partial \beta}\left(\lambda \sum_{i=1}^{c} \|\mathbf{Z}_i\beta_i\|_2^2 \right) = 2\lambda \left[\sum_{i=1}^{c} (\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i \right] \beta.
\end{aligned}$$

Eventually, the derivative over β of $\frac{1}{2}(y - \mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta) + \gamma \sum_{i=1}^{c}\sum_{j=1}^{c} \beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j + \lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2$ is

$$\begin{aligned}
&\frac{\partial}{\partial \beta}\left(\frac{1}{2}(y - \mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta) + \gamma \sum_{i=1}^{c}\sum_{j=1}^{c} \beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j + \lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2 \right) \\
&= -\mathbf{X}^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta) + 2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta - 2\gamma\mathbf{M}\beta + 2\lambda \left[\sum_{i=1}^{c} (\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i \right] \beta.
\end{aligned}$$

Appendix 2: Proof of our objective function is convex function

In the literature [49], there is a description that one function is a convex function as long as it satisfies some certain conditions. Specifically, suppose f is a twice differentiable function, namely, its second derivative or Hessian $\nabla^2 f$ is continuous and exists at each point in $\mathbf{dom}\, f$, where $\mathbf{dom}\, f$ is open. Then, f is a convex function if and only if $\mathbf{dom}\, f$ is convex, and also the Hessian of f is positive semidefinite, i.e., $\nabla^2 f(x) \succeq 0$, all $x \in \mathbf{dom}\, f$. In addition, there is an example which can help us to better explain and prove the convex characteristic of the objective function, as follows.

Example 1 Consider the quadratic function $f : \mathbf{R}^n \to \mathbf{R}$, with $\mathbf{dom}\, f = \mathbf{R}^n$, given by $f(x) = (1/2)x^{\mathrm{T}}\mathbf{P}x + q^{\mathrm{T}}x + r$, where $\mathbf{P}$ is a symmetric matrix of size $n \times n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$. Due to $\nabla^2 f(x) = \mathbf{P}$ for all x, f is convex if and only if $\mathbf{P} \succeq 0$.

Let $g(\beta) = \frac{1}{2}(y - \mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta) + \gamma \sum_{i=1}^{c}\sum_{j=1}^{c} \beta_i^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j\beta_j + \lambda \sum_{i=1}^{c} \|\mathbf{X}\beta - \mathbf{X}_i\beta_i\|_2^2$.

Then according to the aforementioned theorem and example, we can infer that the function $g(\beta)$ is convex function if $\nabla^2 g(\beta) \succeq 0$ is proved to be valid, that is, $\nabla^2 g(\beta)$ is a positive semidefinite matrix. As for the problem of how to determine a matrix is positive semidefinite matrix, as long as this matrix is a real symmetric matrix and all order principal minor determinant are greater than or equal to zero, we can conclude that it is positive semidefinite matrix. From Eq. (7), we can get $\nabla^1 g(\beta) = -\mathbf{X}^{\mathrm{T}}\mathbf{W}(y - \mathbf{X}\beta) + 2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta - 2\gamma\mathbf{M}\beta +$

$2\lambda\left[\sum_{i=1}^{c}(\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i\right]\beta$, and then $\nabla^2 g(\beta) = -\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + 2\gamma\mathbf{X}^{\mathrm{T}}\mathbf{X} - 2\gamma\mathbf{M} + 2\lambda\sum_{i=1}^{c}(\mathbf{Z}_i)^{\mathrm{T}}\mathbf{Z}_i$. Because $\nabla^2 g(\beta)$ satisfies the above determination conditions of positive semidefinite matrix, it is concluded that our objective function is convex function.

References

1. Liu, W., Zha, Z.J., Wang, Y., Lu, K., Tao, D.: p-Laplacian regularized sparse coding for human activity recognition. IEEE Trans. Ind. Electron. **63**(8), 5120–5129 (2016)
2. Xu, Y., Fei, L., Wen, J., Zhang, D.: Discriminative and robust competitive code for palmprint recognition. IEEE Trans. Syst. Man Cybern. Syst. **PP**(99), 1–10 (2016)
3. Chen, G., Tao, D., Wei, L., Liu, L., Jie, Y.: Label propagation via teaching-to-learn and learning-to-teach. IEEE Trans. Neural Netw. Learn. Syst. **28**(6), 1452–1465 (2017)
4. Yong, X., Li, X., Yang, J., Lai, Z., Zhang, D.: Integrating conventional and inverse representation for face recognition. IEEE Trans. Cybern. **44**(10), 1738–1746 (2014)
5. Yong, X., Fang, X., Li, X., Yang, J., You, J., Liu, H., Teng, S.: Data uncertainty in face recognition. IEEE Trans. Cybern. **44**(10), 1950–1961 (2014)
6. Chen, X., Ziarko, W.: Experiments with rough set approach to face recognition. Int. J. Intell. Syst. **26**(6), 499–517 (2011)
7. Fang, Y., Lin, W., Fang, Z., Chen, Z., Lin, C.W., Deng, C.: Visual acuity inspired saliency detection by using sparse features. Inf. Sci. Int. J. **309**(C), 1–10 (2015)
8. Du, B., Wang, Z., Zhang, L., Zhang, L., Liu, W., Shen, J., Tao, D.: Exploring representativeness and informativeness for active learning. IEEE Trans. Cybern. **PP**(99), 1–13 (2015)
9. Liu, W., Ma, T., Xie, Q., Tao, D., Cheng, J.: LMAE: a large margin auto-encoders for classification. Sig. Process. **141**, 137–143 (2017)
10. Liu, W., Tao, D., Cheng, J., Tang, Y.: Multiview Hessian discriminative sparse coding for image annotation. Comput. Vis. Image Underst. **118**(1), 50–60 (2014)
11. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W.: Saliency detection for stereoscopic images. IEEE Trans. Image Process. **23**(6), 2625–2636 (2014)
12. Du, B., Xiong, W., Wu, J., Zhang, L., Zhang, L., Tao, D.: Stacked convolutional denoising auto-encoders for feature representation. IEEE Trans. Cybern. **47**(4), 1017–1027 (2016)
13. Gong, C., Liu, T., Tao, D., Fu, K., Tu, E., Yang, J.: Deformed graph laplacian for semisupervised learning. IEEE Trans. Neural Netw. Learn. Syst. **26**(10), 2261–2274 (2015)
14. Liu, T., Tao, D.: On the performance of manhattan nonnegative matrix factorization. IEEE Trans. Neural Netw. Learn. Syst. **27**(9), 1851–1863 (2016)
15. Du, B., Wang, N., Wang, N., Zhang, L., Zhang, L., Zhang, L.: Hyperspectral signal unmixing based on constrained non-negative matrix factorization approach. Neurocomputing **204**(C), 153–161 (2016)
16. Liu, W., Yang, X., Tao, D., Cheng, J., Tang, Y.: Multiview dimension reduction via Hessian multiset canonical correlations. Inf. Fusion **41**, 119–128 (2017)
17. Liu, T., Gong, M., Tao, D.: Large-cone nonnegative matrix factorization. IEEE Trans. Neural Netw. Learn. Syst. **28**(9), 2129–2142 (2017)
18. Yu, J., Hong, C., Rui, Y., Tao, D.: Multi-task autoencoder model for recovering human poses. IEEE Trans. Ind. Electron. **PP**(99), 1 (2017)
19. Gong, C., Tao, D., Maybank, S.J., Liu, W., Kang, G., Yang, J.: Multi-modal curriculum learning for semi-supervised image classification. IEEE Trans. Image Process. **25**(7), 3249–3260 (2016)
20. Bo, D., Zhang, M., Zhang, L., Ruimin, H., Tao, D.: PLTD: patch-based low-rank tensor decomposition for hyperspectral images. IEEE Trans. Multimed. **19**(1), 67–79 (2017)
21. Liu, W., Zhang, L., Tao, D., Cheng, J.: Support vector machine active learning by Hessian regularization. J. Vis. Commun. Image Represent. **49**, 47–56 (2017)
22. Yang, X., Liu, W., Tao, D., Cheng, J.: Canonical correlation analysis networks for two-view image recognition. Inf. Sci. Int. J. **385**(C), 338–352 (2017)
23. Fang, Y., Wang, Z., Lin, W.: Video saliency incorporating spatiotemporal cues and uncertainty weighting. In: IEEE International Conference on Multimedia and Expo, pP. 1–6 (2013)
24. Bo, D., Zhao, R., Zhang, L., Zhang, L.: A spectral-spatial based local summation anomaly detection method for hyperspectral images. Signal Process. **124**(C), 115–131 (2016)
25. Tao, D., Li, X., Wu, X., Maybank, S.J.: Geometric mean for subspace selection. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 260–274 (2009)
26. Li, L., Liu, S., Peng, Y., Sun, Z.: Overview of principal component analysis algorithm. Optik Int. J. Light Electron Opt. **127**(9), 3935–3944 (2016)
27. Gong, C., Tao, D., Fu, K., Yang, J.: Fick's law assisted propagation for semisupervised learning. IEEE Trans. Neural Netw. Learn. Syst. **26**(9), 2148–2162 (2015)
28. Chen, G., Liu, T., Tang, Y., Jian, Y., Jie, Y., Tao, D.: A regularization approach for instance-based superset label learning. IEEE Trans. Cybern. **PP**(99), 1–12 (2017)
29. Yu, J., Yang, X., Fei, G., Tao, D.: Deep multimodal distance metric learning using click constraints for image ranking. IEEE Trans. Cybern. **PP**(99), 1–11 (2016)
30. Fang, Y., Fang, Z., Yuan, F., Yang, Y., Yang, S., Xiong, N.N.: Optimized multioperator image retargeting based on perceptual similarity measure. IEEE Trans. Syst. Man Cybern. Syst. **47**(11), 2956–2966 (2017)
31. Gong, C., Tao, D., Chang, X., Yang, J.: Ensemble teaching for hybrid label propagation. IEEE Trans. Cybern. **PP**(99), 1–15 (2017)
32. Yong, X., Zhong, A., Yang, J., Zhang, D.: LPP solution schemes for use with face recognition. Pattern Recognit. **43**(12), 4165–4176 (2010)
33. Yu, J., Rui, Y., Tang, Y.Y., Tao, D.: High-order distance-based multiview stochastic learning in image classification. IEEE Trans. Cybern. **44**(12), 2431 (2014)
34. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
35. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. MIT Press, Cambridge (2003)
36. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
37. Wright, J., Ganesh, A., Zhou, Z., Wagner, A., Ma, Y.: Demo: robust face recognition via sparse representation. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–2 (2009)
38. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11), 2106–2112 (2010)
39. Yong, X., Zhang, D., Yang, J., Yang, J.Y.: A two-phase test sample sparse representation method for use with face recognition. IEEE Trans. Circuits Syst. Video Technol. **21**(9), 1255–1262 (2011)
40. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: which helps face recognition? In: IEEE International Conference on Computer Vision, pp. 471–478 (2012)

41. Deng, W., Jiani, H., Guo, J.: Extended SRC: undersampled face recognition via intraclass variant dictionary. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1864–1870 (2012)
42. Tang, X., Feng, G., Cai, J.: Weighted group sparse representation for undersampled face recognition. Neurocomputing **145**(18), 402–415 (2014)
43. Timofte, R., Van Gool, L.: Adaptive and weighted collaborative representations for image classification. Pattern Recognit. Lett. **43**(1), 127–135 (2014)
44. Wu, J., Timofte, R., Van Gool, L.: Learned collaborative representations for image classification. In: IEEE Winter Conference on Applications of Computer Vision, pp. 456–463 (2015)
45. Yong, X., Zhong, Z., Jian, Y., You, J., Zhang, D.: A new discriminative sparse representation method for robust face recognition via regularization. IEEE Trans. Neural Netw. Learn. Syst. **PP**(99), 1–10 (2016)
46. Cai, S., Zhang, L., Zuo, W., Feng, X.: A probabilistic collaborative representation based approach for pattern classification. In: Computer Vision and Pattern Recognition, pp. 2950–2959 (2016)
47. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theor. Comput. Sci. **209**(1–2), 237–260 (1998)
48. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 447 (2016)
49. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2010)
50. Fang, Y., Lin, W., Chen, Z., Tsai, C.M., Lin, C.W.: A video saliency detection model in compressed domain. IEEE Trans. Circuits Syst. Video Technol. **24**(1), 27–38 (2014)
51. Huang, W., Wang, X., Jin, Z., Li, J.: Penalized collaborative representation based classification for face recognition. Appl. Intell. **4**(4), 12–19 (2015)
52. Xu, Y., Zhu, Q., Chen, Y., Pan, J.S.: An improvement to the nearest neighbor classifier and face recognition experiments. Int. J. Innov. Comput. Inf. Control **9**(2), 543–554 (2013)
53. Yong, X., Zhu, Q., Fan, Z., Qiu, M., Chen, Y., Liu, H.: Coarse to fine K nearest neighbor classifier. Pattern Recognit. Lett. **34**(9), 980–986 (2013)
54. Yong, X., Fang, X., You, J., Chen, Y., Liu, H.: Noise-free representation based classification and face recognition experiments. Neurocomputing **147**(1), 307–314 (2015)
55. Yong, X., Fan, Z., Zhu, Q.: Feature space-based human face image representation and recognition. Opt. Eng. **51**(1), 7205 (2012)
56. Yong, X., Li, X., Yang, J., Zhang, D.: Integrate the original face image and its mirror image for face recognition. Neurocomputing **131**(7), 191–199 (2014)
57. ORL: Face database. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. Accessed 1 Mar 2017
58. FERET: Face database. http://www.itl.nist.gov/iad/humanid/feret/feret_master.html. Accessed 1 Mar 2017
59. YaleB: Face database. http://vision.ucsd.edu/content/yale-face-database. Accessed 1 Mar 2017
60. AR: Face database. http://web.mit.edu/emeyers/www/face_databases.html#ar. Accessed 1 Mar 2017

Yali Peng was born in Shaanxi, China, in 1979. She got her M.S. degree from Harbin Engineering University, Harbin, China, in 2005. From 2005 to 2007, she worked as an engineer at the company of DaTang telephone. In 2013, she received her Ph.D. degree from Xidian University of China. Since 2016, she has been an associate professor in the School of Computer Science, the Shaanxi Normal University. Her research interests include face recognition, pattern recognition, image processing, 3D reconstructions, applications of computer vision and computer graphics.

Lingjun Li was born in Henan, China, in 1990. She received the B.S. degrees from the Luoyang Normal University, Luoyang, China, in 2014. She is currently working for her M.S. degree in the School of Computer Science, the Shaanxi Normal University. Her research interests include face recognition, pattern recognition and image processing.

Shigang Liu was born in Jiangxi, China, in 1973. He received the B.S. and M.S. degrees from Harbin Engineering University, Harbin, China, in 1997 and 2001, respectively. In 2005, he received his Ph.D. degrees from Xidian University of China. From 2007 to 2009, he was a post-doc in the Xi'an Jiaotong University. He is currently a professor in the School of Computer Science, the Shaanxi Normal University. His research interests include face recognition, pattern recognition, image processing, 3D reconstructions and computer graphics.

Jun Li received the B.S. degree in Electrical Engineering & Automation from Nanjing Normal University, Nanjing, China, in 2008, the M.S. degree in Control Theory & Engineering and Ph.D. degree in Pattern Recognition & Intelligent Systems from Southeast University, Nanjing, in 2011 and 2016, respectively. He is currently a Post-Doctoral Fellow with the School of Automation, Southeast University. His research interests include multimedia search and computer vision.

Xili Wang was born in Shaanxi Province, China, in 1969. She received the B.S. degree in computer application from Tianjin University, China, in 1991, and received her M.S. degree and Ph.D. degree from Xidian University, China, in 1994 and 2004, respectively. She is currently a professor in the School of Computer Science, Shaanxi Normal University, Xi'an, China. Her research interests include intelligent information processing, image perception and understanding.