Robust face recognition via low-rank sparse representation-based classification

Haishun Du[1]   Qingpu Hu[1]   Dianfeng Qiao[1]   Ioannis Pitas[2]

[1]Institute of *Image Processing and Pattern Recognition*, Henan University, Kaifeng 475004, China

[2]Department of *Informatics*, Aristotle University of Thessaloniki, GR-54006 Thessaloniki, Greece

Abstract: Face recognition has attracted great interest due to its importance in many real-world applications. In this paper, we present a novel low-rank sparse representation-based classification (LRSRC) method for robust face recognition. Given a set of test samples, LRSRC seeks the lowest-rank and sparsest representation matrix over all training samples. Since low-rank model can reveal the subspace structures of data while sparsity helps to recognize the data class, the obtained test sample representations are both representative and discriminative. Using the representation vector of a test sample, LRSRC classifies the test sample into the class which generates minimal reconstruction error. Experimental results on several face image databases show the effectiveness and robustness of LRSRC in face image recognition.

Keywords: Face recognition, image classification, sparse representation, low-rank representation,low-rank sparse representation.

1   Introduction

Face recognition (FR) has been a hot topic in computer vision and pattern recognition due to its importance in many real-world applications, such as public security monitoring, access control and card identification. Although face images usually have high dimensionality, they possibly reside on a low-dimensional subspace[1,2]. Many subspace learning methods, such as principal component analysis (PCA)[3], linear discriminant analysis (LDA)[4], locality preserving projection (LPP)[5], marginal Fisher analysis (MFA)[6], local discriminant embedding (LDE)[7] and sparsity preserving discriminant embedding (SPDE)[8] have been proposed for reducing face image dimensionality. Subsequently, a classifier, such as nearest neighbor classifier (NNC)[9] or support vector machine (SVM), is usually employed for classification. Although subspace learning methods have achieved a good accuracy in FR, they are not robust to face image corruption[10], such as occlusion, disguise or pixel contamination. Especially, when the amount of training samples is small, the learned subspace projections may be wrong or deficient[11].

Recently, a new robust FR framework, using sparse representation-based classification (SRC)[12], was presented. SRC sparsely encodes a test sample over a dictionary consisting of training samples via -norm optimization techniques[13−15]. Then, the test sample is classified into the class that generates minimal reconstruction residual. In the case of test samples suffering from corruption and occlusion, SRC uses an identity matrix as occlusion dictionary to deal with occlusion or corruption of test samples. However, the dimensionality of the identity matrix is usually very high, which renders sparse coding computationally expensive. To solve this issue, Deng et al.[16] proposed an Extended SRC (ESRC) method, in which an

intra-class variation matrix, determined by subtracting the class centroid from the same class samples, is used as occlusion dictionary. The dimensionality of occlusion dictionary used in ESRC is much smaller than that of occlusion dictionary used in SRC. This fact greatly improves computational efficiency[16].

Although SRC and its extensions have been successfully used in FR, their working mechanism has not been clearly revealed. Recently, Zhang et al.[17] proposed an efficient FR method, namely collaborative representation-based classification (CRC). They thought that it is the collaborative representation using all training samples, but not the L_1-norm sparsity constraint, plays a key role in the success of SRC[17]. Therefore, they replaced the L_1-norm by the computationally much more efficient L_2-norm and claimed that CRC can achieve competitive FR performance compared with SRC. I. Naseem et al.[18] presented a new L_2-norm based classification method, namely the linear regression-based classification (LRC), by using the concept that samples from the one subject reside on a linear subspace. LRC linearly encodes a test sample over class-specific training samples by using the least squares method and classifies the test sample based on class-specific reconstruction residuals.

Recently, Liu et al.[19] presented a Low-rank representation (LRR) method to reveal the global structure of data that are drawn from multiple subspaces. Lu et al.[20] incorporated a graph regularization term into the LRR to encode the local structure information of data. L. Ma et al.[21] proposed a discriminative low-rank dictionary learning algorithm for sparse representation. Y. Zhang et al.[22] proposed a structured low-rank representation method for image classification, in which they used the discriminative low-rank representations of training data to learn a dictionary and design a linear multi-classifier.

Inspired by above methods, we present a low-rank sparse representation-based classification (LRSRC) method for robust FR. Given a set of test samples, LRSRC seeks the lowest-rank and sparsest representation matrix over all the

training samples. As observed in [19], low-rank model can reveal the subspace structures of data, while sparsity helps to recognize the data class. Since LRSRC imposes both the lowest-rank and sparsest constraints on the representation matrix, the obtained test sample representations are both representative and discriminative. Using the representation vector of a test sample, LRSRC assigns the test sample to the class that generates minimal reconstruction error. It is worth mentioning that LRSRC can efficiently deal with corrupted test samples due to its accurate corruption estimation. It is worth noting that there are some mainly differences between our work and the works of [21] and [22], though our work is related to them. The mainly differences are as follows. a) Our work focuses on the low-rankness and sparsity of the representation matrix of test samples, while the work of [21], as a dictionary learning method, mainly emphasizes the low-rankness of the learned dictionary; b) Our work directly uses the training data matrix as dictionary to calculate the lowest-rank and sparest representation matrix of test samples, and classifies test samples based on the minimal reconstruction errors, while [22] mainly uses the discriminative low-rank representations of training data to learn a dictionary and design a linear multi-classifier for image classification.

The rest of the paper is outlined as follows. In Section 2, we briefly review some related works, including sparse representation-based classification (SRC) and low-rank representation (LRR). Section 3 presents the propose LRSRC method for robust FR in detail. Experimental results and corresponding discussions are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 Related work

2.1 Sparse representation based classification

The sparse representation-based classification (SRC) was proposed in [12] for robust face recognition. Suppose there are C classes of training samples, let $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{iN_i}] \in \mathbb{R}^{D \times N_i}$ be the matrix formed by N_i training samples of ith class, in which $\boldsymbol{x}_{ij}$ is an one-dimensional vector stretched by the jth sample of the ith class. Then $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C] \in \mathbb{R}^{D \times N}$ is the matrix of all training samples, where $N = \sum_{i=1}^{C} N_i$ is the total number of training samples. Using $\boldsymbol{X}$ as dictionary, a test sample $\boldsymbol{y}$ can be approximately represented as linear combination of the elements of $\boldsymbol{X}$, i.e. $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a}$, in which $\boldsymbol{a}$ is the representation coefficient vector of $\boldsymbol{y}$. If $D < N$, the linear equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a}$ is underdetermined. Therefore, the solution $\boldsymbol{a}$ is not unique. Since the test sample $\boldsymbol{y}$ can be adequately represented by the training samples from the correct class, the representation is obviously sparse when the amount of training samples is large enough. SRC claims that the sparser the representation coefficient vector $\boldsymbol{a}$ is, the easier it is to recognize the test samples class label[12]. This inspires one to find the sparsest solution of $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a}$ by solving the following optimization problem.

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2 + \lambda\|\boldsymbol{a}\|_1, \qquad (1)$$

where λ is a scalar constant. The optimization problem (1) can be efficiently solved by many algorithms, such as basis pursuit[13], l1_ls[14] and alternating direction algorithm[15].

Having obtained the sparsest solution $\hat{\boldsymbol{a}}$, let $\delta_i(\hat{\boldsymbol{a}})$ be a vector, whose only nonzero entries are the entries in $\hat{\boldsymbol{a}}$ that are associated with class i[12]. Using $\delta_i(\hat{\boldsymbol{a}})$, the test sample $\boldsymbol{y}$ can be reconstructed as $\hat{\boldsymbol{y}}_i = \boldsymbol{X}\delta_i(\hat{\boldsymbol{a}})$. Then, the test sample $\boldsymbol{y}$ is classified into the class that minimizes class reconstruction error between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_i$:

$$\text{identity}(\boldsymbol{y}) = \min_i \|\boldsymbol{y} - \boldsymbol{X}\delta_i(\hat{\boldsymbol{a}})\|_2. \qquad (2)$$

In real-world applications, the observations usually are corrupted or occluded. A test sample $\boldsymbol{y}$ is rewritten as

$$\boldsymbol{y} = \boldsymbol{y}_0 + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} = [\boldsymbol{X}, \boldsymbol{I}] \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{e} \end{bmatrix} = \boldsymbol{B}\boldsymbol{w}, \qquad (3)$$

Where $\boldsymbol{B} = [\boldsymbol{X}, \boldsymbol{I}] \in \mathbb{R}^{D \times (D+N)}$, and $\boldsymbol{I} \in \mathbb{R}^{D \times D}$ is the identity matrix. The clean test sample $\boldsymbol{y}_0$ and the corruption error $\boldsymbol{e}$ can be sparsely encoded over the dictionary $\boldsymbol{X}$ and the occlusion dictionary $\boldsymbol{I}$. The sparse solution $\hat{\boldsymbol{w}} = [\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}]^{\mathrm{T}}$ can be computed by solving an optimization problem similar to problem (1). Then the test sample $\boldsymbol{y}$ is classified by the following decision rule:

$$\text{identity}(\boldsymbol{y}) = \min_i \|\boldsymbol{y} - \boldsymbol{X}\delta_i(\hat{\boldsymbol{a}}) - \hat{\boldsymbol{e}}\|_2. \qquad (4)$$

2.2 Low-rank representation

Recently, Liu et al.[19] presented a low-rank representation (LRR) method to recover the subspace structures from data that reside on a union of multiple subspaces. LRR seeks the lowest-rank representation matrix $\boldsymbol{Z}$ of data matrix $\boldsymbol{X}$ over a given dictionary $\boldsymbol{A}$. The rank minimization problem is formulated as follows:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \text{rank}(\boldsymbol{Z}) + \lambda\|\boldsymbol{E}\|_l, \quad s.t. \ \boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{E}, \qquad (5)$$

where λ is scalar constant, and $\|\boldsymbol{E}\|_l$ denotes a certain regularization strategy for characterizing different corruptions, such as $\|\boldsymbol{E}\|_0$ for charactering random corruption, and the $\|\boldsymbol{E}\|_{2,0}$ for describing sample-specific corruption.

If one just considers the case in which the data matrix $\boldsymbol{X}$ entries are random corrupted, the rank optimization problem can be formulated as:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \text{rank}(\boldsymbol{Z}) + \lambda\|\boldsymbol{E}\|_0, \quad s.t. \ \boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{E}. \qquad (6)$$

As finding the solution of optimization problem (6) is NP-hard and computationally expensive, Liu et al.[19] replaced the rank with nuclear norm and the L_0-norm with L_1-norm. So, the optimization problem (6) is relaxed to the following optimization problem:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \|\boldsymbol{Z}\|_* + \lambda\|\boldsymbol{E}\|_1, \quad s.t. \ \boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{E}, \qquad (7)$$

where $\|\boldsymbol{Z}\|_*$ denotes the nuclear norm (i.e., the sum of the singular values) of $\boldsymbol{Z}$. Given an appropriate dictionary $\boldsymbol{A}$ (in [19], the data matrix $\boldsymbol{X}$ is used as dictionary), the augmented Lagrange multipliers (ALM) algorithm[23] can be used to solve the optimization problem (7). Having obtained the optimal solution $(\boldsymbol{Z}^*, \boldsymbol{E}^*)$, the representation matrix $\boldsymbol{Z}^*$ can be used for recovering the underlying subspace structures of data $\boldsymbol{X}$.

3 Low-rank sparse representation based classification

In this section, we present a low-rank sparse representation-based classification (LRSRC) method for robust face recognition. LRSRC firstly seeks a lowest-rank and sparsest representation matrix of the test samples. Then, by using the representation vector of a test sample, LRSRC assigns the test sample to the class that generates minimal reconstruction error.

3.1 low-rank sparse representation

Let the training data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C] \in \mathbb{R}^{D \times N}$ be composed by C classes of training samples, and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M] \in \mathbb{R}^{D \times M}$ be the data matrix of all test samples. Using $\boldsymbol{X}$ as dictionary, a test sample $\boldsymbol{y}_i$ can be approximately represented as linear combination of the elements of $\boldsymbol{X}$, i.e. $\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{z}_i$, where $\boldsymbol{z}_i$ is the representation coefficient vector of $\boldsymbol{y}_i$. Thus, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z}$, where $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_M] \in \mathbb{R}^{N \times M}$ is the representation matrix of $\boldsymbol{Y}$.

SRC[12] imposes the sparsest constraint on the representation vector to make it discriminative. However, it does not consider the subspace structures of data, which is essential to classification tasks. As we know from [19], low-rank model can reveal the subspace structures of data. Therefore, we impose both lowest-rank and sparsest constraints on the representation matrix $\boldsymbol{Z}$ and formulate the optimization problem as follows:

$$\min_{\boldsymbol{Z}} \text{rank}\,(\boldsymbol{Z}) + \beta\|\boldsymbol{Z}\|_0, \quad s.t. \ \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z}, \quad (8)$$

where β is a scalar constant to balance low-rankness and sparsity.

As the optimization problem (8) is non-convex, we substitute the L_1-norm for L_0-norm and the nuclear norm for the rank[19]. Therefore, the optimization problem (8) is relaxed as follows:

$$\min_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{Z}\|_1, \quad s.t. \ \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z}. \quad (9)$$

In many real-world applications, face images are usually corrupted or occluded. Thus, the test data matrix can be rewritten as $\boldsymbol{Y} = \boldsymbol{Y_0} + \boldsymbol{E} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}$, in which $\boldsymbol{Y_0}$ is the clean test data matrix and $\boldsymbol{E}$ is the error matrix. The optimization problem is rewritten as follows:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{Z}\|_1 + \lambda\|\boldsymbol{E}\|_1$$
$$s.t. \ \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \quad (10)$$

where λ is a scalar parameter.

3.2 Optimization

By introducing an auxiliary variable $\boldsymbol{W}$, the problem (10) can be converted into an equivalent optimization problem as follows:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{W}\|_1 + \lambda\|\boldsymbol{E}\|_1$$
$$s.t. \ \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \boldsymbol{Z} = \boldsymbol{W}. \quad (11)$$

The optimization problem (11) can be solved by the linearized alternating direction method with adaptive penalty (LADMAP) [24−26]. We give the augmented Lagrange function of problem (11) as follows:

$$
\begin{aligned}
\text{L}(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}, \boldsymbol{T}_1, \boldsymbol{T}_2, \mu) &= \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{W}\|_1 + \lambda\|\boldsymbol{E}\|_1 \\
&+ \langle \boldsymbol{T}_1, \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z} - \boldsymbol{E} \rangle + \langle \boldsymbol{T}_2, \boldsymbol{Z} - \boldsymbol{W} \rangle \\
&+ \frac{\mu}{2}\left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z} - \boldsymbol{E}\|_F^2 + \|\boldsymbol{Z} - \boldsymbol{W}\|_F^2\right),
\end{aligned}
\quad (12)
$$

where $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are Lagrangian multipliers, and $\mu > 0$ is a penalty parameter. By some simple algebra, the above augmented Lagrangian function is rewritten as follows:

$$
\begin{aligned}
\text{L}(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}, \boldsymbol{T}_1, \boldsymbol{T}_2, \mu) &= \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{W}\|_1 + \lambda\|\boldsymbol{E}\|_1 \\
&+ \text{h}(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}, \boldsymbol{T}_1, \boldsymbol{T}_2, \mu) - \frac{1}{2\mu}\left(\|\boldsymbol{T}_1\|_F^2 + \|\boldsymbol{T}_2\|_F^2\right),
\end{aligned}
\quad (13)
$$

where

$$
\begin{aligned}
&\text{h}\,(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}, \boldsymbol{T}_1, \boldsymbol{T}_2, \mu) \\
&= \frac{\mu}{2}\left(\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z} - \boldsymbol{E} + \frac{\boldsymbol{T_1}}{\mu}\right\|_F^2 + \left\|\boldsymbol{Z} - \boldsymbol{W} + \frac{\boldsymbol{T_2}}{\mu}\right\|_F^2\right).
\end{aligned}
\quad (14)
$$

We minimize the function L with other two variables fixed to update the variables $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}$ alternately. In each iteration, the updating rules are given as follows:

$$
\begin{aligned}
\boldsymbol{Z}_{k+1} &= \arg\min_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_* \\
&+ \langle \nabla_{\boldsymbol{Z}}\text{h}(\boldsymbol{Z}_k, \boldsymbol{W}_k, \boldsymbol{E}_k, \boldsymbol{T}_{1,k}, \boldsymbol{T}_{2,k}, \mu_k), \boldsymbol{Z} - \boldsymbol{Z}_k \rangle \\
&+ \frac{\mu_k \text{L}}{2}\|\boldsymbol{Z} - \boldsymbol{Z}_k\|_F^2 \\
&= \arg\min_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_* + \frac{\mu_k \text{L}}{2}\|\boldsymbol{Z} - \boldsymbol{A}_k\|_F^2 \\
&= \Theta_{(\mu_k \text{L})^{-1}}\,(\boldsymbol{A}_k),
\end{aligned}
\quad (15)
$$

where

$$
\begin{aligned}
\boldsymbol{A}_k &= \boldsymbol{Z}_k + \boldsymbol{X}^T\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z}_k - \boldsymbol{E}_k + \boldsymbol{T}_{1,k}/\mu_k\right)\Big/\text{L} \\
&- (\boldsymbol{Z}_k - \boldsymbol{W}_k + \boldsymbol{T}_{2.k}/\mu_k)/\text{L},
\end{aligned}
$$

$\text{L} = \|\text{X}\|_2^2$, Θ is the singular value thresholding operator[27], and $\nabla_{\boldsymbol{Z}}\text{h}$ is the partial differential of function h with respect to $\boldsymbol{Z}$.

$$
\begin{aligned}
\boldsymbol{W}_{k+1} &= \arg\min_{\boldsymbol{W}} \beta\|\boldsymbol{W}\|_1 \\
&+ \frac{\mu_k}{2}\left\|\boldsymbol{Z}_{k+1} - \boldsymbol{W}_k + \frac{\boldsymbol{T}_{2,k}}{\mu_k}\right\|_F^2 \\
&= \text{S}_{\beta\mu_k^{-1}}\left(\boldsymbol{Z}_{k+1} + \frac{\boldsymbol{T}_{2,k}}{\mu_k}\right)
\end{aligned}
\quad (16)
$$

$$
\begin{aligned}
\boldsymbol{E}_{k+1} &= \arg\min_{\boldsymbol{E}} \lambda\|\boldsymbol{E}\|_1 \\
&+ \frac{\mu_k}{2}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z}_{k+1} - \boldsymbol{E} + \frac{\boldsymbol{T}_{1,k}}{\mu_k}\right\|_F^2 \\
&= \text{S}_{\lambda\mu_k^{-1}}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z}_{k+1} + \frac{\boldsymbol{T}_{1,k}}{\mu_k}\right)
\end{aligned}
\quad (17)
$$

In (16) and (17), S is the soft-thresholding operator[28]. The complete procedure is outlined in **Algorithm 1**.

Algorithm 1 LADMAP algorithm for low-rank sparse representation

Input: training data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C] \in \mathbb{R}^{D \times N}$; test data matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M] \in \mathbb{R}^{D \times M}$;
     and parameters $\lambda > 0$; $\beta > 0$.

Initialize: $\boldsymbol{Z}_0 = \boldsymbol{W}_0 = \boldsymbol{E}_0 = \boldsymbol{T}_{1,0} = \boldsymbol{T}_{2,0} = 0$, $\mu_0 = 0.1$, $\mu_{max} = 10^{10}$, $\rho_0 = 1.1$, $\varepsilon_1 = 10^{-2}$, $\varepsilon_2 = 10^{-6}$, $\mathrm{L} = \|\boldsymbol{X}\|_2^2$,
     $maxiter = 1000$, $k = 0$.

while no converged and $k \le maxiter$ **do**

1: update the variable $\boldsymbol{Z}$ according to (15);

2: update the variable $\boldsymbol{W}$ according to (16);

3: update the variable $\boldsymbol{E}$ according to (17).

4: update the Lagrange multipliers:
    $\boldsymbol{T}_{1,k+1} = \boldsymbol{T}_{1,k} + \mu_k (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z}_{k+1} - \boldsymbol{E}_{k+1})$;
    $\boldsymbol{T}_{2,k+1} = \boldsymbol{T}_{2,k} + \mu_k (\boldsymbol{Z}_{k+1} - \boldsymbol{W}_{k+1})$.

5: update the parameter μ:
    $\mu_{k+1} = \min(\rho\mu_k, \mu_{\max})$, where
$$\rho = \begin{cases} \rho_0, & \mu_k \max\left(\sqrt{\mathrm{L}}\|\boldsymbol{Z}_{k+1} - \boldsymbol{Z}_k\|_F, \|\boldsymbol{W}_{k+1} - \boldsymbol{W}_k\|_F, \|\boldsymbol{E}_{k+1} - \boldsymbol{E}_k\|_F\right) \Big/ \|\boldsymbol{Y}\|_F < \varepsilon_1, \\ 1, & otherwise. \end{cases}$$

6: check the convergence condition:
    $\mu_{k+1}\max\left(\sqrt{\mathrm{L}}\|\boldsymbol{Z}_{k+1} - \boldsymbol{Z}_k\|_F, \|\boldsymbol{W}_{k+1} - \boldsymbol{W}_k\|_F, \|\boldsymbol{E}_{k+1} - \boldsymbol{E}_k\|_F\right)\Big/\|\boldsymbol{Y}\|_F < \varepsilon_1$ and $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Z}_{k+1} - \boldsymbol{E}_{k+1}\|_F/\|\boldsymbol{Y}\|_F < \varepsilon_2$.

7: $k \leftarrow k + 1$

end while

Output: an optimal solution $(\boldsymbol{Z}_k, \boldsymbol{W}_k, \boldsymbol{E}_k)$.

3.3 Classification

Given a training data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C] \in \mathbb{R}^{D \times N}$ and a test data matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M] \in \mathbb{R}^{D \times M}$, we can find the lowest-rank and sparsest representation matrix $\hat{\boldsymbol{Z}} = [\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_M]$ and the error matrix $\hat{\boldsymbol{E}} = [\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \cdots, \hat{\boldsymbol{e}}_M]$ of $\boldsymbol{Y}$ by Algorithm 1. Let $\delta_i(\hat{\boldsymbol{z}}_j)$ be a vector whose only nonzero entries are the entries in $\hat{\boldsymbol{z}}_j$ that are associated with class i[12]. Using $\delta_i(\hat{\boldsymbol{z}}_j)$, the test sample $\boldsymbol{y}_j$ can be reconstructed, i.e. $\hat{\boldsymbol{y}}_j^i = \boldsymbol{X}\delta_i(\hat{\boldsymbol{z}}_j) + \hat{\boldsymbol{e}}_j$. Then, the test sample $\boldsymbol{y}_j$ is classified into the class that minimizes the class reconstruction error between $\boldsymbol{y}_j$ and $\hat{\boldsymbol{y}}_j^i$:

$$\text{identity}(\boldsymbol{y}_j) = \arg\min_i \|\boldsymbol{y}_j - \boldsymbol{X}\delta_i(\hat{\boldsymbol{z}}_j) - \hat{\boldsymbol{e}}_j\|_2. \quad (18)$$

The proposed LRSRC algorithm is summarized in **Algorithm 2**.

4 Experiments

We perform experiments on Extended Yale B[29], CMU PIE[30], and AR databases[31] to evaluate the performance of our proposed LRSRC method. We compare LRSRC with some state-of-the-art methods, including nearest neighbor classifier (NN)[9], SRC[12], ESRC[16], CRC[17] and LRC[18]. Note that PCA is used to reduce the dimension of face image and the eigenface features are used for all the methods.

4.1 Extended Yale B database

The Extended Yale B database[29] consists of 2414 face images of 38 individuals. Each individual contains about 64 images, taken under various laboratory-controlled lighting conditions. Fig. 1(a) shows several example images of one subject in the Extended Yale B database. In our experiments, each image is manually cropped and resized to 32×32 pixels. We randomly select l (=16, 24, 32) samples per individual for training, and the remainder for testing.

For each given l, we independently perform all the methods five times, and report the average five-fold recognition rates. The eigenface feature dimensions are set as 50, 100, 150, 200, 300 and 500 for the comparison of recognition rates across different methods. We set LRSRC parameters as $\lambda = 10$, $\beta = 0.09, 0.1$ and 0.08 for the 16, 24, and 32 training samples per individual, respectively. Fig. 2 shows the curves of recognition rate versus feature dimension. Table 1 gives the best recognition rates of all the methods.

From Fig. 2, one can see that our method exceeds all the other methods in most of dimensions under different training sets. For example, with a feature dimension of 200, our method outperforms all the other methods by more than 3% improvement of recognition rate under three different training sets. Meanwhile, it can be seen from Table 1 that the best recognition rates of our method are always greater than those of the other methods. Especially, our method achieves the best recognition rate of 94.87% when 16 samples per individual are used for training, while SRC merely arrives at the best recognition rate of 94.26% even when 32 samples per individual are used for training. This indicates
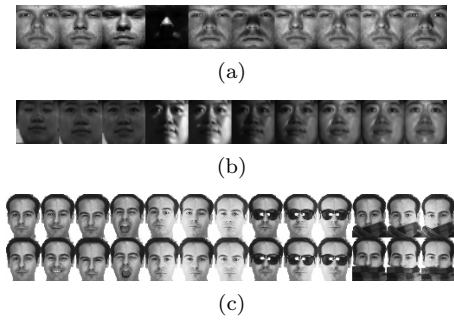

(a)

(b)

(c)

Fig. 1 Example images of the three databases: (a) Extended Yale B database; (b) CMU PIE database; (c) AR database.

Algorithm 2 Low-rank sparse representation-based classification (LRSRC)

Input: training data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C] \in \mathbb{R}^{D \times N}$ from C classes; test data matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M]$
   $\in \mathbb{R}^{D \times M}$; parameters $\lambda > 0$, $\beta > 0$.

1: normalize the columns of $\boldsymbol{X}$ to have unit L_2-norm.

2: Solve the following optimization problem to obtain the representation matrix $\hat{\boldsymbol{Z}}$ and the error matrix $\hat{\boldsymbol{E}}$ of test data matrix $\boldsymbol{Y}$ by Algorithm 1:
$$\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{Z}\|_1 + \lambda\|\boldsymbol{E}\|_1 \quad s.t. \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}$$

3: for $j = 1 : M$ do
$$\text{identity}(\boldsymbol{y}_j) = \arg\min_i \|\boldsymbol{y}_j - X\delta_i(\hat{\boldsymbol{z}}_j) - \hat{\boldsymbol{e}}_j\|_2, \quad i = 1, \cdots, C.$$
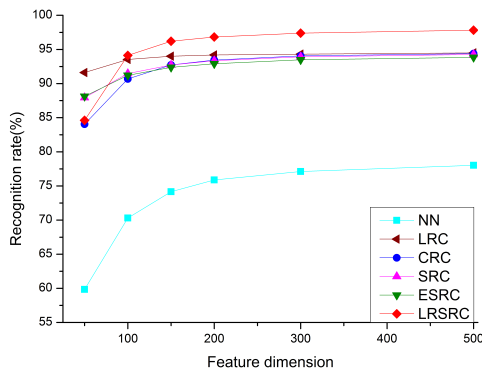   end for

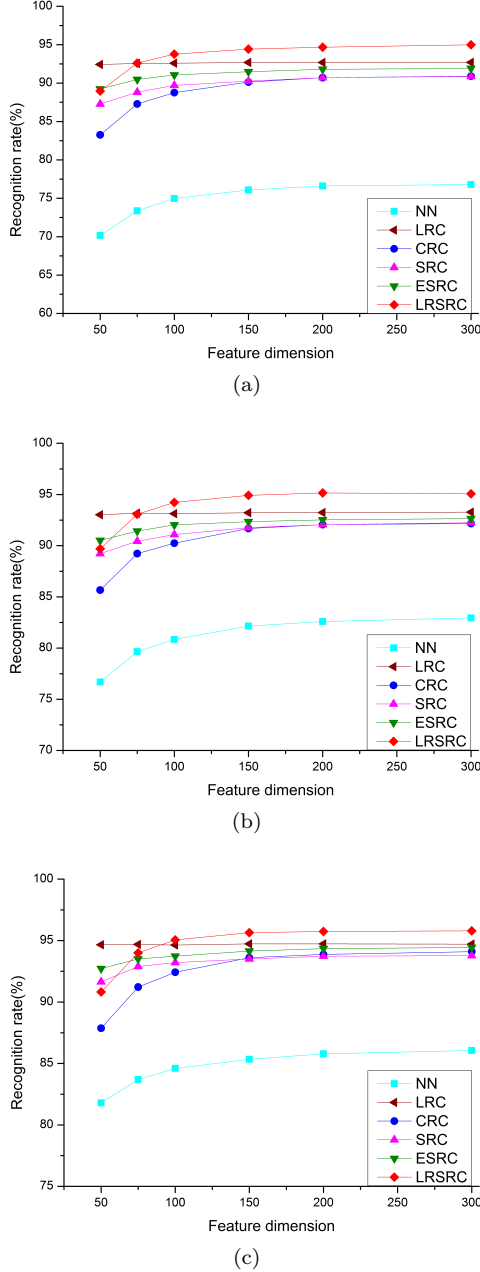**Output**: the class labels of all the test samples.
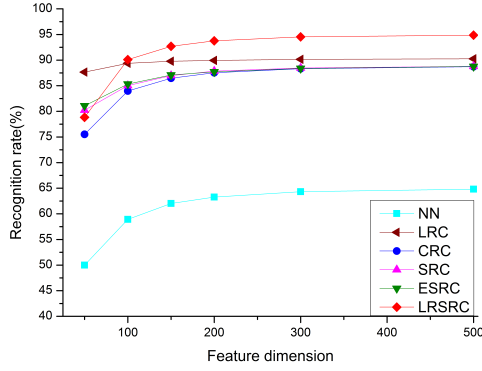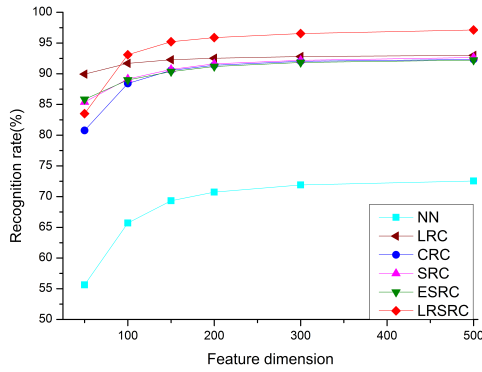
Fig. 2 Recognition rate versus feature dimension of all the methods on Extended Yale B database: (a) 16 training samples per individual; (b) 24 training samples per individual; (c) 32 training samples per individual.

Table 1 The best recognition rates (%) and the corresponding dimensions (in the parentheses) of all the methods on Extended Yale B database.

Method	$Train16$	$Train24$	$Train32$
NN[9]	64.82(500)	72.54(500)	78.03(500)
LRC[18]	90.25(500)	93.00(500)	94.49(500)
CRC[17]	88.70(500)	92.33(500)	94.39(500)
SRC[12]	88.78(500)	92.64(500)	94.26(500)
ESRC[16]	88.76(500)	92.22(500)	93.84(500)
LRSRC	**94.87**(500)	**97.14**(500)	**97.83**(500)

that our method can achieve a good recognition performance, when the number of training samples is small. From the experimental results, we confirm the effectiveness and robustness of our method even if face images are corrupted by illumination variations.

4.2   CMU PIE database

The CMU PIE database[30] contains over 40,000 face images of 68 individuals. Images of each individual were acquired across 13 different poses, under 43 different illumination conditions, and with 4 different expressions. Fig. 1(b) shows some example images of one subject in the CMU PIE database. Here, we use a near frontal pose subset, namely C07, for experiments, which contains 1629 images of 68 individuals, each individual with about 24 images. In our experiment, all images are manually cropped and resized to be 32×32 pixels. A random subset with l $(= 8, 10, 12)$ samples per individual is selected for training, and the rest for testing. For each given l, we independently run all the methods five times, and report the average five-fold recognition rates. The eigenface feature dimensions are selected as 50, 75, 100, 150, 200 and 300 for the comparison of face recognition rates across different methods. The LRSRC parameters are set as $\lambda = 10$, $\beta = 0.09$ for these experiments. The curves of recognition rate versus feature dimension are presented in Fig. 3. The best recognition rates of all the methods are shown in Table 2.

From Fig. 3, one can clearly see that our method outperforms all the other methods in most of dimensions under three training sets. In detail, the recognition rates of our methods are about 1.5% greater than those of the other methods in most of dimensions for 8 and 10 training samples per individual, and 1% for 12 training samples per individual. Furthermore, one can know from Table 2 that the best recognition rate of our method is always greater than

those of the other methods for each training set. In addition, one can also find from Table2 that SRC achieves the best recognition rate of 93.78% when 12 samples per individual are used for training, which is still lower than that of our method, even when 8 samples per individual are used for training. It indicates again that our method can obtain good recognition performance, when the number of training samples is small. The experimental results further verify the effectiveness and robustness of our method for dealing with face images corrupted by illumination variations and expression changes.

Table 2   The best recognition rates (%) and the corresponding dimensions (in the parentheses) of all the methods on CMU PIE database.

Method	$Train8$	$Train10$	$Train12$
NN[9]	76.79(300)	82.95(300)	86.05(300)
LRC[18]	92.70(300)	93.28(300)	94.74(200)
CRC[17]	90.89(300)	92.16(300)	94.10(300)
SRC[12]	90.86(300)	92.24(300)	93.78(300)
ESRC[16]	91.91(300)	92.64(300)	94.44(300)
LRSRC	**94.99**(300)	**95.15**(200)	**95.79**(300)

4.3   AR database

The AR database[31] includes over 4000 color face images of 126 individuals. For each individual, 26 images were taken in two separated sessions. In each session, there are 13 images for each person, among which 3 images are occluded by sunglasses, 3 by scarves, and the remaining 7 with illumination and expression variations (and thus are named as neutral images). All images are cropped to 50×40 pixels and converted into gray scale. As done in [10], we choose a subset containing 50 male subjects and 50 female subjects in the experiments. Fig. 1(c) shows whole images of one subject in the AR database. Following the scheme in [32], we consider three scenarios to evaluate the performance of our method.

Sunglasses: In this scenario, we take into account the situation in which both the training and test samples are corrupted by sunglass occlusion. Sunglasses produce about 20% occlusion of the face image. We use 7 neutral images and 1 image with sunglasses from the first session for training, 7 neutral images from the second session and the remaining 5 images with sunglasses from two sessions for testing.

Scarf: In this scenario, we consider corruptions of both the training and test images, because of scarf occlusion. Scarf generates around 40% occlusion of the face image. Similarly, we choose 8 training images consisting of 7 neutral images plus 1 image with scarf from the first session for training, and 12 test images consisting of 7 neutral images from the second session and the remaining 5 images with scarf from two sessions for testing.

Mixed (Sunglasses + Scarf): In this scenario, we consider the case in which both the training and test images are occluded by sunglasses and scarf. We select 7 neutral images plus 2 occluded images (one image with sunglass and another with scarf) from the first session for training and the remaining 17 images in two sessions for testing.

In this experiment, we select eigenface feature dimensions as 50, 100, 150, 200, 300 and 500 for the comparison of recognition rates across different methods. The LRSRC parameters are set as $\lambda = 10$, $\beta = 0.09$. Fig. 4 gives the recognition results of three different scenarios. Table 3 summarizes the best recognition rates of all the methods under three different scenarios.

From Fig. 4, one can see that the proposed method consistently gives the highest recognition rate in most of dimensions under three different scenarios. At the same time, one can find from Table 3 that the best recognition rate of our method is highest for each scenario. Specifically, our method outperforms the best competitor SRC by more th-

Fig. 3 Recognition rate versus feature dimension of all the methods on CMU PIE database: (a) 8 training samples per individual; (b) 10 training samples per individual; (c) 12 training samples per individual.

Table 3 The best recognition rates (%) and the corresponding dimensions (in the parentheses) of all the methods on the AR database.
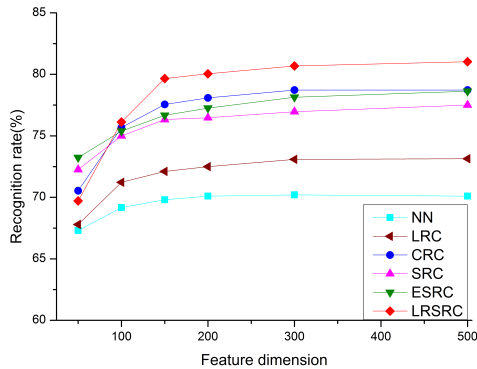
Method	*Sunglass*	*Scarf*	*mixed*
NN[9]	71.18(300)	67.08(300)	70.20(300)
LRC[18]	72.85(500)	70.35(500)	73.14(500)
CRC[17]	78.40(300)	76.60(500)	78.73(500)
SRC[12]	78.96(500)	75.35(500)	77.50(500)
ESRC[16]	78.47(500)	76.53(500)	78.63(300)
LRSRC	**80.21**(500)	**79.31**(300)	**81.03**(300)
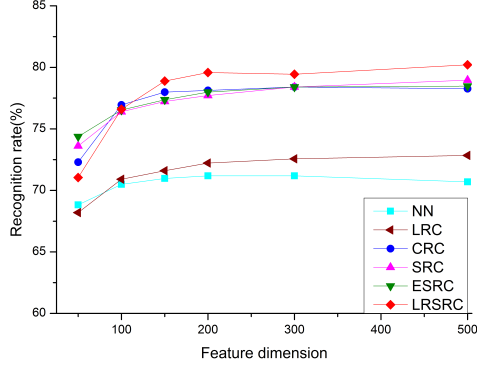
(a)
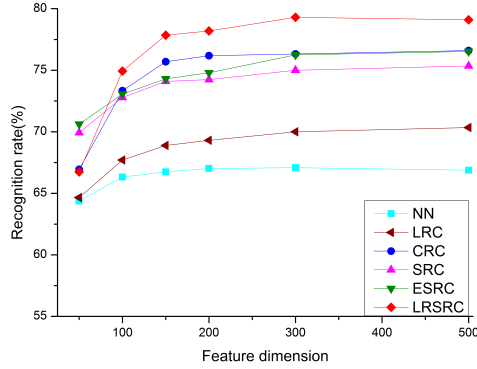
(b)

(c)

Fig. 4. Recognition rate curves of all the methods on the AR database: (a) Sunglass; (b) Scarf; (c) Mixed.

an 1% improvement of the best recognition rate for the Sunglass scenario, 3.5% improvement of the best recognition rate for the Scarf and Mixed scenarios. Moreover, one

can easily find from Table 3 that our method also exceeds the other four methods under three different scenarios. The experiment results demonstrate the effectiveness and robustness of our method for processing occluded face images.

5 Conclusions

In this paper, we present a low-rank sparse representation-based classification (LRSRC) method for robust face recognition. LRSRC can efficiently find the lowest-rank and sparsest representation matrix of a set of test samples over all training samples, and then classifies each test sample into the correct class by using its representation vector. Experimental results on Extended Yale B, CMU PIE, and AR databases confirm that our proposed method is effective and robust, even when face images are corrupted by illumination variations, expression changes, disguises, or occlusions. It must be pointed out that our proposed LRSRC mainly performs the classifications based on the entire test set to be corporated. In our future work, we will investigate how to extend LRSRC to the single test sample scenarios.

References

[1] J.B. Tenenbaum, V. D. Silva, J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, vol. 290, no. 12, pp. 2319–2322, 2000.

[2] S.T. Roweis, L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, vol. 290, no. 12, pp. 2323–2326, 2000.

[3] M. Turk, A. Pentland. Face recognition using eigenfaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–591, 1991.

[4] P.N. Belhumeur, J.P. Hespanha, D.J. kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[5] X. He, S.Yan, Y. Hu, P. Niyogi, H. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[6] S. Yan, D. Xu, B. Zhang, Q. Yang, S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[7] H.T. Chen, H.W. Chang, T.L. Liu. Local discriminant embedding and its variants. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 846–853, 2005.

[8] G.Q. Wang, L.X. Li, X.B. Guo. Sparsity preserving discriminant embedding for face recognition. *Chinese Journal of Scientific Instrument*, vol. 35, no. 2, pp. 305–312, 2014.

[9] R. Duda, P. Hart, D. Stork. *Pattern Classification*, 2nd ed., John Wiley & Sons, 2001.

[10] F. De la Torre, M.J. Black. Robust principal component analysis for computer vision. In *Proceedings of Eighth IEEE International Conference on Computer Vision (IC-CV)*, pp. 362–369, 2001.

[11] Z.Z Feng, M. Yang, L. Zhang, Y. Liu, D. Zhang. Joint discriminative dimensionality reduction and dictionary learning for face recognition. *Pattern Recognition*, vol. 46, pp. 2135–2143, 2013.

[12] J. Wright, A.Y. Yang, A. Ganesh, S. Sastry, Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[13] S.S. Chen, D.L. Donoho, M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[14] S.J. Kim, K. Koh, M. lustig, S. Boyd, D. Gorinevsky. An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[15] J. Yang, Y. Zhang. Alternating direction algorithm for l1-problems in compressive sensing. *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[16] W. Deng, J. Hu, J. Guo. Extended SRC: undersampled face recognition via intra-class variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1864–1870, 2012.

[17] L. Zhang, M. Yang, X. Feng, Y. Ma, D. Zhang. Collaborative representation based classification for face recognition. *CoRR*, vol. abs/1204.2358v2, arXiv:1204.2358v2, 2014.

[18] I. Naseem, R. Togneri, M. Bennamoun. Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.

[19] G. Liu, Z. Lin, S. Yan, Y. Yu, Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[20] X.Q. Lu, Y.L. Wang, Y. Yuan. Graph-regularized low-rank representation for destriping of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 4009–4018, 2013.

[21] L. Ma, C.H. Wang, B.H. Xiao, W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *Proceedings of of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2586–2593, 2012.

[22] Y.M.Z. Zhang, Z.L. Jiang, L.S. Davis. Learning Structured Low-rank Representation for Image Classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 676–683, 2013.

[23] D. Bertsekas. *Constrained optimization and Lagrange multiplier methods*, vol. 410, Academic Press New York, 1982.

[24] Z. Lin, R. Liu, Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Proceedings of of NIPS*, 2011.

[25] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, N. Yu. Non-negative low rank and sparse graph for semi-supervised learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2328–2335, 2012.

[26] Y. Zhang, Z. Jiang, L. Davis. Learning structured low-rank representations for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 676–683, 2013.

[27] J. Cai, E. Cands, Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[28] Z. Lin, M. Chen, L. Wu. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, UIUC Technical Report, UILU-ENG-09-2215, 2009.

[29] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[30] T. Sim, S. Baker, M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[31] A.M. Martinez, R. Benavente. The AR face database. CVC Technical Report 24, 1998.

[32] C.F. Chen, C.P. Wei, Y.C.F. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2618–2625, 2012.
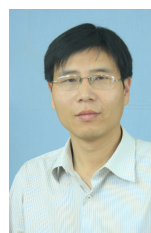
Haishun Du   received his Ph.D. degree from Southeast University, Nanjing, China, in 2007. Now, he is an associate professor at the College of Computer and Information Engineering, Henan University, China. His research interests include pattern recognition, computer vision and digital image processing.

E-mail: jddhs@henu.edu.cn (Corresponding author)

ORCID iD: 0000-0003-0883-8118
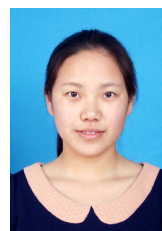
Qingpu Hu   received her B.S. degree from Henan University, China, in 2013. Now, she is working for her M.S. degree in Pattern Recognition & Intelligent Systems at Henan University, China. Her research interests include pattern recognition and computer vision.

E-mail: huqingpu_henu@126.com

Dianfeng Qiao   received his B.S. degree from Henan University, China, in 2013. Now, he is working for his M.S. degree in Control Theory & Control Engineering at Henan University, China. His research interests include pattern recognition and fault detection.

E-mail: qdfeng_henu@126.com

Ioannis Pitas   received the Diploma of Electrical Engineering in 1980 and the PhD degree in Electrical Engineering in 1985, both from the University of Thessaloniki, Greece. Since 1994 he has been a Professor at the Department of Informatics, University of Thessaloniki, Greece.   He is an IEEE fellow (2007), IEEE Distinguished Lecturer (2012) and EURASIP fellow (2013). His current interests are in the areas of intelligent digital media, digital imageideo processing, multimedia signal processing, human centered computing computational intelligence and computer vision.

E-mail: pitas@aiia.csd.auth.gr