Pattern Recognition 62 (2017) 87–98

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Regularized coplanar discriminant analysis for dimensionality reduction ☆

Ke-Kun Huang [a], Dao-Qing Dai [b,*], Chuan-Xian Ren [b]

[a] School of Mathematics, JiaYing University, Meizhou, Guangdong, 514015, China
[b] Intelligent Data Center and Department of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form
26 May 2016
Accepted 22 August 2016
Available online 26 August 2016

Keywords:
Dimensionality reduction
Sparse representation classifier
Face recognition
Hyperspectral image classification
Coplanar discriminant analysis

ABSTRACT

The dimensionality reduction methods based on linear embedding, such as neighborhood preserving embedding (NPE), sparsity preserving projections (SPP) and collaborative representation based projections (CRP), try to preserve a certain kind of linear representation for each sample after projection. However, in the transformed low-dimensional space, the linear relationship between the samples may be changed, which cannot make the linear representation-based classifiers, such as sparse representation-based classifier (SRC), to achieve higher recognition accuracy. In this paper, we propose a new linear dimensionality reduction algorithm, called Regularized Coplanar Discriminant Analysis (RCDA) to address this problem. It simultaneously seeks a linear projection matrix and some linear representation coefficients that make the samples from the same class coplanar and the samples from different classes not coplanar. The proposed regularization term balances the bias from the optimal linear representation and that from the class mean to avoid overfitting the training data, and overcomes the matrix singularity in solving the linear representation coefficients. An alternative optimization approach is proposed to solve the RCDA model. Experiments are done on several benchmark face databases and hyperspectral image databases, and results show that RCDA can obtain better performance than other dimensionality reduction methods.

1. Introduction

In many fields of pattern classification, such as face recognition, hyperspectral image classification, object categorization, action recognition and target detection, the original data is usually provided in high-dimensional form, while the underlying structure in many cases can be characterized by a small number of features. Dimensionality reduction (DR) is necessary and helpful for classification. In the past decades, many useful techniques for DR have been developed, such as principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2], Fisherface [3], maximum margin criterion (MMC) [4], regularized discriminant analysis (RDA) [5], locality preserving projections (LPP) [6], and marginal fisher analysis (MFA) [7].

In recent years, researchers proposed a number of methods to improve the above traditional DR methods. Lu et al. proposed a parametric regularized LPP [8], which results in better locality preserving power. Deng et al. proposed a transform-invariant PCA [9], characterizing accurately the intrinsic structures of the human face that are invariant to the in-plane transformations. Lu et al. proposed a sparse exponential family PCA method [10], to achieve simultaneous dimension reduction and variable selection for better interpretation of the results. Oh et al. presented a generalized mean PCA [11], overcoming the problem that PCA is prone to outliers included in the training set. Abou-Moustafa et al. presented a Pareto LDA [12], to maximize each pairwise distance to maximally separate all class mean. Ghassabeh et al. proposed a fast incremental LDA [13], which accelerates the convergence rate of the incremental LDA algorithm. Ren et al. proposed the outlier Suppressing LDA [14], exploring the importance of the sample itself in building the optimal subspace. Wang et al. proposed a semi-supervised LDA [15], which can use limited number of labeled data and a quantity of the unlabeled ones for training. Ji et al. proposed a relevance MFA [16], to model the pairwise constraints of relevance-link and irrelevance-link into the relevance graph and irrelevance graph.

The above DR methods all consider the property about

☆This work is supported in part by National Science Foundation of China under Grants 61375033, 61403164, 61572536 and 11631015, in part by the Fundamental Research Funds for the Central Universities under Grant 161GZD16, and in part by the Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (2013LYM_0085).

* Corresponding author.
E-mail addresses: kkcocoon@163.com (K.-K. Huang), stsddq@mail.sysu.edu.cn (D.-Q. Dai), rchuanx@mail.sysu.edu.cn (C.-X. Ren).

http://dx.doi.org/10.1016/j.patcog.2016.08.024
0031-3203/

Euclidean distance between pairwise points, while there are some DR methods which consider the property about the distance between a sample and the linear combination of other samples, called linear embedding property, such as neighborhood preserving embedding (NPE) [17], sparsity preserving projections (SPP) [18,19] and collaborative representation based projections (CRP) [20]. NPE [17] represents each data point as a linear combination of the neighboring points. Then it finds an optimal embedding such that the neighborhood structure can be preserved in the subspace. Recently, many discriminative versions of NPE have been proposed, such as neighbourhood preserving discriminant embedding (NPDE) [21] and double adjacency graphs-based discriminant neighborhood embedding (DAG-DNE) [22]. Different from NPE, where the nearest neighbors are manually chosen, SPP [18,19] automatically constructs a graph. It aims to preserve the sparse linear reconstructive relationship of the data, which is achieved by minimizing a l_1 regularization objective function. Because SPP is unsupervised too, recently, some works consider its discriminative version, such as LDSNPE [23], GDSNPE [24], discriminative SPP [25] and sparse discriminative multi-manifold embedding (SDMME) [26]. Like SPP, CRP [20] aims to preserve the collaborative representation based reconstruction relationship of data. CRP is much faster than SPP since CRP calculates the objective function with l_2 regularization while SPP calculates the objective function with l_1 regularization.

NPE, SPP and CRP try to preserve a certain kind of linear representation for each sample after projection. However, in the transformed low-dimensional space, the linear relationship between the samples may be changed. In other words, the linear representation coefficients in the original high-dimensionality space for a sample may be different from those in the transformed low-dimensional space. But the classifiers work in the transformed space. Recently, sparse representation based classifier (SRC) has been successfully used in pattern classification [27,28], and there are many works to extend SRC [29–33]. SRC first codes a testing sample as a sparse linear representation of all the training samples, and then classifies the testing sample by evaluating which class leads to the minimum representation error. SRC supposes that the sparse nonzero representation coefficients concentrate on the training samples with the same class label as the testing sample. In other words, a testing sample should be linearly represented by the samples from the same class better than those from other classes to achieve higher recognition accuracy. The dimensionality reduction methods based on linear embedding meet the characteristic of SRC at a certain degree, but their performances are still limited, because the linear relationship between the samples may be changed in the transformed low-dimensional space.

To achieve better performance based on SRC, Yang et al. proposed a method called SRC steered discriminative projection (SRCDP) [34]. Observing that SRC adopts a class reconstruction residual-based decision rule, SRCDP maximizes the ratio of between-class reconstruction residual to within-class reconstruction residual in the transformed low-dimensional space. But the within-class reconstruction residual is too big to be minimized by projection, and it is influenced by the samples from different classes, which limits SRCDP to achieve better performance. To design a more effective DR method for SRC, we propose a new method called regularized coplanar discriminant analysis (RCDA).

The contributions of this paper are listed as follows:

1. We propose the coplanar projection model, which simultaneously finds a projection matrix and some linear representation coefficients such that the sample from the same class in the same hyperplane as much as possible.
2. We propose the model of regularized coplanar discriminant analysis (RCDA). RCDA makes the samples from the same class coplanar and the samples from different classes not coplanar. The linear representation coefficients are regularized by the proposed mean l_2 norm, which can balance the bias from the optimal linear representation and that from the class mean to avoid overfitting the training data, and overcomes the matrix singularity in solving the linear representation coefficients.
3. We propose an optimization algorithm to solve the model of RCDA, which alternatively calculates the projection matrix and the linear representation coefficients.

The remainder of the paper is organized as follows. In Section 2, we describe the proposed method in detail. In Section 3, we give some analyses of the proposed method. The experimental results are given in Section 4. Finally, the conclusion is provided in Section 5.

2. The proposed method

In this section, we describe the proposed method in detail. First, we discuss the basic idea of the proposed method in Section 2.1, only concerning the within-class representation for convenience. Then we propose the RCDA model by considering a supervised version of the model in Section 2.2. The corresponding optimization algorithm is proposed in Section 2.3.

2.1. RCDA: basic idea

The objective of the proposed method is to find a linear projection matrix that reduces the error of the within-class linear representation while preserves the error of the between-class linear representation in the transformed space.

2.1.1. Coplanar projection

To facilitate describing our method, we first discuss the within-class representation of ith training sample $\mathbf{x}_i$. Let

$$\mathbf{X}_i = \left\{ \mathbf{x}_j | c_j = c_i \quad \text{and} \quad j \neq i \right\} \in \mathbf{R}^{m \times n_i^w}$$

be the n_i^w training samples from the same class as $\mathbf{x}_i$ except $\mathbf{x}_i$. We want to simultaneously find a projection matrix $\mathbf{W} \in \mathbf{R}^{m \times d}$ and within-class linear representation coefficients $\boldsymbol{\beta}_i^w$ for each training sample such that the error of within-class linear representation is minimized after linear transformation, i.e:

$$\min_{\mathbf{W}, \boldsymbol{\beta}^w} \sum_{i=1}^{n} \left\| \mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{X}_i \boldsymbol{\beta}_i^w \right\|_2^2 \tag{1}$$

where $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{I}$ is the identity matrix, and $\boldsymbol{\beta}^w$ denotes all the linear representation coefficients $\left\{ \boldsymbol{\beta}_i^w | i = 1, \ldots, n \right\}$.

Note that the linear representation coefficients are calculated in the transformed lower-dimension space where the classifier actually works, instead of original space. Because the linear relationship between the samples may be changed after linear transformation, it will be more accurate to calculate the linear representation coefficients in the transformed space. If each sample can be linearly represented by the samples from the same class, then the samples from the same class are in the same hyperplane. So the coplanarity can be measured by the error of within-class linear representation. Model (1) finds some linear projection directions that make the training samples from the same class as much as possible in the same hyperplane after projection, so we call it *coplanar projection* model.

It should be mentioned that model (1) is a new model. The most similar work is NPE [17]. The linear representation

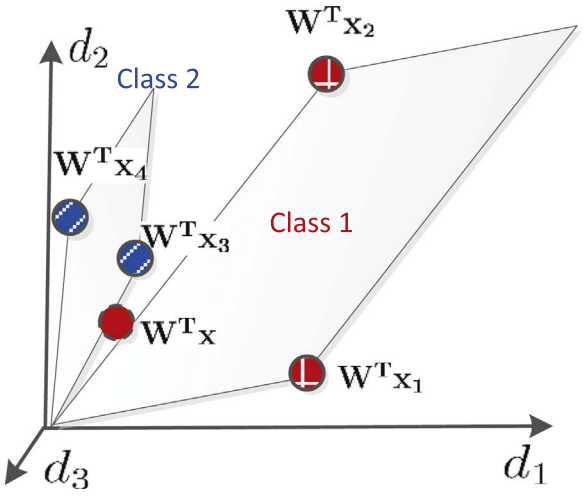

Fig. 1. An example when model (1) overfits the training samples. $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to class 1, while $\mathbf{x}_3$ and $\mathbf{x}_4$ belong to class 2. Model (1) makes the samples from the same class in the same plane, but the samples from class 1 are still far away from each other. If there is a testing sample $\mathbf{x}$ that belongs to class 1 but deviates a little from the plane of class 1 in the transformed space, it is easy to be classified into class 2 because it is close to class 2.

coefficients of NPE are calculated in the original space, while those of model (1) are calculated in the lower-dimensional transformed space where the classifier practically works. Model (1) simultaneously learns the projection matrix and the linear representation coefficients. So we use the term "*coplanar projection*" to represent our method, which is distinguished from the methods based on linear embedding such as NPE, SPP and CRP where the linear representation coefficients are calculated in the original space.

The coefficients β_i^w are influenced by the projection matrix $\mathbf{W}$, model (1) is hard to be optimized. We can get a suboptimal solution by alternatively calculating $\mathbf{W}$ and β_i^w until a predefined criterion is achieved.

2.1.2. Mean constrained projection

Model (1) may overfit the training samples. Though the transformed training samples from the same class are in the same hyperplane, they may be still far away from each other in the transformed space. Then, if a new transformed testing sample slightly deviates from the hyperplane of the same class, it may be wrongly classified. Fig. 1 shows an example in this case.

So, on the other hand, we also hope that the samples from the same class be clustered. In other words, the distance between each sample and the mean of the samples from the same class should be as small as possible, i.e.,

$$\min_{\mathbf{W}} \sum_{i=1}^{n} \left\| \mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{X}_i\tilde{\beta}_i^w \right\|_2^2 \tag{2}$$

where $\tilde{\beta}_i^w = \left[\frac{1}{n_i^w}, \frac{1}{n_i^w}, \ldots, \frac{1}{n_i^w}\right]^T \in \mathrm{R}^{n_i^w \times 1}$. In fact, model (2) is similar to the within-class objective of LDA.

2.1.3. Regularized coplanar projection

Models (1) and (2) are different, so they cannot be optimized at the same time. To balance the two models, we propose the following model to find a normalized projection matrix $\mathbf{W} \in \mathbf{R}^{m\times d}$ such that the error of within-class linear representation with mean l_2 regularization is minimized after linear transformation:

$$\min_{\mathbf{W},\beta^w} \sum_{i=1}^{n} \left(\left\| \mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{X}_i\beta_i^w \right\|_2^2 + \lambda \left\| \beta_i^w - \tilde{\beta}_i^w \right\|_2^2 \right) \tag{3}$$

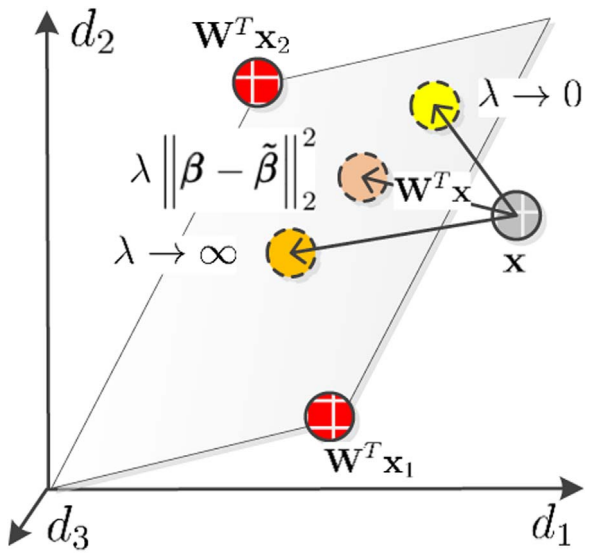

Fig. 2. Comparison of models (1), (2) and (3). The mean l_2 regularization, i.e., model (3), balances the error between $\mathbf{x}$ and the point of optimal linear representation (yellow point, $\lambda \to 0$) and that from class mean (orange point, $\lambda \to \infty$), which provides a more stable representation. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

It means that the linear combination $\mathbf{W}^T\mathbf{x}_i$ is hoped to be closed to the mean of $\mathbf{W}^T\mathbf{X}_i$. If $\lambda \to 0$, model (3) will become model (1); If $\lambda \to \infty$, model (3) will become model (2). λ balances the bias from the optimal linear representation and that from the class mean. On the other hand, when $\mathbf{W}$ is fixed, model (3) has a least square solution for β_i^w, and the l_2 regularization term makes the solution stable by avoiding to inverse a singular matrix, so it provides a more stable representation.

Fig. 2 shows the comparison of models (1), (2) and (3). The length of arrow denotes the distance between the corresponding two points. We can find that $\mathbf{x}$ is the closest to the point of the optimal linear representation (yellow point, $\lambda \to 0$), so it is the easiest to find a projection to cut down the distance. But the yellow point is the farthest from the mean (orange point, $\lambda \to \infty$), which implies that the samples from the same class may be scattered and it is bad for generalization for new testing sample. Model (3) balances the error between $\mathbf{x}$ and the point of optimal linear representation and that from the class mean to avoid overfitting the training data, and overcomes the matrix singularity in solving the linear representation coefficients, which provides a more stable representation.

2.2. RCDA: the model

Based on the previous subsection, where only within-class representation is discussed, we propose the RCDA model by considering a supervised version of model (3), where we want to find a linear projection that optimally reduces the error of the within-class representation while preserves the error of the between-class representation in the transformed low-dimensional space.

2.2.1. Within-class coplanar compactness

We rewrite model (3) into a matrix form for a compact representation, i.e., the sum of the errors of the within-class linear representation with mean l_2 regularization in the lower dimensional subspace, called within-class coplanar compactness, can be characterized by the term:

$$\sum_{i=1}^{n}\left(\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{X}_i\boldsymbol{\beta}_i^w\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_i^w-\tilde{\boldsymbol{\beta}}_i^w\right\|_2^2\right)=\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)+\lambda\left\|\mathbf{B}_w-\tilde{\mathbf{B}}_w\right\|_F^2, \tag{4}$$

where

$$\mathbf{S}_w=\mathbf{X}\left(\mathbf{I}-\mathbf{B}_w^T\right)\left(\mathbf{I}-\mathbf{B}_w\right)\mathbf{X}^T, \tag{5}$$

$\mathbf{X}\in\mathbf{R}^{m\times n}$ are all the n training samples, $\mathbf{X}_i$ are the samples of the same class as $\mathbf{x}_i$ except $\mathbf{x}_i$, and $\mathbf{B}_w\in\mathbf{R}^{n\times n}$ is the weight matrix composed by $\boldsymbol{\beta}_i^w$. Suppose $\boldsymbol{\eta}_i$ is the index vector of the samples of class c_i in $\mathbf{X}$ except $\mathbf{x}_i$, then $\mathbf{B}_w\left[\boldsymbol{\eta}_i,\,i\right]=\boldsymbol{\beta}_i^w$ and the other elements are equal to zero. Similarly, $\tilde{\mathbf{B}}_w\left[\boldsymbol{\eta}_i,\,i\right]=\tilde{\boldsymbol{\beta}}_i^w$ where $\tilde{\boldsymbol{\beta}}_i^w=\left[\frac{1}{n_i^w},\frac{1}{n_i^w},\ldots,\frac{1}{n_i^w}\right]^T\in\mathrm{R}^{n_i^w\times 1}$, and n_i^w is the number of samples in $\mathbf{X}_i$.

2.2.2. Between-class coplanar separability

Similarly, the sum of the errors of the between-class linear representation with mean l_2 regularization in the lower dimensional subspace, called between-class coplanar separability, can be characterized by the term:

$$\sum_{i=1}^{n}\sum_{c=1}^{C}\left(\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{X}^c\boldsymbol{\beta}_{i,c}^b\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_{i,c}^b-\tilde{\boldsymbol{\beta}}_c^b\right\|_2^2\right)=\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)+\lambda\sum_{c=1}^{C}\left\|\mathbf{B}_c-\tilde{\mathbf{B}}_c\right\|_F^2, \tag{6}$$

where

$$\mathbf{S}_b=\mathbf{X}\left(\sum_c\left(\mathbf{I}-\mathbf{B}_c^T\right)\left(\mathbf{I}-\mathbf{B}_c\right)\right)\mathbf{X}^T, \tag{7}$$

$\mathbf{X}^c$ is the samples of class c, C is the number of class, and $\mathbf{B}_c\in\mathrm{R}^{n\times n}$ is the weight matrix composed by $\boldsymbol{\beta}_{i,c}^b$. Suppose $\boldsymbol{\eta}_c$ is the index vector of the samples of class c in $\mathbf{X}$, then $\mathbf{B}_c\left[\boldsymbol{\eta}_c,\,i\right]=\boldsymbol{\beta}_{i,c}^b$. Similarly, $\tilde{\mathbf{B}}_c\left[\boldsymbol{\eta}_c,\,i\right]=\tilde{\boldsymbol{\beta}}_c^b$ where $\tilde{\boldsymbol{\beta}}_c^b=\left[\frac{1}{n_c},\frac{1}{n_c},\ldots,\frac{1}{n_c}\right]^T\in\mathrm{R}^{n_c\times 1}$, and n_c is the number of samples of class c.

2.2.3. The model of regularized coplanar discriminant analysis

We simultaneously seek a linear projection matrix $\mathbf{W}$, the within-class linear representation coefficients $\boldsymbol{\beta}_i^w$ and the within-class linear representation coefficients $\boldsymbol{\beta}_{i,c}^b$ to minimize the within-class coplanar compactness and maximize the between-class coplanar separability at the same time. According to Eqs. (4) and (6), the RCDA model can be summarized as follows:

$$\min_{\mathbf{W}}\frac{\min_{\boldsymbol{\beta}^w}\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)+\lambda\left\|\mathbf{B}_w-\tilde{\mathbf{B}}_w\right\|_F^2}{\min_{\boldsymbol{\beta}^b}\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)+\lambda\sum_{c=1}^{C}\left\|\mathbf{B}_c-\tilde{\mathbf{B}}_c\right\|_F^2} \tag{8}$$

where $\mathbf{W}\in\mathbf{R}^{m\times d}$, $\mathbf{W}^T\mathbf{W}=\mathbf{I}$, $\boldsymbol{\beta}^w$ denotes $\left\{\boldsymbol{\beta}_i^w|i=1,\ldots,n\right\}$, and $\boldsymbol{\beta}^b$ denotes $\left\{\boldsymbol{\beta}_{i,c}^b|i=1,\ldots,n,\,c=1,\ldots,C\right\}$.

The proposed method finds some linear projection directions that make the training samples of the same class coplanar as much as possible and make the training samples of different classes not coplanar. The coefficients are regularized by mean l_2 regularization. So the proposed method is called Regularized Coplanar Discriminant Analysis (RCDA).

2.3. RCDA: the optimization algorithm

Similar to Eqs. (1) and (3), the optimal solution to the model in Eq. (8) is hard to find. We propose a suboptimal algorithm that alternatively calculates $\mathbf{W}$, $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$ until a predefined criterion is achieved. Specifically, we first calculate $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$ by initializing $\mathbf{W}=\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Second, we attain $\mathbf{W}$ using $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$. Third, we calculate new $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$ by $\mathbf{W}$, and so on.

2.3.1. Optimize $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$

When $\mathbf{W}$ is fixed, we need to optimize $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$. Because the solution algorithm for $\boldsymbol{\beta}^w$ and that for $\boldsymbol{\beta}^b$ are similar, we only consider the former. Because each term of Eq. (4) is positive, minimizing Eq. (4) is equivalent to the following equation:

$$\min_{\boldsymbol{\beta}_i^w}f(\boldsymbol{\beta}_i^w)\doteq\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{X}_i\boldsymbol{\beta}_i^w\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_i^w-\tilde{\boldsymbol{\beta}}_i^w\right\|_2^2 \tag{9}$$

for $i=1,\ldots,n$. We calculate the lower-dimensional data by the linear projection $\mathbf{Y}_i=\mathbf{W}^T\mathbf{X}_i$ and $\mathbf{y}_i=\mathbf{W}^T\mathbf{x}_i$ first, then the solution of Eq. (9) can be found by setting the derivative of $f(\boldsymbol{\beta}_i^w)$ to be zero:

$$\frac{\partial f}{\partial\boldsymbol{\beta}_i^w}=-2\mathbf{Y}_i^T\left(\mathbf{y}_i-\mathbf{Y}_i\boldsymbol{\beta}_i^w\right)+2\lambda\left(\boldsymbol{\beta}_i^w-\tilde{\boldsymbol{\beta}}_i^w\right)=-2\mathbf{Y}_i^T\mathbf{y}_i+2\mathbf{Y}_i^T\mathbf{Y}_i\boldsymbol{\beta}_i^w+2\lambda\boldsymbol{\beta}_i^w-2\lambda\tilde{\boldsymbol{\beta}}_i^w=0,$$

namely,

$$\boldsymbol{\beta}_i^w=\left(\mathbf{Y}_i^T\mathbf{Y}_i+\lambda\mathbf{I}\right)^{-1}\left(\mathbf{Y}_i^T\mathbf{y}_i+\lambda\tilde{\boldsymbol{\beta}}_i^w\right). \tag{10}$$

2.3.2. Optimize $\mathbf{W}$

When $\boldsymbol{\beta}^w$ and $\boldsymbol{\beta}^b$ are fixed, similar to LDA [2], the solution of the objective of model (8) can be found by the following eigenvalue problem with Lagrange optimization:

$$\mathbf{S}_b\mathbf{W}=\boldsymbol{\Lambda}\mathbf{S}_w\mathbf{W}, \tag{11}$$

where $\boldsymbol{\Lambda}\mathbf{S}$ is the diagonal Lagrange multiplier matrix, and the columns of $\mathbf{W}$ are the d eigenvectors of the matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$ corresponding the top d eigenvalues.

Note that though the value of $\mathbf{W}$ at each iteration is different, the size of $\mathbf{W}$ at each iteration is equal except the initial one, i.e., $\mathbf{W}\in\mathbf{R}^{m\times d}$. The process can be repeated until some predefined iteration criterion is achieved. In the iterative process, the weight matrices are calculated in the transformed space where SRC practically works, thus it can reflect the linear relationship between each training sample and the samples from each class more accurately, and achieve better performance. The detail algorithm is summarized as Algorithm 1.

Algorithm 1. The algorithm for solving the RCDA model.

Require:
The training set $\mathbf{X}$ and corresponding class labels; Iteration times T.
Ensure:
The projection matrix $\mathbf{W}$ of RCDA;
1: Let $\mathbf{W}=\mathbf{I}$ and $t=0$.
2: Calculate $\boldsymbol{\beta}_i^w$ and $\boldsymbol{\beta}_{i,c}^b$ according to Eq. (10);
3: Calculate $\mathbf{S}_w$ and $\mathbf{S}_b$ according to Eqs. (5) and (7), respectively;
4: Calculate $\mathbf{W}$, i.e., the d leading eigenvectors of the matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$;
5: If $t>=T$, go to step 7;
6: Set $t=t+1$, go to step 2;
7: **Return** $\mathbf{W}$;

3. Analysis and comparison

3.1. Theoretical analysis

Theorem 1. *If $\lambda \to \infty$, the $\boldsymbol{\beta}$ obtained by the proposed method converges to $\tilde{\boldsymbol{\beta}}$, i.e., $\boldsymbol{\beta}_i^w \to \tilde{\boldsymbol{\beta}}_i^w$ and $\boldsymbol{\beta}_{i,c}^b \to \tilde{\boldsymbol{\beta}}_c^b$. If $\lambda \to 0$, the $\boldsymbol{\beta}$ obtained by the proposed method converges to the optimal linear representation coefficient, i.e., $\boldsymbol{\beta} \to (\mathbf{Y}^T\mathbf{Y})^{-1}(\mathbf{Y}^T\mathbf{y})$.*

Proof. According to Eq. (10), we have:

$$\lim_{\lambda\to\infty}\boldsymbol{\beta} = \lim_{\lambda\to\infty}\left(\mathbf{Y}^T\mathbf{Y} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{Y}^T\mathbf{y} + \lambda\tilde{\boldsymbol{\beta}}\right) = \lim_{\lambda\to\infty}\left(\frac{\mathbf{Y}^T\mathbf{Y}}{\lambda} + \mathbf{I}\right)^{-1}\left(\frac{\mathbf{Y}^T\mathbf{y}}{\lambda} + \tilde{\boldsymbol{\beta}}\right)$$

$$= \mathbf{I}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}.$$

Obviously, $\lim_{\lambda\to 0}\boldsymbol{\beta} = (\mathbf{Y}^T\mathbf{Y})^{-1}(\mathbf{Y}^T\mathbf{y})$. □

From Theorem 1, we can find that if $\lambda \to \infty$, the proposed method become a LDA-like method; if $\lambda \to 0$, the proposed method become an optimal linear embedding method. λ balances the bias from the optimal linear representation and that from the class mean, which provides a more stable representation.

3.2. Computational complexity

It takes $O(k^3n)$ and $O(k^3n^2)$ to construct $\mathbf{B}_w$ and $\mathbf{B}_b$, respectively, where k is the number of samples per subject and n is the number of training samples. The time complexity of the eigen decomposition step using Eq. (11) is $O(m^3)$, where m is the feature dimension. Experimental results will show that the proposed method is fast not only in training process, but also in testing process. Even at the same dimension, SRC is faster with the features extracted by RCDA than with those extracted by other DR methods.

3.3. Comparison of related works

NPDE [21], a discriminative version of NPE [17], is similar to our work. NPDE also takes into account the within-class linear

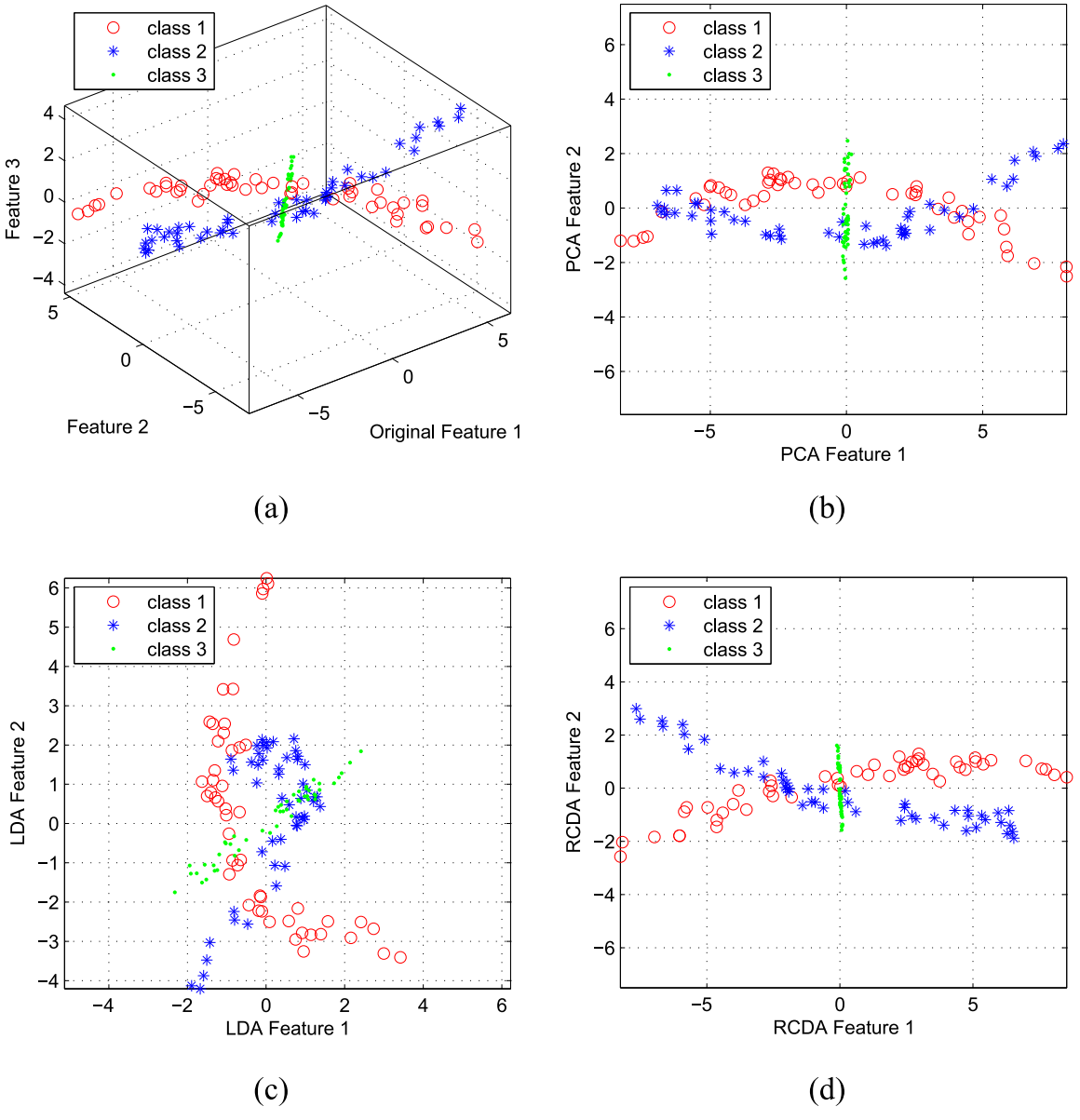

Fig. 3. A synthetic example. (a) Original 3D synthetic data. (b) PCA projection (recognition accuracy: 64%). (c) LDA projection (60%). (d) RCDA projection (90%). RCDA makes the data from the same class as co-linear as possible, rather than makes the data crowded.

representation as well as the between-class one. But NPDE learns the within-class linear representation for each sample with its k nearest neighbors from the same class, while our method uses all samples from the same class with l_2 regularization, which does not need to consider the size of neighborhood. Furthermore, the linear representation of NPDE is calculated in the original space, while that of our work is calculated in the transformed low-dimensional space where the classifier practically works.

Another similar work is GDSNPE [24], a discriminant version of SPP [18]. It retains the sparse characteristic and emphasizes the discriminative information by incorporating MMC. Its sparse characteristic is only modeled by the within-class linear representation with l_1 regularization, while our work also considers the between-class linear representation, and uses l_2 regularization. Furthermore, the separability of GDSNPE is characterized by the inter-class and intra-class geometrical structure like LDA and MFA, while our method makes the samples from the same class as much as possible in the same hyperplane.

CRP [20] is also related with our work. For each training sample, CRP first calculates the linear representation with l_2 regularization by all the other training samples to formulate the l_2 graph. Then it preserves the collaborative representation based reconstruction relationship of data. Our method is different from CRP as follows. First, CRP is an unsupervised method, while our method is supervised. Second, our method uses mean l_2 regularization, instead of l_2 regularization. Third, the linear representation of CRP is calculated in the original space, while that of our method is calculated in the transformed low-dimensional space; Fourth, we simultaneously learn the linear projection matrix and linear representation coefficients, while CRP calculates them independently.

The most similar work is SRCDP [34], which is also a DR method aiming for SRC, and characterizes the scatters in the transformed low-dimensional space. But the within-class representation of SRCDP is the sparse reconstruction residual, which introduces considerable within-class residual, while the representation of our work is linear reconstruction with l_2 regularization, which is significantly different from SRCDP.

4. Experimental results

In this section, we perform experiments on a synthetic data, two benchmark face databases and two hyperspectral image databases, to demonstrate the performance of the proposed RCDA method, and compare it with several traditional DR algorithms, such as PCA [1], LDA [2], LPP [6] and MFA [7], as well as several new algorithms based on linear embedding methods, such as NPE [17], SPP [18], CRP [20], LDSNPE [23], GDSNPE [24] and SRCDP [34]. If no specification, we use SRC as the classifier for each DR method.

4.1. Experiment on a simple instance

In this subsection, we present a numerical instance using synthetic data of three dimensions to visually demonstrate the property of RCDA, and compare it with PCA and LDA. The synthetic data is designed as follows. Firstly, we create some 3D data $\{(x_i, y_i, z_i)\}$ around an arc in xy plane, and then add a random number to each coordinate of every point according to Eq. (12), where $R=12$ is the radius, θ_i is uniformly distributed between 0 and $\pi/2$, and ϵ is uniformly distributed between -0.5 and 0.5. The 3D data are regarded as class 1:

$$\begin{cases} x_i = R\cos\theta_i + \epsilon \\ y_i = R\sin\theta_i + \epsilon \\ z_i = \epsilon \end{cases} \tag{12}$$

Then we create the data of class 2 similar to class 1, except that we set θ_i between π and $\pi + \pi/2$, and rotate it of 0.5 radian about the vector from the origin through the point (1, 1, 0). The data of class 3 are randomly distributed around a line. Last, we translate the data of each class respectively, so that the mean of each class is equal to the origin. The data of class 1 and class 2 are similar to the piece of double helix of DNA, and the data of class 3 are through the middle of them. The synthetic 3D data are shown as in Fig. 3 (a).

We project the synthetic 3D data into the 2D subspace of PCA, LDA and RCDA, respectively. Fig. 3(b)–(d) shows the effects after projection. As can be seen, the data after PCA projection is the most scattered, but there are 4 intersections across classes. LDA wants to make the points of different class far from each other and keep the points of same class close to each other, but they are still mingled by each other. RCDA makes the data from the same class as co-linear as possible, rather than makes the data crowded. So, SRC can get higher recognition accuracy after RCDA projection. We create another data of 3 classes taken as the testing data, and use SRC to classify them. The recognition accuracies are 64, 60 and 90 percent for PCA, LDA and RCDA, respectively.

4.2. Results on AR face database

In this subsection, we first compare the proposed method with other DR methods on AR face database. Then we analyze the contribution of each component of the proposed, such as l_2 regularization term, discriminant analysis and iteration.

The AR face database [35] consists of over 4000 facial images from 126 subjects (70 men and 56 women). These images suffer different facial variations, including various facial expressions, illumination variations and occlusion by sunglasses or scarf. In our experiment, we use a subset that consists 119 subjects. For each subject, 14 images with only illumination changes and expressions

Fig. 4. Images of one person in AR database.

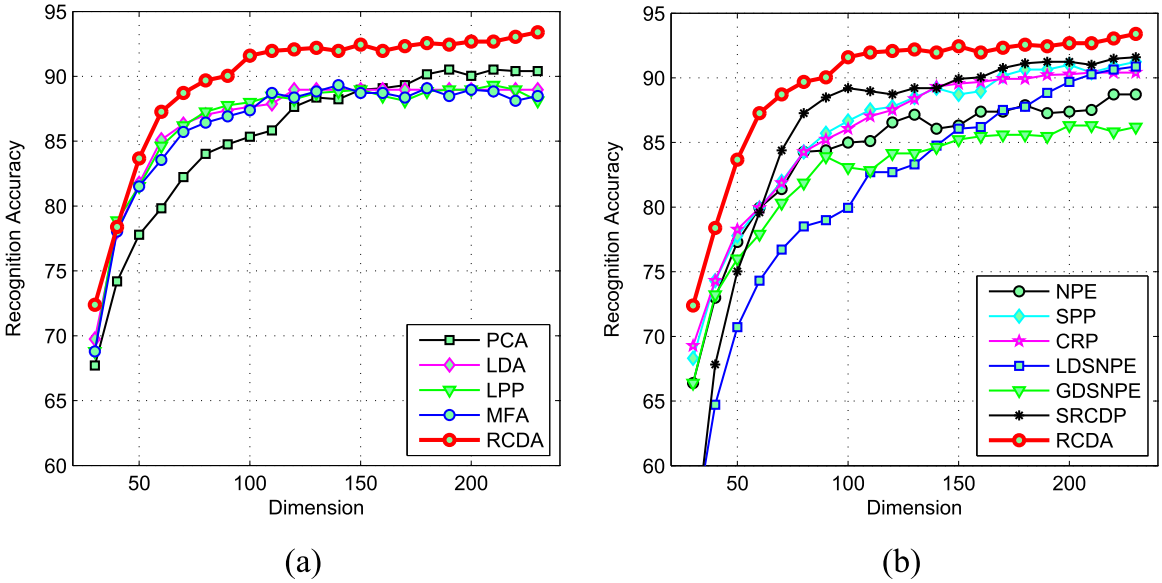

Fig. 5. Recognition accuracy versus dimension on AR database for different DR methods. (a) Comparison of the proposed method and the traditional methods. (b) Comparison of the proposed method and the linear embedding methods.

Table 1
Max recognition accuracy (%) and corresponding dimensions, training time and testing time (s) on AR database for different DR methods.

Method	Recognition accuracy	Dimensions	Training time (s)	Testing time (s)
RAW	91.84	2576	–	6512
PCA [1]	90.52	190	**14**	**31**
LDA [2]	88.96	118	15	61
LPP [6]	89.32	210	15	63
MFA [7]	89.32	140	15	60
NPE [17]	88.72	220	15	64
SPP [18]	91.24	230	137	48
CRP [20]	90.40	230	63	50
LDSNPE [23]	90.88	230	105	49
GDSNPE [24]	86.31	200	121	57
SRCDP [34]	91.60	230	128	66
RCDA	**93.40**	230	40	42

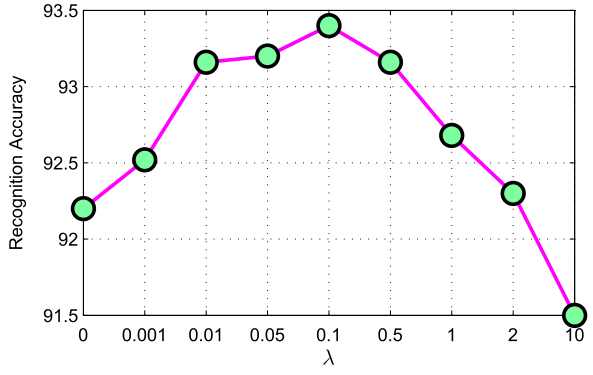

Fig. 6. Recognition accuracy versus λ for the proposed method on AR face database. λ balances the optimal linear representation and the mean.

are selected and taken into two sessions. The gray-scale images are resized to 56×46. Fig. 4 shows the face images of the first subject in the AR database, where the images in the first row are from session 1, and the images in the second row are from session 2. The statuses of the images in the first row are: neutral, smile, anger, scream, left light on, right light on, all sides light on, and the images in the second row are taken under the same conditions. Since the database has naturally been partitioned into two sessions, we use the images from session 1 for training, and images from session 2 for testing, i.e., of each subject, there are 7 samples for training and 7 samples for testing.

4.2.1. Comparison of different DR methods

We first compare our method with other DR methods. Their parameters are set as follows. The original samples are first projected to a PCA subspace for each DR method. Because the recognition accuracies of original LPP and NPE are much lower than those of other methods, we use their supervised versions, which only consider the neighbors from the same class. For LPP, the cosine kernel is used to construct the weight matrix. The sizes of within-class and between-class neighbor of LPP, MFA and NPE are set to 5. For LDSNPE, we set $\mu = 10$ proposed in [23]. For GDSNPE, we set $\gamma = 1$ proposed in [24]. For CRP, we set $\lambda = 0.01$ proposed in [20]. For SRCDP and RCDA, we retain 260 PCA dimensions for AR face database, and 360 for other face databases. The dual augmented lagrange multiplier (DALM) [36] method is used to solve all the l_1 minimization problems, where $\lambda = 0.001$ for all databases.

Fig. 5 shows the recognition accuracy versus dimension on AR database for each method. It can be seen that SPP also attains good performance for preserving the sparsity, and SRCDP gets better performance because it is a SRC-oriented method. The recognition accuracy of the proposed method is higher than those of the traditional methods such as PCA, LDA, LPP and MFA, and is better than those of the linear embedding methods such as NPE, SPP, CRP, LDSNPE, GDSNPE and SRCDP.

Table 1 shows comparison of maximum recognition accuracy, corresponding dimensions, training time and testing time for different DR methods. The testing time is the total classification time for all testing samples. Each method is evaluated at the dimension corresponding its highest recognition accuracy. As we can see, when we do not use any DR method, the testing time is 6512 s because the l_1 minimization of SRC is time-consuming especially at

Table 2
The role of l_2 regularization.

Method	CDA	$RCDA_1$	$RCDA_2$	RCDA
Recognition accuracy	92.20	92.92	92.68	**93.40**
Training time (s)	45	2032	45	45

Table 3
The role of iteration.

Iteration times	0	1	2	3
Recognition accuracy	92.33	93.40	93.28	93.52
Training time (s)	30	45	59	72

high dimension. Thus we need to reduce the dimension. The testing time of the proposed method is the least of all methods except PCA, and the performance of the proposed method is the best of all, including with the original features (RAW), which show that our method not only reduces the complexity, but also improves the performance. Though the dimension of LDA is less than that of RCDA, the testing time of LDA is much more than that of RCDA. This indicates RCDA can get better features for quickly performing SRC.

Because each method uses PCA as its first step, the training time of PCA is the minimal. The training time of other methods are only a little different from that of PCA, except the four sparse methods, namely, SPP, LDSNPE, GDSNPE, and SRCDP. The reason is that the l_1 minimization problem is time-consuming, especially at higher dimensions. But RCDA uses l_2 regularization, and calculates the representations class by class, so it costs little training time.

4.2.2. The role of the regularization term

If $\lambda \to \infty$, then $\boldsymbol{\beta} \to \tilde{\boldsymbol{\beta}}$, i.e., in this case we hope that the errors between each sample and the mean of the samples from the same class be as small as possible. But the errors maybe so big that it is hard to find a good projection to cut down them. If $\lambda \to 0$, we only consider the errors between each sample and its optimal linear representation. The errors are smaller and it is easy to find a good projection. But the optimal linear representation maybe far from the class mean, so the samples from the same class maybe scattered and it is harmful to classify new testing sample after projection. The l_2 regularization term can balance the errors between $\mathbf{x}$ and its optimal linear representation and the biases from class mean. Fig. 6 shows the recognition accuracy versus λ for the proposed method. It can be found that when $\lambda \to \infty$ or $\lambda \to 0$, the performance decreases. We can set a proper value for λ to get higher recognition accuracy. Though the optimal λ is different in different database, we simply set $\lambda = 0.1$ in Eq. (8) for all face databases.

The l_2 regularization term can be replaced by l_1, but in this case it takes much more training time, and the recognition accuracy is less than the proposed method. We can remove the l_2 regularization term, but the performance is decreased and the matrix in Eq. (10) may be singular, especially when two training samples are very similar. We can also set $\tilde{\boldsymbol{\beta}} = \mathbf{0}$ instead of $\tilde{\boldsymbol{\beta}} = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]^T$. Table 2 shows the performance comparison of four cases: without regularization (supervised version of model (1), denoted by CDA), with l_1 regularization (RCDA_1), with mean l_2 regularization (RCDA) and with l_2 regularization ($\tilde{\boldsymbol{\beta}} = \mathbf{0}$, RCDA_2). Note that they are only different in the regularization term, and the other components are the same as model (8). It can be found that the training times of CDA and RCDA are equal, but the training time of RCDA_1 is much longer. The recognition accuracy of RCDA is the highest.

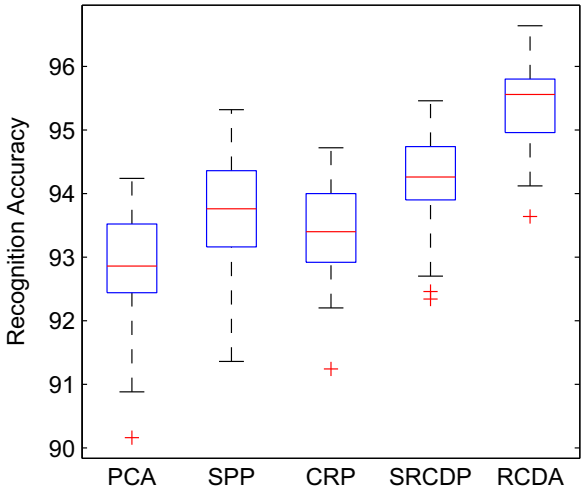

Fig. 7. Boxplot of the recognition accuracy for different DR methods on the random AR database.

Table 4
Paired t-test results for the proposed method and other method

Method	Mean recognition accuracy	Standard deviation	P-value
PCA	92.85	0.92	1.1e−18
SPP	93.70	0.91	2.7e−12
CRP	93.33	0.87	1.5e−15
SRCDP	94.18	0.80	2.1e−10
RCDA	95.37	0.72	–

4.2.3. The role of discriminant analysis

The discriminant analysis not only enables the training samples from the same class as much as possible in the same hyperplane, but also makes the training samples from different classes as much as possible not in the same hyperplane. When there is no discriminant analysis, the performance decreases. We do an experiment to compare the method that does not use class information but use coplanar analysis, i.e., for each training sample, this method uses all other training samples to represent it. The recognition accuracy without discriminant analysis is 89.33 percent, while that with discriminant analysis is 93.40 percent. It can be found that the recognition accuracy of the proposed method is much better than the one without discriminant analysis.

4.2.4. The role of iteration

In Algorithm 1, we propose an iterative method to solve the RCDA model. When there is no iteration, i.e., without step 6 in Algorithm 1, the performance decreases. The linear representation is iteratively calculated in the transformed space where the classifier practically works, thus it can achieve better performance. Table 3 shows the recognition accuracy and corresponding training time under 0, 1, 2 and 3 iterations. As we can see, the performance with iteration is better than that without iteration. But it cannot improve the performance when there are more iterations, but it takes more computational cost, so we only perform 1 iteration.

4.2.5. Comparison of different DR methods on the random AR database

To ensure that our results will not depend on any special choice of the training data, for each subject, seven images are randomly selected for training and the rest were for testing. Once we select a random training sample set, they are used for all methods. We repeat the experiment 30 times. For more concise, the proposed method only is compared with four typical methods: PCA, SPP, CRP and SRCDP.

To present more statistics and the stability of the recognition results, we use boxplot to show the classification performance of different DR methods on the random AR database, as in Fig. 7. The box has lines at the lower quartile, median, and upper values of 30 recognition accuracies. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. We can

Fig. 8. Testing images with occlusion on AR database. First row: real occlusion by sunglasses and scarf; Second row: random occlusion by gray box.

Table 5
Comparison of different methods for occlusion on AR database.

Method	Real occlusion	Random occlusion	Clean	All
Baseline	11.98	55.29	68.92	47.07
PCA	73.65	81.97	88.66	81.82
SPP	74.95	82.81	88.56	82.46
CRP	74.64	82.43	88.64	82.26
SRCDP	74.65	81.99	91.97	83.28
RCDA	**76.39**	**85.32**	**92.52**	**85.16**

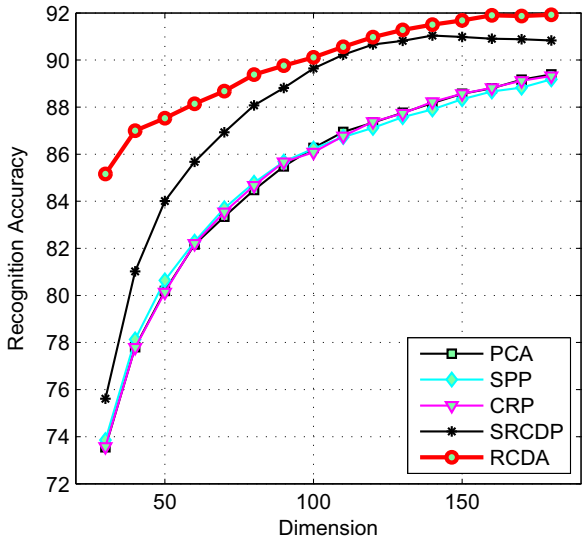

Fig. 9. Mean recognition accuracy versus dimension for different DR methods on extended Yale B database.

Table 6
Mean recognition accuracy and corresponding time for different l_1 solvers.

Method	OMP	L1Magic	L1LS	SPAM	DALM
Recognition accuracy (%)	91.96	92.07	92.10	92.11	92.10
Recognition time (s)	60	61	437	21	14

find that the median recognition accuracy of RCDA is the highest of all methods.

We also perform a paired t-test for the null hypothesis that data in the difference between RCDA and each method are a random sample from a normal distribution with mean 0 and unknown variance, against the alternative that the mean is not 0. The mean recognition accuracy, standard deviation and p-value are shown in Table 4. We can see all the p-values are much lower than 0.05, which indicates that the advantages brought by our method is significant.

4.2.6. Results with occlusion

Because we use SRC as the classifier and one major advantage of SRC is its robustness to occlusion, we perform experiment with occlusion. However, the original SRC is not robust to large contiguous occlusion [27]. To address this problem, the extended SRC was proposed [37], which applies an extended dictionary to represent the possible variation between training images and testing ones, and achieves better performance. Here we use the extended SRC as the classifier.

Besides 7 samples without occlusion for each subject of each session on AR database, there are 6 samples with occlusion by sunglasses or scarf, shown in the first row of Fig. 8. Of 119 subjects, we randomly select 19 subjects, and use all the 247 samples from these subjects of session 1 to construct the extended dictionary. Then we use 700 samples without occlusion from the other 100 subjects of session 1 for training. The remaining occluded samples of session 1 and 700 samples of session 2 for testing. To demonstrate the performance of the proposed method for random occlusion, we randomly select a box in each clean testing image, where the area of the box is 10 percent of that of the image. Then the color of the box is set to gray to construct a random occlusion, shown in the second row of Fig. 8. We repeat the experiment 10 times, and report the mean classification accuracy.

Table 5 shows the mean recognition accuracies of different methods versus different testing sets. The four testing sets are the 600 samples with real occlusion, the 700 samples with random occlusion, the 700 clean samples and all the 2000 testing samples, respectively. The method "Baseline" refers to the nearest neighbor classifier. We can find that the occlusion is very challenging because the baseline method get poor performance. However, based on the extended SRC, the DR methods significantly improve the performance. It should be pointed that they almost do not decline the performance for non-occluded testing samples. In particular, RCDA gets the best performance of all methods.

4.3. Results on extended Yale B face database

Extended Yale B database [38] contains about 2414 frontal face images of 38 individuals. We use the cropped and normalized 54×48 face images, which are taken under extreme illumination conditions [39]. For each subject, we randomly select 10 images for training, and the rest for testing. We repeat the experiment 10 times in each condition.

Fig. 9 shows the mean recognition accuracy versus dimension on the Extended Yale B database for different DR methods. As we can see, PCA, SPP and CRP get similar performance, and SRCDP is better than the two methods, while RCDA gets the highest performance.

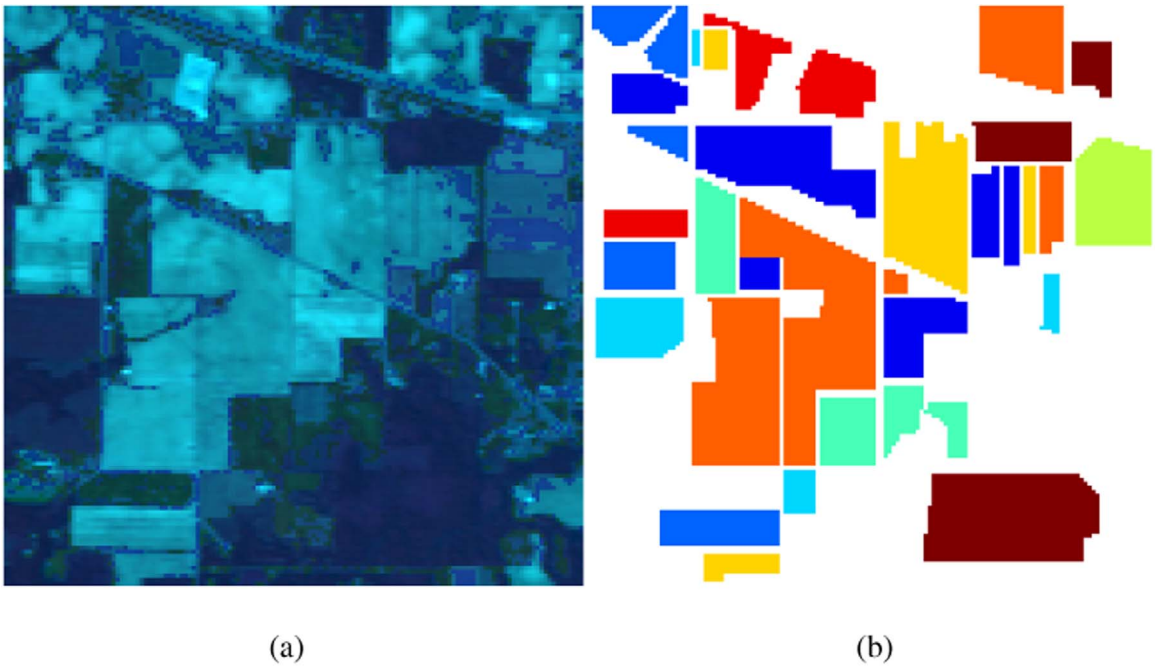

Fig. 10. (a) False color image of Indian Pines (using bands 80, 30 and 20), (b) ground truth of the labeled area with nine classes.

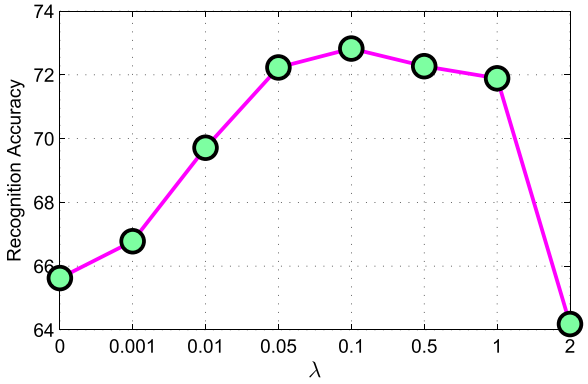

Fig. 11. Recognition accuracy versus λ for the proposed method on Indian Pines hyperspectral database with 60 training samples per class. Compared with Fig. 6, when there are more training samples per class, a proper λ can get significantly better performance than $\lambda = 0$ or $\lambda \to \infty$.

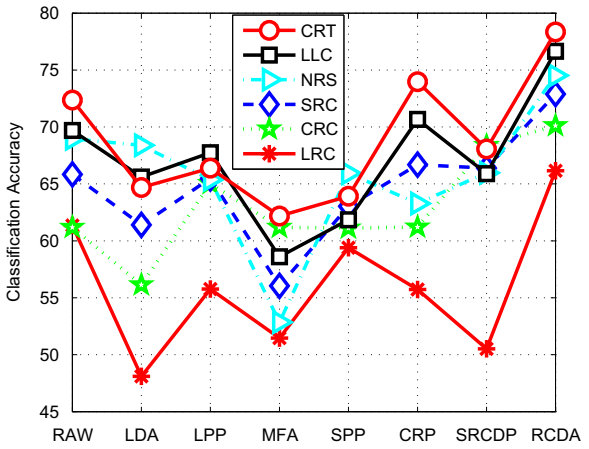

Fig. 12. Classification accuracy versus different classifier on Indian Pines hyperspectral database. For each classifier, RCDA significantly improve the classification accuracy compared with other DR methods as well as RAW.

Table 7
Classification accuracies versus different number of training samples on Indian Pines hyperspectral database.

Method	Number of training samples per class				
	20	30	40	50	60
RAW	61.47	62.94	64.98	66.31	66.73
LDA	43.16	55.03	60.06	63.17	63.48
LPP	30.37	44.99	58.34	63.33	65.89
MFA	23.16	47.03	54.02	58.11	58.38
SPP	57.97	60.29	61.49	62.26	62.94
CRP	63.50	64.98	66.72	67.53	68.08
SRCDP	54.18	58.04	60.63	64.58	64.83
RCDA	**64.02**	**67.58**	**69.95**	**71.45**	**72.82**

There are many methods to solve the l_1-minimization problem of SRC, such as Orthogonal Matching Pursuit (OMP) [40], l_1-Magic [41], l_1-regularized Least Squares (L1LS) [42], SPArse Modeling (SPAM) [43] and Dual Augmented Lagrange Multiplier (DALM) [36]. Here we evaluate the proposed algorithm with different l_1 solvers. Table 6 shows the results. We find that the recognition accuracy of OMP is a little lower than other methods, and L1Magic, L1LS, SPAM and DALM get similar recognition accuracy. DALM is the fastest, so we use DALM to solve the l_1-minimization problem of SRC.

4.4. Results on hyperspectral image databases

Hyperspectral images are widely used in remote sensing, automatic target recognition and surveillance task. Along with the increasing demand of remote sensing data, the sensor technology has been developed to improve the spatial and spectral resolution [44]. Because there are limited training samples available, DR is a necessary preprocessing for hyperspectral image classification [45–47]. In this subsection, we perform experiments on two hyperspectral image databases, to show that our method is also good at hyperspectral image classification.

Table 8
Classification accuracy versus different number of training samples on Salinas hyperspectral dataset.

Method	Number of training samples per class				
	10	20	30	40	50
RAW	94.99	95.98	96.24	96.81	96.99
LDA	88.54	88.33	94.93	95.28	96.82
LPP	87.77	85.79	91.70	95.09	95.88
MFA	87.91	86.07	89.95	90.68	92.42
SPP	95.10	96.09	97.10	97.30	97.39
CRP	93.57	94.89	96.67	96.66	96.98
SRCDP	94.71	96.42	96.94	97.05	97.24
RCDA	**96.31**	**96.62**	**97.52**	**97.75**	**97.98**

4.4.1. Results on Indian Pines hyperspectral database

The first hyperspectral data employed was acquired using National Aeronautics and Space Administrations (NASA) Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor and was collected over northwest Indiana's Indian Pine test site. The image represents a rural scenario with 145×145 pixels and 220 bands in the 0.4–2.5 μm region of the visible and infrared spectrum with a spatial resolution of 20 m. In this experiment, a total of 202 bands is used after removal of water absorption bands. From the 16 different land-cover classes in the image, 7 classes are discarded due to their insufficient number of training samples. A three-band false color image and its ground-truth are shown in Fig. 10.

On hyperspectral databases, there are many samples per class, while the number of class is less than that on face databases. To avoid overfitting the data, we do not learn the between-class linear representation, i.e., we set $\mathbf{B}_c = \mathbf{0}$ in Eq. (8). Furthermore, when there are more training samples per class, the parameter λ of the proposed method is more important. Fig. 11 shows the recognition accuracy versus λ for the proposed method on Indian Pines hyperspectral database with 60 training samples per class. When there are more training samples per class, an proper λ can get significantly better performance than $\lambda = 0$ or $\lambda \to \infty$. λ balances the optimal linear representation and the mean.

The classification accuracies versus different number of training samples on the database are reported in Table 7, where "RAW" denotes the original data without DR. Since the Indian Pines database has a relatively large number of densely sampled spectral bands, there are substantial redundancies in it. We use different DR methods to reduce the dimensionality. As Table 7, we can see that the proposed method significantly outperforms other baseline algorithms such as LDA, LPP, MFA, SPP and SRCDP. Note that the classification accuracy of the proposed method is even significantly higher than that of RAW without DR. But the classification time of RAW is much more than that of the proposed method. Compared with the results on face databases, the improvement is more significant on hyperspectral image databases. Because there are more training samples per class, the regularization term of RCDA is more useful.

Besides SRC, there are some other linear representation-based classifiers which are attracted a great deal of attentions in recent years, such as collaborative representation classifier (CRC) [48], linear representation classifier (LRC) [49], nearest regularized subspace (NRS) [50], locality-constrained linear classifier (LLC) [51], and collaborative representation with Tikhonov regularization (CRT) [52]. In fact, RCDA can also improve the performance compared with other DR methods based on the linear representation-based classifiers. Fig. 12 shows the recognition accuracy versus different classifier on Indian Pines hyperspectral database with 60 training samples per class. We can find that for each classifier, RCDA significantly improve the classification accuracy compared with other DR methods as well as RAW.

4.4.2. Results on Salinas hyperspectral database

The second experimental hyperspectral data was collected by the 224-band AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution. The area covered comprises 512 lines by 217 samples. As with Indian Pines scene, we discarded the 20 water absorption bands. This image was available only as at-sensor radiance data. It includes vegetables, bare soils, and vineyard fields. Salinas ground-truth contains 16 classes.

The classification accuracy versus different number of training samples on Salinas database are reported in Table 8. We can find that the proposed method outperforms other baseline algorithms.

5. Conclusion

Based on the experiments, we can draw a number of conclusions as follows. The recognition accuracy of RCDA is not only higher than those of other DR methods based on SRC, but also better than the original high-dimensional features. The testing time of RCDA is faster than those of other DR methods and especially much faster than that of the original high-dimensional features, and the training time of RCDA is a little more than PCA. Each step of RCDA, including l_2 regularization, coplanar discriminant analysis, and iterative computing, makes a contribution to improve performance. When there are more training samples per class, the regularization term of RCDA is more useful. The improvements by RCDA are more significant on hyperspectral image databases than those on face databases, because there are more training samples per class on hyperspectral image databases. RCDA is a linear DR method, while sometimes the best features must be transformed by non-linear method, which is our future work.

References

[1] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.
[2] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, New Jersey, 2012.
[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[4] H.F. Li, T. Jiang, K.S. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Netw. 17 (1) (2006) 157–165.
[5] D.Q. Dai, P.C. Yuen, Face recognition by regularized discriminant analysis, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 37 (4) (2007) 1080–1085.
[6] X.F. He, S.C. Yan, Y.X. Hu, P. Niyogi, H.J. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[7] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[8] J.W. Lu, Y.P. Tan, Regularized locality preserving projections and its extensions for face recognition, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 40 (3) (2010) 958–963.
[9] W.H. Deng, J.N. Hu, J.W. Lu, J. Guo, Transform-invariant PCA: a unified approach to fully automatic face alignment, representation, and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2014) 1275–1284.
[10] M. Lu, J.Z. Huang, X.N. Qian, Sparse exponential family principal component analysis, Pattern Recognit. 60 (12) (2016) 681–691.
[11] J. Oh, N. Kwak, Generalized mean for robust principal component analysis, Pattern Recognit. 54 (6) (2016) 116–127.
[12] K.T. Abou-Moustafa, F.D.L. Torre, F.P. Ferrie, Pareto models for discriminative multiclass linear dimensionality reduction, Pattern Recognit. 48 (5) (2015) 1863–1877.
[13] Y.A. Ghassabeh, F. Rudzicz, H.A. Moghaddam, Fast incremental LDA feature extraction, Pattern Recognit. 48 (6) (2015) 1999–2012.
[14] C.X. Ren, D.Q. Dai, X.F. He, H. Yan, Sample weighting: an inherent approach for outlier suppressing discriminant analysis, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3070–3083.
[15] S. Wang, J.F. Lu, X.J. Gu, H.S. Du, J.Y. Yang, Semi-supervised linear discriminant analysis for dimension reduction and classification, Pattern Recognit. 57 (9) (2016) 179–189.
[16] Z. Ji, Y.W. Pang, Y. Yuan, J. Pan, Relevance and irrelevance graph based marginal Fisher analysis for image search reranking, Signal Process. 121 (4) (2016) 139–152.

[17] X.F. He, D. Cai, S.C. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: IEEE International Conference on Computer Vision (ICCV), vol. 2, 2005, pp. 1208–1213.
[18] L.S. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.
[19] B. Cheng, J.C. Yang, S.C. Yan, Y. Fu, T.S. Huang, Learning with l1-graph for image analysis, IEEE Trans. Image Process. 19 (4) (2010) 858–866.
[20] W.K. Yang, Z.Y. Wang, C.Y. Sun, A collaborative representation based projections method for feature extraction, Pattern Recognit. 48 (1) (2015) 20–27.
[21] P.Y. Han, A.T.B. Jin, F.S. Abas, Neighbourhood preserving discriminant embedding in face recognition, J. Vis. Commun. Image Represent. 20 (8) (2009) 532–542.
[22] C.T. Ding, L. Zhang, Double adjacency graphs-based discriminant neighborhood embedding, Pattern Recognit. 48 (5) (2015) 1734–1742.
[23] G.F. Lu, Z. Jin, J. Zou, Face recognition using discriminant sparsity neighborhood preserving embedding, Knowl.-Based Syst. 31 (2012) 119–127.
[24] J. Gui, Z.A. Sun, W. Jia, R.X. Hu, Y.K. Lei, S.W. Ji, Discriminant sparse neighborhood preserving embedding for face recognition, Pattern Recognit. 45 (8) (2012) 2884–2893.
[25] Q.X. Gao, Y.F. Huang, H.L. Zhang, X. Hong, K. Li, Y. Wang, Discriminative sparsity preserving projections for image recognition, Pattern Recognit. 48 (8) (2015) 2543–2553.
[26] P.Y. Zhang, X.G. You, W.H. Ou, C.L.P. Chen, Y.M. Cheung, Sparse discriminative multi-manifold embedding for one-sample face identification, Pattern Recognit. 52 (4) (2016) 249–259.
[27] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[28] M.S. Cui, S. Prasad, Class-dependent sparse representation classifier for robust hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 53 (5) (2015) 2683–2695.
[29] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 625–632.
[30] R. He, W.S. Zheng, B.G. Hu, Maximum correntropy criterion for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1561–1576.
[31] C.X. Ren, D.Q. Dai, H. Yan, Robust classification using L2,1-norm based regression model, Pattern Recognit. 45 (7) (2012) 876–888.
[32] Z.R. Lai, D.Q. Dai, C.X. Ren, K.K. Huang, Discriminative and compact coding for robust face recognition, IEEE Trans. Cybern. 45 (9) (2015) 1900–1912.
[33] K.K. Huang, D.Q. Dai, C.X. Ren, Z.R. Lai, Learning kernel extended dictionary for face recognition, IEEE Trans. Neural Netw. Learn. Syst. (99) (2016) 1–13, in press.
[34] J. Yang, D.L. Chu, L. Zhang, Y. Xu, Sparse representation classifier steered discriminative projection with applications to face recognition, IEEE Trans. Neural Netw. Learn. Syst. 24 (7) (2013) 1023–1035.
[35] A.M. Martinez, R. Benavente, The AR Face Database, Technical Report, CVC #24 June 1998.
[36] A.Y. Yang, Z.H. Zhou, A.G. Balasubramanian, S.S. Sastry, Y. Ma, Fast l_1-minimization algorithms for robust face recognition, IEEE Trans. Image Process. 22 (8) (2013) 3234–3246.
[37] W.H. Deng, J.N. Hu, J. Guo, Extended SRC: undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1864–1870.
[38] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[39] K.C. Lee, J. Ho, D.J. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.
[40] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations, J. Constr. Approx. 13 (1997) 57–98.
[41] E.J. Cands, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2) (2004) 489–509.
[42] S.J. Kim, K. Koh, M. Lustig, S. Boyd, An interior-point method for large-scale l1-regularized least squares, IEEE J. Sel. Top. Signal Process. 1 (4) (2010) 606–617.
[43] J. Mairal, F. Bach, J. Ponce, Sparse modeling for image and vision processing, Found. Trends Comput. Graph. Vis. 8 (2) (2014) 85–283.
[44] K.K. Huang, D.Q. Dai, A new on-board image codec based on binary tree with adaptive scanning order in scan-based mode, IEEE Trans. Geosci. Remote Sens. 50 (10) (2012) 3737–3750.
[45] W. Li, S. Prasad, J.E. Fowler, L.M. Bruce, Locality-preserving dimensionality reduction and classification for hyperspectral image analysis, IEEE Trans. Geosci. Remote Sens. 50 (4) (2012) 1185–1198.
[46] M.S. Cui, S. Prasad, Angular discriminant analysis for hyperspectral image classification, IEEE J. Sel. Top. Signal Process. 9 (6) (2015) 1003–1015.
[47] X.S. Wang, Y. Gao, Y.H. Cheng, A non-negative sparse semi-supervised dimensionality reduction algorithm for hyperspectral data, Neurocomputing 188 (5) (2016) 275–283.
[48] L. Zhang, M. Yang, X.C. Feng, Sparse representation or collaborative representation: which helps face recognition? in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 471–478.
[49] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2106–2112.
[50] W. Li, E.W. Tramel, S. Prasad, J.E. Fowler, Nearest regularized subspace for hyperspectral classification, IEEE Trans. Geosci. Remote Sens. 52 (1) (2014) 477–489.
[51] J.J. Wang, J.C. Yang, K. Yu, F.J. Lv, T. Huang, Y.H. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3360–3367.
[52] W. Li, Q. Du, M.M. Xiong, Kernel collaborative representation with Tikhonov regularization for hyperspectral image classification, IEEE Geosci. Remote Sens. Lett. 12 (1) (2015) 48–52.

Ke-Kun Huang received the B.Sc., M.Sc. and Ph.D. degrees in Applied Mathematics from Sun Yat-sen University, Guangzhou, China, in 2002, 2005 and 2016, respectively. He is currently an Associate Professor with the Department of Mathematics, Jiaying University, Meizhou, China. His research interests include image processing and face recognition.

Dao-Qing Dai received the B.Sc. degree in Mathematics from Hunan Normal University, Changsha, China, in 1983, the M.Sc. degree in mathematics from Sun Yat-sen University, Guangzhou, China, in 1986, and the Ph.D. degree in Mathematics from Wuhan University, Wuhan, China, in 1990. He is currently a Professor with School of Mathematics, Sun Yat-sen University. His current research interests include image processing, statistical pattern recognition, and bioinformatics.

Chuan-Xian Ren received the Ph.D. degree from School of Mathematics, Sun Yat-Sen University, Guangzhou, China, in 2010. He is currently an Associate Professor with School of Mathematics, Sun Yat-sen University. His current research interests include image processing and face recognition.