This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

Robust Sparse Linear Discriminant Analysis

Jie Wen, Xiaozhao Fang*, Jinrong Cui, Lunke Fei, Ke Yan, Yan Chen, Yong Xu

***Abstract*—Linear discriminant analysis (LDA) is a very popular supervised feature extraction method and has been extended to different variants. However, classical LDA has the following problems: 1) The obtained discriminant projection does not have good interpretability for features. 2) LDA is sensitive to noise. 3) LDA is sensitive to the selection of number of projection directions. In this paper, a novel feature extraction method called robust sparse linear discriminant analysis (RSLDA) is proposed to solve the above problems. Specifically, RSLDA adaptively selects the most discriminative features for discriminant analysis by introducing the $l_{2,1}$ norm. An orthogonal matrix and a sparse matrix are also simultaneously introduced to guarantee that the extracted features can hold the main energy of the original data and enhance the robustness to noise, and thus RSLDA has the potential to perform better than other discriminant methods. Extensive experiments on six databases demonstrate that the proposed method achieves the competitive performance compared with other state-of-the-art feature extraction methods. Moreover, the proposed method is robust to the noisy data.**

***Index Terms*—Linear discriminant analysis, feature selection, feature extraction, data reconstruction.**

I. Introduction

FEATURE selection and extraction play important roles in pattern classification and have received much attention in recent years [1]. Especially for gene expression and image analysis, the original data usually have very high dimensions and contain large redundant features or noises. In this case, how to select and extract the most discriminative features for different classification tasks is a challenge work [2, 3]. In fields of pattern classification and machine learning, feature selection and extraction have proven to be effective tools in reducing complexity, improving efficiency and enhancing the classification performance [4-6]. Feature selection aims to select a few of the most important or relevant features from the original features to efficiently represent original data for a given task [7]. It does not change the feature value but allows data to be better classified and more efficiently stored. Feature extraction tries to learn a projection matrix that can transform the original high dimensional data into a low dimensional subspace [8, 9]. Both of the feature selection and feature extraction can be viewed as the subspace learning methods to some extent because they aim to find a low dimensional representation in a new space to represent the original high dimensional data.

In the past few decades, various feature extraction methods have been proposed. In this branch, the most well-known method is principal component analysis (PCA) [10], which seeks to learn a projection that can preserve the main energy of data. In view of its good property in data reconstruction and energy preservation, it is widely used as a data preprocessing technique for data analysis [11-13]. Locality preserving projection (LPP) [14], sparsity preserving projections (SPP) [15], and neighborhood preserving embedding (NPE) [16] are also the popular feature extraction methods which learn their projections from different geometric structures of original data. Compared with PCA, these methods aim to preserve geometric structures of original data in the subspaces. Although the above methods have their advantages in feature extraction, they would not be suitable for classification problems since features extracted by these methods do not contain discriminability [17].

Linear discriminant analysis (LDA) is one of the most favored methods to extract discriminative features for pattern classification [18, 19]. LDA uses the label information to learn a discriminant projection that can greatly enlarge the between-class distance and reduce the within-class distance so as to improve the classification accuracy. Various extensions of LDA have also been developed to enhance the performance and efficiency. For example, orthogonal LDA (OLDA) [20], uncorrelated LDA (ULDA) [21], and two-dimensional linear discriminant analysis (2DLDA) [22] are proposed to address the small sample size problem of naive LDA. Compared with OLDA and ULDA which transform an image into a vector for projection learning, 2DLDA can be directly applied on the image matrix, which can make use of the structure information of image for feature extraction. To solve the problem that LDA fails to deal with data of non-Gaussian distribution, marginal Fisher analysis (MFA) [23], discriminative locality alignment (DLA) [24], and manifold partition discriminant analysis (MPDA) [25] are proposed. These three methods seek to learn a more general discriminant projection by utilizing both the neighbor information and label information. However, the LDA based methods mentioned above all use the l_2 norm to calculate the scatter matrixes, which is likely to magnify errors and leads these methods to be sensitive to outliers.

This work was supported in part by the National Natural Science Foundation of China under Grant nos.61300032 and 61772141, in part by the National Natural Science Foundation of China Youth Fund under Grant nos. 61702110 and 61703169, in part by the Guangdong Province high-level personnel of special support program under Grant no. 2016TX03X164, and in part by the Shenzhen Fundamental Research fund under Grant no. JCYJ20160331185006518. *(Jie Wen and Xiaozhao Fang contributed equally to this work.) (Corresponding author: Xiaozhao Fang (e-mail: xzhfang168@126.com).)*

Jie Wen, Ke Yan, and Yong Xu are with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China; and also with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China. (Email: wenjie@hrbeu.edu.cn; yanke401@163.com; yongxu@ymail.com).

Xiaozhao Fang and Lunke Fei are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China, (e-mail: xzhfang168@126.com; flksxm@126.com).

Jinrong Cui is with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510000, China (E-mail:tweety1028@163.com).

Yan Chen is with the Shenzhen Sunwei Intelligent Corporation, Shenzhen 518000, China (Email:jadechenyan@gmail.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

To address this issue, Li et al. proposed to use a rotational invariant l_1 norm to measure the two scatter matrices for discriminant projection learning [26]. In the method of Li et al., a weighting parameter is used to balance the importance of the two scatter matrices. However, it is difficult to find the optimal weighting value for different tasks, which limits its application. Wang et al. also proposed an improved LDA method which uses the l_1 norm rather than the l_2 norm in the Fisher criterion function [27]. However, their methods are inefficiency since each projection vector needs to be iteratively solved. Recent years, many deep learning based feature extraction methods have also been proposed and aroused much attention [28, 29]. For example, Dorfer et al. extended the classical LDA into the deep neural network and proposed the deep linear discriminant analysis (DeepLDA) for object classification [28]. DeepLDA seeks to learn a model that can concentrate as much discriminative power as possible on the $C-1$ directions, where C is the class number. DeepLDA achieves very good performance on the large-scale image datasets. However, it needs large amount of training samples to train the feature extraction net. In addition, it is too difficult to interpret the model with the complex network structures. For example, we do not know that which types of features or which sub-areas of the image play the dominant role to the classification. Compared with the deep learning based methods, conventional methods are more suitable to the tasks with small scale databases. In this work, we mainly focus on the conventional feature extraction methods.

Although some conventional methods show their superiority in the real-world applications, most of them have a common problem, *i.e.*, they do not have the ability to perform feature selection. In real-world applications, there are many redundant features in original data. In other words, some features are harmful to the purpose of classification [30]. To overcome this issue, the sparse constraint is exploited in subspace learning methods to select important features and remove redundant information for feature extraction. For example, sparse discriminant analysis (SDA) [31], sparse linear discriminant analysis (SLDA) [17], and sparse uncorrelated linear discriminant analysis (SULDA) [32] are proposed to learn a sparse discriminant subspace for feature extraction. The $l_{2,1}$ norm based sparse technique is also helpful for efficient feature selection owing to the row-sparsity property. For example, Li et al. proposed a novel robust structured subspace learning (RSSL) method to achieve an appropriate latent subspace for data representation where $l_{2,1}$ norm is adopted in the formulations of loss function and regularization term to make algorithm robust to the outliers and noise and to select discriminative features for label prediction [33]. Tao et al. proposed a method to select the discriminative features by imposing a row-sparsity constraint on the transformation matrix of LDA via the $l_{2,1}$ norm regularization [34].

Using the sparse constraint to select the most important features for feature extraction is available. However, the above methods still have many shortcomings. First, methods with the l_1 norm constraint cannot uncover the intuitive difference across features. In other words, we still do not know which categories of features are the most important for the given task since the selected features are different in different projection vectors. Second, most of these methods are not robust to noise. Third, existing LDA based methods are somewhat sensitive to the selection of the number of dimensions since the discriminability of each projection direction is fixed. This indicates that these methods cannot adaptively preserve more discriminant information according to the selected number of projection directions.

To solve the above issues and obtain the desired subspace, we propose a robust feature extraction method based on LDA in this paper. The proposed method uses the $l_{2,1}$ norm to constrain the projection matrix, which is able to simultaneously perform feature extraction and feature selection. In order to improve the robustness to noise, we introduce a sparse error term to fit noise during learning. Most importantly, to hold the main energy of the original data in the discriminant subspace, the proposed method introduces an orthogonal matrix to connect the original features and transformed features so that the transformed data can preserve the main discriminant information. Different from traditional feature extraction methods that only preserve properties of reconstruction or discrimination, the proposed approach can be viewed as the method that integrates PCA and LDA into a joint learning framework. In this way, it not only can extract the most discriminative features, but also holds the main energy of the original data with respect to the number of projection directions. These factors enable the transformed data to be more discriminative, and thus guarantee the proposed method to obtain a better performance than other methods. In brief, the proposed method has the following properties.

1) RSLDA can simultaneously select and extract the most discriminative features for classification.
2) The transformed data in the discriminant subspace hold the main energy and thus have the minimum loss of the discriminant information. In this way, the proposed method is insensitive to the selection of number of dimensions. The subsequent experiments also verify that the proposed method is more flexible in selecting the dimension and can obtain a very outstanding performance with low dimensions.
3) Compared with other LDA based methods, our method is more robust to noise.

The remainder of the paper is organized as follows. In Section II, we briefly introduce some related works. In Section III, we present the RSLDA method and its solution in detail. In Section IV, the rationales of the proposed method are provided. In Section V, the classification performance of the proposed method is evaluated with some state-of-the-art supervised learning methods. In Section VI, the conclusion is offered.

II. Related works

In this section, we briefly introduce a feature extraction method, *i.e.*, LDA, and a feature selection method via the $l_{2,1}$ norm constraint, which are much related to our work.

For convenience, we first introduce some notations used through the paper. We define $X=[x_1,x_2,\ldots,x_n]\in R^{m\times n}$ as a training set with n samples, each sample x_i is represented

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

as a column vector $x_i = [x_{1,i}, x_{2,i}, \ldots, x_{m,i}]^T \in R^m$. For the image sample, we pre-transform it into the column vector by stacking the image columns. For a matrix $A \in R^{m\times d}$, its l_1 norm, $l_{2,1}$ norm, and l_F norm are calculated as $\|A\|_1 = \sum_{j=1}^{d}\sum_{i=1}^{m}|a_{ij}|$, $\|A\|_{2,1} = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{d}a_{ij}^2}$, and $\|A\|_F = \sqrt{\sum_{j=1}^{d}\sum_{i=1}^{m}a_{ij}^2}$, respectively. For a vector $b = [b_1, b_2, \ldots, b_m]$, the l_2 norm is defined as $\|b\|_2 = \sqrt{\sum_{i=1}^{m}b_i^2}$.

A. Linear discriminant analysis

Suppose there are c pattern classes, n_i denotes the number of samples of the ith class, $n = \sum_{i=1}^{c} n_i$ is the total number of all samples, column vector $x_j^i \in R^m$ denotes the jth sample of the ith class. LDA tries to find a projection vector which is able to enlarge the distance of samples from different classes and reduce the distance of samples from the same class. LDA uses the following Fisher criterion to obtain this projection vector [35]

$$a = \arg\max_a \frac{a^T S_b a}{a^T S_w a} \tag{1}$$

where S_b and S_w are the between-class and within-class scatter matrices, respectively. S_b and S_w are calculated as follows.

$$S_b = \frac{1}{n}\sum\nolimits_{i=1}^{c} n_i(u_i - u)(u_i - u)^T \tag{2}$$

$$S_w = \frac{1}{n}\sum\nolimits_{i=1}^{c}\sum\nolimits_{j=1}^{n_i}(x_j^i - u_i)(x_j^i - u_i)^T \tag{3}$$

where $u_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^i$ denotes the mean feature of samples of the ith class, $u = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} x_j^i$ is the mean feature of all samples. Generally, problem (1) is equivalent to the following optimization problem [20, 30]

$$a = \arg\min_{a^T a=1} a^T(S_w - \lambda S_b)a \tag{4}$$

where λ is a small positive constant.

By solving Eq.(4), we can observe that the optimal projection vector a is the eigenvector corresponding to the minimum eigenvalue of $S_w a = \lambda S_b a$. Generally, a single projection vector is not enough to distinguish multiple classes. In real-world applications, we usually select a set of projection vectors which satisfy the optimal Fisher criterion $A = \arg\min_{A^T A=I} Tr(A^T(S_w - \lambda S_b)A)$ for multi-class classification. Projection matrix A is selected as a set of eigenvectors corresponding to the first k smallest eigenvalues of $S_w A = \lambda S_b A$. Let $A = [a_1, a_2, ..., a_k] \in R^{m\times k}$ be the set of the selected k eigenvectors, we can obtain discriminative feature vector $y_j^i \in R^k$ of each sample by $y_j^i = A^T x_j^i$.

B. Feature selection method via the $l_{2,1}$ norm

Xiang et al. presented a feature selection method via the $l_{2,1}$ norm constraint [36]. Suppose $X = [x_1, x_2, \ldots, x_n] \in R^{m\times n}$ is a set of n samples. The ith sample x_i is represented as a column vector $x_i = [x_{1,i}, x_{2,i}, \ldots, x_{m,i}]^T \in R^m$. Suppose Q is the learned projection matrix and $q_{i,\cdot}$ is the ith row vector of Q, $i \in [1, m]$. Projected sample y_i is

$$y_i = Q^T x_i \tag{5}$$

$$\begin{aligned} y_{1,i} &= q_{1,1}x_{1,i} + q_{2,1}x_{2,i} + \cdots + q_{m,1}x_{m,i} \\ y_{2,i} &= q_{1,2}x_{1,i} + q_{2,2}x_{2,i} + \cdots + q_{m,2}x_{m,i} \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_{m,i} &= q_{1,m}x_{1,i} + q_{2,m}x_{2,i} + \cdots + q_{m,m}x_{m,i} \end{aligned} \tag{6}$$

It is obvious that if some rows of projection matrix Q are equal to zero, then some features corresponding to these rows can be regarded as unimportant or redundant features and thus can be removed. As introduced in the previous section, feature selection is to select the most discriminative features from the original data for classification. The discriminative power of the ith category of features can be represented by $\|q_{i,\cdot}\|_2$, where $\|\cdot\|_2$ is the l_2 norm. Suppose $\bar{Q} = \left[\|q_{1,\cdot}\|_2, \|q_{2,\cdot}\|_2, \ldots, \|q_{m,\cdot}\|_2\right]^T$, then the task of selecting d most discriminative features from m features is equivalent to the following constraint problem

$$\left\|\bar{Q}\right\|_0 = d \tag{7}$$

Solving the optimization problem with the l_0 norm constraint is a NP hard problem. Fortunately, the solution obtained by using the l_1 norm constraint also contains sufficient sparsity and has the approximate solution to the l_0 norm constraint [37]. Thus, we can use the l_1 norm constraint to select these most important features. $\bar{Q}$ with the l_1 norm constraint is equivalent to matrix Q with the $l_{2,1}$ norm constraint

$$\|Q\|_{2,1} = \left\|\bar{Q}\right\|_1 = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{m}q_{ij}^2} \tag{8}$$

The above analysis demonstrates the good feature selection property of the $l_{2,1}$ norm constraint. Inspired by this motivation, Xiang et al. integrated this constraint into the linear regression framework to adaptively select those important features for classification [36].

III. The Proposed Method

LDA aims at learning a projection that decreases the distance of samples from the same class and increases the separability of samples from different classes. However, it has some obvious drawbacks. First, the learned projection matrix does not have good interpretability for features due to each new feature of a sample is linearly combined by all features and most of projection coefficients are nonzero. This also indicates that LDA does not have the ability to select the most useful features from the redundant data. Second, LDA selects k eigenvectors corresponding to the first k smallest eigenvalues as the projection for feature extraction while the number of k is data-dependent. This leads the classification accuracy of LDA to be sensitive to the selection of reduced dimensions. Third, many LDA based methods are sensitive to the presence of noise. In this paper, we propose a novel and robust sparse discriminative feature extraction method to solve the above problems.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

A. Model of the proposed method

In real-world applications, the acquired data are usually high-dimensional and contain large amounts of redundant features. Thus it is necessary to select those important features from the original complicated data for discriminant analysis so that the negative influence of such redundant features can be effectively reduced. As presented in the previous section, imposing the sparse norm constraint, such as the l_1 and $l_{2,1}$ norms, on the projection can make the model perform feature selection. While different from the l_1 norm, the $l_{2,1}$ norm has a good row-sparsity property, which can make the learned projection have better interpretability for features. Inspired by this motivation, we propose to learn a more robust discriminant subspace by utilizing this constraint as follows

$$\min_{Q} Tr\left(Q^T\left(S_w - uS_b\right)Q\right) + \lambda_1\|Q\|_{2,1} \tag{9}$$

where $Q \in R^{m\times d}$ $(d < m)$ is the discriminative projection matrix. S_b and S_w are the between-class and within-class scatter matrices, respectively. λ_1 is a trade-off parameter, u is a small positive constant used to balance the importance of S_b and S_w.

By using the $l_{2,1}$ norm constraint, model (9) has the ability to adaptively assign large projected weights to the category of important features. However, similar to the conventional LDA, model (9) is still sensitive to the selection of reduced number d. If d is very small, the learned projection cannot preserve the discriminative information as much as possible, which leads to a low classification accuracy. In this paper, we provide an efficient way to address this issue. Motivated by the energy preserving property of PCA, we introduce a variant of the PCA constraint into the projection learning model as follows [38]

$$\begin{aligned} &\min_{P,Q} Tr\left(Q^T\left(S_w - uS_b\right)Q\right) + \lambda_1\|Q\|_{2,1} \\ &s.t.\ \ X = PQ^TX,\ \ P^TP = I \end{aligned} \tag{10}$$

where the constraints of $X = PQ^TX$ and $P^TP = I$ can be viewed as a variant of PCA to some extent, which ensures the original data can be recovered well [39]. $P \in R^{m\times d}$ is an orthogonal reconstruction matrix. By taking into account the reconstruction relationship between the transformed samples and original samples, the transformed data can preserve the main energy of the original data as much as possible with respect to the reduced dimensions. In this way, RSLDA not only learns a discriminative subspace, but also has minimum information loss to some extent through the joint optimization framework, and thus has the potential to perform better.

In real-world applications, the data or image may be corrupted by noise. In this paper, we mainly focus on the case of random noise. We use a sparse term to compensate the noise so that the negative effect can be reduced to some extent. Thus, the objective function of RSLDA can be rewritten as follows

$$\begin{aligned} &\min_{P,Q,E} Tr\left(Q^T\left(S_w - uS_b\right)Q\right) + \lambda_1\|Q\|_{2,1} + \lambda_2\|E\|_1 \\ &s.t.\ \ X = PQ^TX + E,\ \ P^TP = I \end{aligned} \tag{11}$$

where λ_2 is also a trade-off parameter and determines the importance of the corresponding term. E denotes errors and is used to model the random noise. $\|\cdot\|_1$is the l_1 norm.

B. Solution to the proposed learning model

In this section, we present an iterative method to solve the optimization problem of RSLDA by using the alternating direction method of multipliers (ADMM) [40]. We first convert problem (11) into the following Lagrangian function

$$\begin{aligned} L(P,Q,E,Y) =& Tr\left(Q^T\left(S_w - uS_b\right)Q\right) + \lambda_1\|Q\|_{2,1} \\ &+ \lambda_2\|E\|_1 + \left\langle Y, X - PQ^TX - E\right\rangle \\ &+ \frac{\beta}{2}\left\|X - PQ^TX - E\right\|_F^2 \\ =& Tr\left(Q^T\left(S_w - uS_b\right)Q\right) + \lambda_1\|Q\|_{2,1} \\ &+ \lambda_2\|E\|_1 - \frac{1}{2\beta}\|Y\|_F^2 \\ &+ \frac{\beta}{2}\left\|X - PQ^TX - E + \frac{Y}{\beta}\right\|_F^2 \end{aligned} \tag{12}$$

where β is a penalty parameter, Y is the Lagrangian multiplier. Then P,Q,E can be alternately solved by minimizing the Lagrangian function L with other variables fixed. The solution scheme is as follows.

Step 1. Update Q: fix P, E and update Q by minimizing the following problem

$$\begin{aligned} L(Q) =& Tr\left(Q^T\left(S_w - uS_b\right)Q\right) + \lambda_1\|Q\|_{2,1} \\ &+ \frac{\beta}{2}\left\|X - PQ^TX - E + \frac{Y}{\beta}\right\|_F^2 \end{aligned} \tag{13}$$

Define $X - E + \frac{Y}{\beta} = M$, Q can be calculated by the derivative of $L(Q)$ with respect to Q

$$\frac{\partial L(Q)}{\partial Q} = 2(S_w - uS_b)Q + \lambda_1 DQ + \beta(XX^TQ - XM^TP) \tag{14}$$

where D is defined as $D = \begin{bmatrix} \frac{1}{\|q_1\|_2} & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\|q_m\|_2} \end{bmatrix}$. q_i is the ith row of Q and $Q = \begin{bmatrix} q_1 \\ \vdots \\ q_m \end{bmatrix}$. Let $\partial L(Q)/\partial Q = 0$, then we obtain

$$Q = \left(2\left(S_w - uS_b\right) + \lambda_1 D + \beta XX^T\right)^{-1}\left(\beta XM^TP\right) \tag{15}$$

Step 2. Update P: fix Q and update E, P by minimizing the following problem

$$\min_{P^TP=I}\left\|X - PQ^TX - E + \frac{Y}{\beta}\right\|_F^2 \tag{16}$$

Let $X - E + \frac{Y}{\beta} = M$. Problem (16) is converted to

$$\begin{aligned} &\min_{P^TP=I}\left\|M - PQ^TX\right\|_F^2 \\ &= \min_{P^TP=I} Tr\left(M^TM - 2M^TPQ^TX\right) \\ &= \max_{P^TP=I} Tr\left(M^TPQ^TX\right) \\ &= \max_{P^TP=I} Tr\left(P^TMX^TQ\right) \end{aligned} \tag{17}$$

Problem (17) is an Orthogonal Procrustes problem and can be simply solved. Suppose $SVD\left(MX^TQ\right) = USV^T$, then

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

P is obtained as $P = UV^T$ [39], where SVD denotes the operation of singular value decomposition.

Step 3. Update E: we fix P, Q and update E by solving the following problem

$$\min_E \lambda_2 \|E\|_1 + \frac{\beta}{2} \left\| X - PQ^T X + \frac{Y}{\beta} - E \right\|_F^2 \tag{18}$$

If we define $e = \frac{\lambda_2}{\beta}$ and $E_0 = X - PQ^T X + \frac{Y}{\beta}$, according to the shrinkage operator [41], problem (18) has the following closed form solution

$$E = shrink(E_0, e) \tag{19}$$

where $shrink$ denotes the shrinkage operator.

Step 4. Update Y, β: Y and β are respectively updated by using the following formulas

$$Y = Y + \beta(X - PQ^T X - E) \tag{20}$$

$$\beta = \min(\rho\beta, \beta_{\max}) \tag{21}$$

where ρ and $\beta_{\max}$ are the constant.

The proposed method is summarized in Algorithm 1.

Algorithm 1 RSLDA (solving (11))

Input: data matrix X, parameters λ_1, λ_2.

Initialization: $Q = 0$; $E = 0$; $Y = 0$; $\beta = 0.1$;$\rho = 1.01$; $P = \arg\min_P Tr\left(P^T (S_w - uS_b) P\right)$ $s.t.$ $P^T P = I$; $\beta_{\max} = 10^5$;$u = 10^{-4}$;

while not converged **do**

1. Update Q by using (15).
2. Update P by solving (17).
3. Update E by using (19).
4. Update Y, β by (20) and (21), respectively.

end while

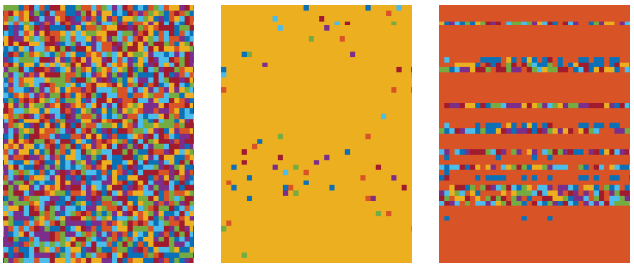
Output: P, Q, E

IV. Analysis of RSLDA

In the previous section, the proposed method and its solution are introduced. In this section, we will present the rationales of the proposed method in detail and then analyze the computational complexity and convergence of the proposed method.

A. Rationale of RSLDA

(1) Compared with conventional methods: In real-world applications, the acquired original data or images usually have large dimensions, which contain a lot of redundant features that are harmful to classification. Moreover, different features have different discriminative powers in different tasks. So the learned projection should better have this feature selection ability. For this goal, a $l_{2,1}$ norm constraint is integrated into the projection learning model of LDA. As analyzed in the previous section, the $l_{2,1}$ norm has the row-sparsity property which has the potential to adaptively assign large projected weights to the important features during learning. Fig.1 shows the first 50 rows of the projection matrices obtained by LDA [18], SLDA [17], and the proposed method. SLDA is a feature extraction method with the l_1 norm sparse constraint. We can observe that the projection matrix obtained by the proposed method has the good row-sparsity property while those of LDA and SLDA do not have. From Fig.1(c), it is obvious to see that which categories of features are the most discriminative features for classification. This proves that the projection matrix obtained by the proposed method has better interpretability than those of LDA and SLDA. Fig.1 also proves that RSLDA has the ability to select the most discriminative features from the original data for feature extraction. This has the potential to improve the discriminability of the new subspace.



(a) LDA (b) SLDA (c) RLSDA

Fig. 1. Comparison of projection matrices obtained by LDA, SLDA and the proposed method on the Extended Yale B face database in which 15 samples of each class are randomly selected as training samples. Note: we only show the first 50 rows of their projection matrices for comparison. For vividly comparison, we choose colormap of 'Lines' in the above figures.

As introduced in Section II, LDA selects the eigenvectors corresponding to the first k smallest eigenvalues as the projection for discriminant analysis. In fact, the number value of k of the selected dimensions is a key variable to determine the performance of LDA. The main reason is that these k eigenvectors cannot preserve enough discriminative information for classification, especially $k \ll c - 1$ (c is the class number). To overcome this problem, the proposed method integrates a variant of PCA constraint into the projection learning model of LDA. Similar to PCA, this reconstruction constraint makes the projection hold the main energy of the original data and thus can ensure the minimum loss of information. By this novel integration of PCA and LDA, RSLDA is able to catch as much discriminative information as possible in each dimension for classification. This also indicates that the proposed method is more flexible to select the number of dimensions than LDA. Different from PCA and LDA which only hold the main energy or discriminability of data, RSLDA shares both advantages of them. Thus the proposed method can learn the optimal subspace, and the extracted features can be viewed as the best representation of the original data in the discriminative subspace. This encourages the method to obtain a better performance.

In real-world applications, images may be corrupted by different factors such as illuminations and occlusions due to the uncontrollability of image acquisition, which may degrade the performance of classification. To address this challenge issue, a sparse error term is introduced to the objective function. Fig.2 shows the image recovery performance of the proposed method. It can be seen that the proposed method can greatly reduce the influence of random corruption of noise. Different from other methods, such as LDA and SLDA, which learn the projection from the original noisy data, RSLDA has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

the potential to extract features from the latent clear data. So RSLDA is more robust to noise.

In conclusion, RSLDA has many good advantages compared with other methods. By the effective integration of these good factors, RSLDA can learn a more robust projection for feature extraction so as to obtain a better classification performance.

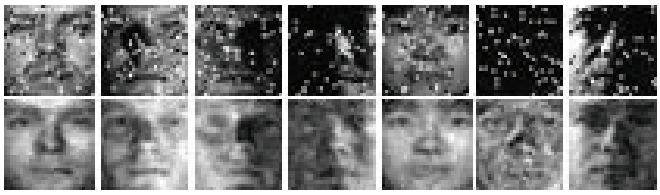


Fig. 2. Results of image recovery of the proposed method under noisy condition. Images in the first row are corrupted by random noise, images in the second row are recovery results of the proposed method. (Note: each corrupted image in the first row is transformed into a column vector x to calculate its recovery vector by formula $\hat{x} = PQ^T x$. Then we reshape vector $\hat{x}$ into the image matrix for displaying.)

(2) Compared with deep learning methods: Deep learning has received much attention in recent years owing to its good performance in extracting discriminative features from samples adaptively [28, 29, 42, 43]. In this branch, DeepLDA is one of the most representative works, which extends the classical LDA into the deep neural network [28]. Similar to the proposed method, DeepLDA also tries to learn a discriminative projection that produces high inter-class and low intra-class variances. The biggest difference between DeepLDA and the proposed method is that DeepLDA performs the LDA in a latent feature space while the proposed method is directly conducted on the original space. The significant advantage of DeepLDA is that it can learn the latent representations with higher discriminability by the guiding of LDA, which is conducive to produce more discriminative features for classification. However, DeepLDA usually needs large amount of samples with label information to train a general network model, with the results that it cannot deal with tasks with limited training samples [43]. This is mainly because that limited samples will make the model overfitting. In addition, DeepLDA is not suitable to deal with the classification tasks with too many categories, especially for the case that some categories have only a few samples. This is mainly because that DeepLDA requires a large number of samples (at least 10-20 samples per category) in each sub-training stage to guarantee the scatter matrix to be non-singular, which greatly improve the requirement for computing equipment. Compared with DeepLDA, the proposed method is very simple and can effectively deal with classification tasks with limited training samples per class. Moreover, the proposed method has good interpretability and is robust to noise. Inspired by the motivation of DeepLDA, it is possible to integrate the proposed method into the framework of deep convolutional network so that the deep model may preserve more discriminative information. In addition, the proposed method not only can be directly applied on the image to extract discriminative features, but also has the potential to further improve the discriminability of other types of features, such as local binary patterns (LBP), scale-invariant feature transform (SIFT), histogram of oriented gradient (HOG), and deep features, etc [44, 45]. Therefore, the proposed method is valuable and can be flexibly applied in many fields.

B. Computational complexity and convergence analysis

For RSLDA presented in Algorithm 1, the most computational steps are step 1 and step 2. In step 1, the major computational cost is the matrix inverse operation. For a $m \times m$ matrix, the computational complexity of inverse operation is $O(m^3)$. Thus, the computational complexity to calculate Q is $O(m^2 n + m^3 + \max(m^2, mn)d)$. In step 2, the major computational cost is the singular value decomposition (SVD) of matrix. For a $m \times n$ matrix, the computational complexity of the SVD operation is $O(n^3)$. Thus, the computational complexity of step 2 is $O(\max(m^2, mn)d + d^3)$. So the whole computational complexity of the proposed method is $O(\tau(m^2 n + m^3 + 2\max(m^2, mn)d + d^3))$, where τ is the iteration number. For simplicity, we suppose that $m \gg n$, thus the computational complexity of the proposed method is $O(\tau(m^2 n + m^3 + 2m^2 d + d^3))$.

Problem (12) is a typical non-convex optimization problem, thus it is not realistic to achieve the global optimal solution. By using the ADMM-style method, a local optimal solution can be achieved. We experimentally show the convergence characteristic of the proposed method. From Fig.3, we can see that the objective value decreases obviously. This means that the proposed method converges fast. Especially in Fig.3(a) and (d), the proposed method can converge within about 10 iterations. Fig.3 also shows the classification accuracy versus the iteration step on different datasets. From Fig.3, especially in figures (c) and (d), we can see that the classification accuracy is stable and fast reaches a stable point within 20 iteration steps. As shown in Fig.3 (b), for the AR database in which some faces are taken with occlusions (sun glasses and scarf), the proposed method can also fast converge to a stable accuracy within 40 iteration steps. Therefore, the proposed method is effective and the overall computational cost is not expensive.

V. Experiments and Analysis

In this section, we use six benchmark databases, including the COIL20 image database[1] [46], AR face database[2] [47], Extended Yale B face database[3] [48], CMU face database[4] [49], Caltech-256 database [50], and PubFig83 web face database [51] to evaluate the effectiveness of the proposed method. KNN and some supervised learning methods, including SVM [52], LDA [18], SLDA [17], OLDA [20], ULDA [21], MFA [23], DLA [24], SULDA [32], and MPDA [25] are chose to compare with the proposed method. For SVM, we utilize the cross validation strategy to select the best parameters, *i.e.*, penalty term and kernel bandwidth of radial basis function, and then report the best accuracy in the following experiments. In each database, we randomly select samples from each

[1]Available at http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[2]Available at http://www2.ece.ohio-state.edu/∼aleix/ARdatabase.html.

[3]Available at http://vision.ucsd.edu/ iskwak/ExtYaleDatabase/ExtYaleB.html.

[4]Available at http://www.ri.cmu.edu/projects/project_418.html.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

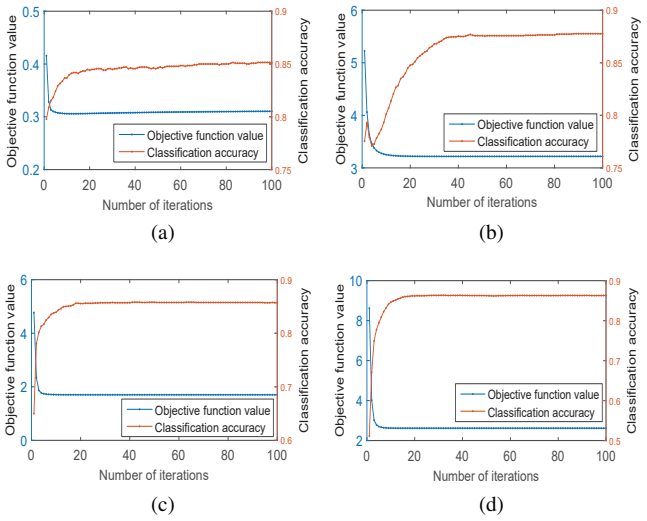


Fig. 3. Convergence and classification accuracy versus the iterations of the proposed method on (a) the COIL20 database, (b) the AR database, (c) the Extended Yale B database, and (d) the CMU PIE database, in which four, four, 10, and 10 samples of each class are randomly selected as the training samples of the corresponding database, respectively.

class as the training set and perform every methods 30 times. Then we report the mean classification accuracy (%) for comparing. For the above supervised learning methods, the nearest neighbor (NN) classifier is used to obtain the final classification accuracies of different methods. All experiments are implemented on Matlab 2015a and Windows 7, with Inter Core i7-4970 CPU and 16GB RAM.

A. Parameters selection

The proposed method contains two parameters, *i.e.*, λ_1, λ_2 to be set in advance. Thus in this section we will discuss the sensitivity of the above two parameters. We first select the two parameters from a candidate set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$, and then perform the proposed method with different combinations of these parameters. Fig.4 shows the classification accuracy versus the combination of parameters on the COIL20 database, AR database, Extended Yale B database, and the CMU PIE database. As can be seen from Fig.4, the proposed method can obtain a satisfactory performance as long the values of two parameters are small and within a feasible range. For the COIL20 database, the best performance can be achieved when the two parameters are close to 0.001. For the AR database, the best two parameters are also close to 0.001. For the Extended Yale B and the CMU PIE databases, the best performance is achieved when λ_1, λ_2 are close to 0.0001. Fig. 4 also indicates that the two parameters are significant to learn the discriminative projection and directly determine the classification performance. As far as we know, there is still an open problem to adaptively select their optimal parameters for different classification tasks. Thus, in the experiment we adopt a simple strategy to find the approximate optimal values for these parameters. We first fix parameter λ_1 in advance to find a candidate interval where the optimal parameter λ_2 may exist. Then, we further fix parameter λ_2 in the candidate interval to find the candidate interval of λ_1. By using this search strategy, we can finally obtain the optimal parameters λ_1 and λ_2 in the 2D candidate space with a fixed step length.

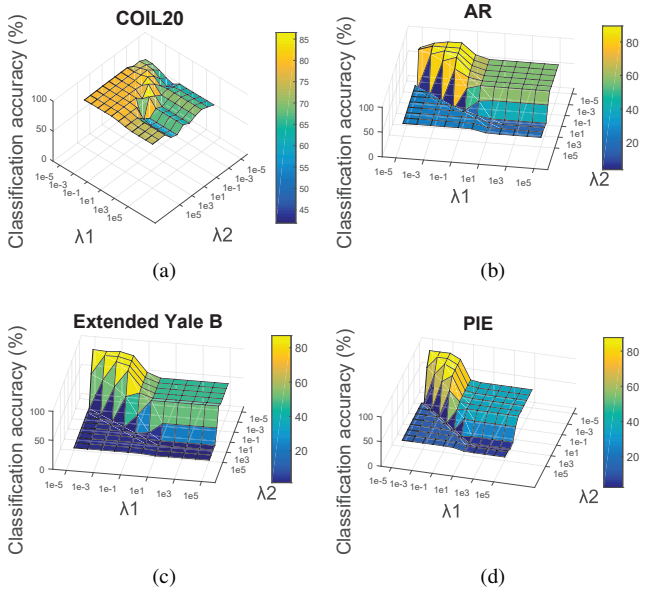


Fig. 4. The classification accuracy of the proposed method versus parameters λ_1, λ_2 on (a) the COIL20 database, (b) the AR database, (c) the Extended Yale B database, and (d) the CMU PIE database, in which four, four, 10, and 10 samples of each class are randomly selected from the corresponding database as training samples, respectively.

B. Experiments on the COIL20 image database

The COIL20 image database contains 1440 images. There are 20 objects and each object provides 72 images which are taken at pose intervals of 5 angle degree. Fig.5 shows some images of the COIL20 image database. Each original image is normalized to the size of 128×128. In the experiments, each image is resized to a 32 by 32 matrix in advance, and then PCA is used to further reduce the dimensions (preserve 95% energy) of the images to improve the computational efficiency. For each class, 4, 6, 8, and 12 samples are randomly selected as training samples and the remaining samples are treated as test samples.

Fig. 5. Some typical images of the COIL20 image database.

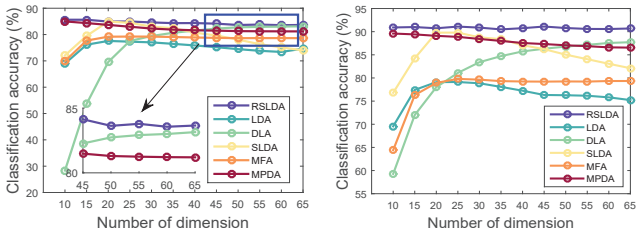
Table I shows the experimental results of different methods on the COIL20 image database. From Table I, we can find that the proposed method achieves the highest classification accuracies where are much higher than those of the LDA, ULDA, SULDA, and MFA. Specially, KNN also achieves a good performance on this database and performs better than LDA with limited samples per class. This is possibly because that the original data already contain sufficiently clear and discriminative features which are suitable for classification. While utilizing LDA may make the extracted features loss

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE COIL20 IMAGE DATABASE.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
4	81.91	83.95	77.57	84.96	77.10	53.94	83.15	79.23	74.12	85.00	**85.63**
6	86.58	90.16	79.13	89.80	84.61	70.85	87.73	79.80	75.08	89.60	**91.11**
8	89.31	93.05	87.56	92.47	89.16	80.84	90.00	80.63	67.50	91.39	**93.34**
12	92.72	95.45	93.33	95.71	93.28	88.13	93.52	92.26	82.08	94.40	**95.92**

some discriminative information, which leads to the decreasing of accuracies. Compared with LDA, the proposed method takes into account the data reconstruction. This constraint enables the extracted features of RSLDA to preserve as much discriminative power as possible, which promotes the proposed method to achieve a better performance. Fig.6 shows the classification results versus the number of dimensions of different discriminant analysis methods on the COIL20 database, in which 4 and 6 samples are randomly selected from each class as the training set and the remaining samples are treated as the test set. From Fig.6, we can see that compared with other supervised learning methods, RSLDA obtains the best classification results in each dimension. Moreover, RSLDA achieves the outstanding performance with few number of dimensions, about 10 for the database. This indicates that the proposed method is able to preserve the main discriminant information for classification.

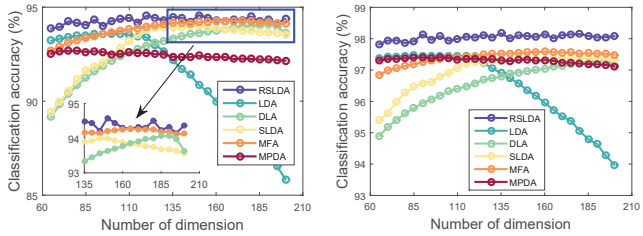


(a) (b)

Fig. 6. Classification accuracy (%) versus the number of dimensions of different supervised learning methods on the COIL20 database, in which (a) 4 samples, (b) 6 samples are randomly selected from each class as training set, respectively. (Note: local area marked by the 'blue rectangle' is magnified and the corresponding magnified image is pointed out by the 'black arrow'.)

the proposed method obtains the best performance. With the increase of dimension number, especially when the dimension number is larger than the class number, the classification accuracy of LDA decreases dramatically, while the proposed method can still obtain consistent good performance. This indicates that the dimension selection of the proposed method is more flexible than other methods. This also demonstrates that the reconstruction constraint is very useful to improve the discriminability of features in the subspace.

Fig. 7. Some typical images of the AR face database.



(a) (b)

Fig. 8. Classification accuracy (%) versus the number of dimensions of different supervised learning methods on the AR database, in which (a) 6 samples, (b) 12 samples are randomly selected from each class as the training set, respectively. (Note: local area marked by the 'blue rectangle' is magnified and the corresponding magnified image is pointed out by the 'black arrow'.)

C. Experiments on the AR face database

The AR face database contains more than 4000 color face images of 126 subjects with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). In the experiments, we use a subset which contains 3120 images from 120 subjects to test the above methods. Fig.7 shows some images of the AR face database. In the experiments, each image is converted to a 40 by 50 gray image in advance. And then PCA is used to preserve 95% energy to improve the computational efficiency. For each subject, 4, 6, 8, and 12 samples are randomly selected as training samples and the remaining samples are treated as test samples.

Table II shows the experimental results of different methods on the AR face database. Fig.8 shows the classification results versus the number of dimensions of different supervised learning methods on the AR face database, in which 6 and 12 samples are randomly chosen from each class as the training set and the remaining samples are treated as the test set. From Table II and Fig.8, it is obvious to see that

D. Experiments on the Extended Yale B face database

The Extended Yale B face database contains 38 subjects and each subject provides 64 face images with different illumination conditions. Some faces of the Extended Yale B face database are shown in Fig.9. Each image is cropped and converted to a 32 by 32 gray image in advance. To improve the computational efficiency, PCA is used to preserve 98% energy in the experiment. For each subject, we randomly select 10, 15, 20, and 25 samples as training samples and the remaining samples are treated as test samples.

Table III shows the experimental results of different methods on the Extended Yale B database. Fig.10 shows the classification results versus the number of dimensions of different methods on the Extended Yale B face database, in which 15 and 25 samples are randomly chosen from each class as the training set and the remaining samples are treated as the test set. From Table III, we can find that the proposed method can obtain competitive performance compared with other supervised learning methods, especially are much better

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE AR FACE DATABASE.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
4	53.88	69.51	87.33	89.83	90.11	86.16	88.23	88.71	27.34	87.94	**90.40**
6	62.92	82.75	93.60	94.00	94.35	92.56	94.07	94.27	83.75	92.68	**94.57**
8	69.15	89.22	95.56	95.83	96.08	95.00	95.79	96.14	91.02	94.47	**96.24**
12	77.43	95.51	97.47	97.38	97.37	97.02	97.31	97.59	95.95	97.41	**98.17**

TABLE III
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE EXTENDED YALE B FACE DATABASE.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
10	43.29	72.34	82.01	83.77	86.18	82.49	**87.78**	87.25	84.61	83.67	87.46
15	50.77	81.40	87.57	88.97	90.38	88.20	91.17	91.02	88.72	86.82	**91.43**
20	56.08	86.38	90.24	91.74	92.56	91.02	93.09	92.72	91.66	90.38	**93.26**
25	59.86	89.24	91.94	93.31	93.78	92.63	94.26	93.56	92.14	91.79	**94.53**

TABLE IV
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE CMU PIE FACE DATABASE.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
10	34.28	63.90	81.09	85.74	87.38	85.74	85.47	85.51	82.97	80.02	**87.60**
15	42.83	75.97	87.53	90.41	91.32	88.15	89.91	90.77	88.11	84.75	**91.64**
20	49.71	82.55	90.57	92.60	93.20	90.79	91.98	93.02	90.72	89.64	**93.59**
25	55.40	86.30	92.41	93.92	94.33	92.41	93.20	94.33	92.50	91.35	**94.57**

than KNN, SVM, and ULDA, etc. From Fig.10, we can find that DLA, LDA, and SLDA are sensitive to the number of dimensions, while the proposed method and MFA are more robust to the selection of dimensions.

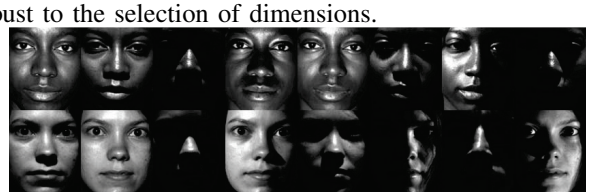


Fig. 9. Some typical images of the Extended Yale B face database.

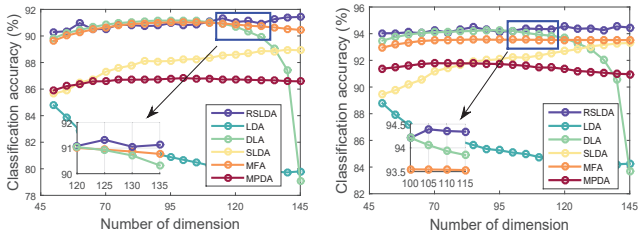


Fig. 10. Classification accuracy (%) versus the number of dimensions of different supervised learning methods on the Extended Yale B database, in which (a) 15 samples, (b) 25 samples are randomly selected from each class as the training set, respectively. (Note: local area marked by the 'blue rectangle' is magnified and the corresponding magnified image is pointed out by the 'black arrow'.)

E. Experiments on the CMU PIE face database

The CMU PIE face database contains 41368 face images from 68 subjects with different poses, illumination conditions, and facial appearances. In the experiments, we use a subset which has 11554 images from 68 subjects to test the above methods. Each image is converted to a 32 by 32 gray image. To improve the computational efficiency, PCA is used to preserve 98% energy in the experiment. Fig.11 shows some typical face images of the CMU PIE face database. We randomly select 10, 15, 20, and 25 samples per subject as training samples and treat the remaining samples as test samples. Experimental results of different methods are shown in Table IV. Fig.12 shows the influence of the number of dimensions of different discriminant analysis methods on the CMU PIE face database, in which 10 and 25 samples are randomly chosen from each class as the training set and the remaining samples are treated as the test set.

From Table IV and Fig.12, it can be seen that the proposed method achieves much better performance than KNN, SVM, LDA, and ULDA. Compared with SLDA, OLDA, and MFA, the classification accuracies of RSLDA are very stable in all dimensions. This proves that RSLDA is able to capture as much discriminant information as possible for classification associated with the number of subspace dimensions. Thus, the proposed method is superior to these methods to some extent.

Fig. 11. Some typical images of the CMU PIE face database.

F. Experiments on the Caltech-256 database

The Caltech-256 database is a challenging classification set which contains a total of 30,608 images with complicated background. There are 257 categories that consist of 256 object categories and a background category. Each object group has 80-827 images. Some typical images are shown in Fig.13. Following the experimental settings in [45], we also compare different methods on the deep learning features of the Caltech-256 database. In this work, we conduct the experiment on the deep convolutional activation features (DeCAF-6) of the Caltech-256 database which are available at https://sites.google.com/site/crossdataset/home/files [53]. To

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

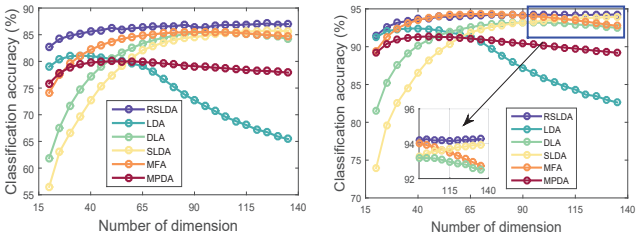


Fig. 12. Classification accuracy (%) versus the number of dimensions of different supervised learning methods on the CMU PIE database, in which (a) 10 samples, (b) 25 samples are randomly selected from each class as the training set, respectively. (Note: local area marked by the 'blue rectangle' is magnified and the corresponding magnified image is pointed out by the 'black arrow'.)

improve the computational efficiency, PCA is applied on the deep features to preserve 98% energy. Then we randomly select 15, 30, 45, and 60 samples from each class as training samples and treat the remaining samples as test samples, respectively.

Fig. 13. Some typical images of the Caltech-256 object database.

Experimental results of different methods on the DeCAF-6 of Caltech-256 database are shown in Table V. From this table, one can see that KNN obtains competitive performance compared with some supervised learning methods, such as LDA, ULDA, and SULDA, etc. This indicates that features extracted by the deep learning method already have higher discriminative power. In addition, we can also find that the manifold based feature extraction methods, *i.e.*, DLA, MFA, and MPDA, perform worse than the other methods. This is mainly because the deep convolutional neural network ignores to preserve the local geometric structures in the stage of deep feature extraction. In other words, the deep features, *i.e.*, DeCAF-6, do not capture the intrinsic nearest neighbor relationships of samples. Table V also shows that the proposed method obtains the best performance on the DeCAF-6 of the Caltech-256 database. This proves that the proposed method has the potential to further improve the discriminative power of the deep learning features.

Figure 14 shows the classification accuracies of different methods with respect to the number of feature dimensions on the DeCAF-6 of the Caltech-256 database, in which 15 and 45 samples are randomly selected from each class as the training set, respectively. From Fig.14, it is obvious to see that the proposed method obtains consistent better performance than the other methods in all dimensions. With the increase of the feature dimension, the classification accuracies of LDA and MPDA dramatically decrease, while the proposed method is very stable. This proves that the proposed method is insensitive to the selection of feature dimensions to some extent.

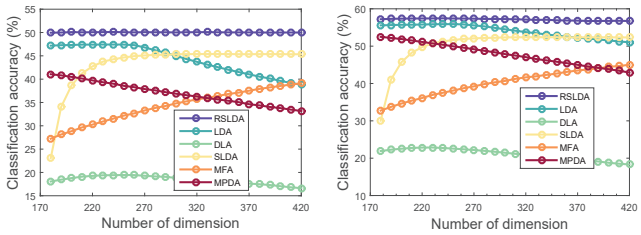


Fig. 14. Classification accuracy (%) versus the number of dimensions of different supervised learning methods on the deep features of the Caltech-256 database, in which (a) 15 samples, (b) 45 samples are randomly selected from each class as the training set, respectively.

G. Experiments on the PubFig83 database

PubFig83 database [51] is a very challenge large-scale face database, in which all images are collected from the web with different illuminations, poses, expressions, and backgrounds, etc. In this subsection, we adopt a subset of PubFig83[5] which totally contains 13002 color images (8720 training samples and 4282 test samples) provided by 83 persons to evaluate the effectiveness of the proposed method [54]. Each person provides 46-231 face images. The size of each image is 250×250. Typical images in the PubFig83 database are shown in Fig.15. Following the experimental settings presented in [54], we also compare different methods on the subset of features (the first 1536 dimensions of descriptors) which are composed of HOG, LBP, and Gabor wavelet features of images for a fair comparison. Besides the compared methods mentioned above, three typical deep learning methods, *i.e.*, DeepLDA [28], Alexnet [43], and VGG [55], are also evaluated on this database. For DeepLDA, we use the network model similar to that of the mnist database for training. For Alexnet, we also directly use the 8720 images for training without any pre-trained models. It is a pity that VGG does not converge during training without the pre-trained model. Thus we use the pre-trained model of VGG to conduct experiment and report the classification accuracy for comparison. Experimental results are shown in Table VI.

From Table VI, it is obvious that VGG achieves the highest classification accuracy. However, the other two deep learning methods, *i.e.*, DeepLDA and Alexnet, perform worse than the conventional supervised learning methods. This proves that it is difficult to obtain a perfect deep network model with a few training samples. This is mainly because that the limited training samples will lead to overfitting or non-convergence. Although the deep convolutional networks have the potential to extract the most discriminative features for classification, they need large amounts of samples or pre-trained models. Compared with the conventional supervised methods, the proposed method can obtain the highest accuracy on this database, about 6% higher than LDA. This proves that the proposed method is able to improve the discriminability of features extracted by some unsupervised approaches.

[5]The subset of PubFig83 database and their corresponding features are available at http://www.briancbecker.com/blog/research/pubfig83-lfw-dataset/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

TABLE V
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE DEEP FEATURES OF THE CALTECH-256 DATABASE.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
15	45.55	48.12	47.43	45.40	48.07	44.74	37.33	44.88	45.11	41.06	**50.11**
30	50.02	52.59	52.33	51.37	53.45	51.37	42.02	49.70	51.45	49.08	**55.02**
45	52.38	53.66	55.90	53.15	55.48	54.15	45.06	50.22	54.24	52.50	**57.63**
60	53.51	55.86	56.64	54.68	57.64	55.75	47.37	52.12	55.72	54.45	**58.26**

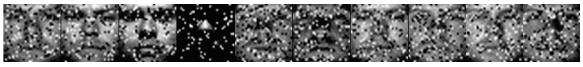
Fig. 15. Some typical images of the PubFig83 database.

TABLE VI
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON THE PUBFIG83 DATABASE.

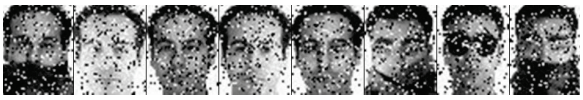
Alg.	Acc.	Alg.	Acc.
KNN	63.35	MFA	78.47
SVM	82.60	SULDA	81.26
LDA	77.95	MPDA	67.89
SLDA	79.44	DeepLDA	44.35
OLDA	82.88	Alexnet	64.00
ULDA	81.75	VGG	**96.25**
DLA	76.09	RSLDA	84.78

H. Experiments on the random pixel corruptions

In this section, we also test the robustness of the proposed method on the Extended Yale B face database and AR face database with random pixel corruptions. In this experiment, we randomly add salt and pepper noise to each image. The corrupt degree of noise is 15%. To improve the efficiency, we also use PCA to reduce the dimension and preserve 95% energy of data in advance. Some typical images which are degraded by the random noise are shown in Fig.16.



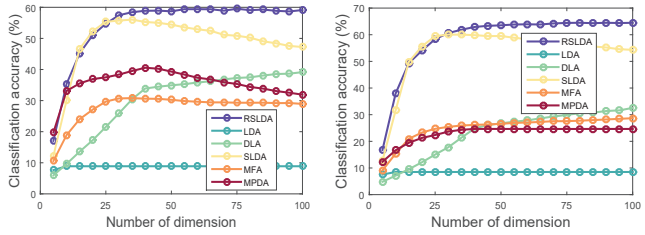
(a) Noisy images of Extended Yale B face database.



(b) Noisy images of AR face database.

Fig. 16. Typical images degraded by random noise.

Table VII shows the experimental results of different methods on the Extended Yale B face database in which the images are corrupted by the 'salt and pepper' noise. We can see that the proposed method can obtain the competitive performance in comparison with the state-of-the-art methods. From Fig.17, we can find that the accuracies of SLDA and the proposed method increase obviously when the number of subspace dimensions increases from 5 to 25. This is mainly because few dimensions cannot hold much more intuitive discriminative information from the noisy data. Fig.17 also shows that the accuracy of SLDA decreases with the further increasing of the reduced dimension. This is mainly because the influence of noise will be enlarged with the increase of the reduced dimension. While the proposed method still obtains the best performance in each dimension. This also proves that by integrating the reconstruction constraint and sparse feature selection constraint, the proposed method can learn the optimal discriminative projection in each dimension so as to obtain a better classification performance.



(a) (b)

Fig. 17. Classification accuracy (%) versus the number of dimensions of different supervised learning methods on the Extended Yale B database with random pixel corruptions, in which (a) 10 samples, (b) 15 samples are randomly selected from each class as training set, respectively.

Table VIII shows the experimental results of different supervised learning methods on the AR face database with random pixel corruptions. From Table VIII, we can see that SLDA and the proposed method achieve much better performance than the remaining compared methods. This indicates that traditional LDA based methods without sparse constraint cannot catch sufficiently useful information of data for discriminant analysis under the random noise. This also proves that the sparse constraint can extract intuitive features and remove the redundancy information of data, and thus is beneficial to obtain a better performance. Table VIII also shows that the proposed method obtains a better performance than SLDA. Compared with SLDA which extracts features directly from the original data, the proposed method extracts features from the latent clear data of original data by introducing a sparse error term to compensate noises. Thus the proposed method can learn a more robust subspace and is reasonable to obtain a better performance than SLDA. From the experimental results on the two noisy face databases, we can see that the proposed method has the potential to reduce the negative influence of noise and effectively improves the classification accuracy.

VI. CONCLUSION

In this paper, we propose a novel supervised feature extraction method called RSLDA that integrates feature selection. By using a $l_{2,1}$ sparse norm to constrain the discriminative projection matrix, the proposed method can simultaneously select and extract the most discriminative features for classification. Moreover, to hold the main energy of original features, a data reconstruction term with an orthogonal constraint is introduced. This reconstruction constraint ensures the minimum loss of discriminative information so as to improve

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2799214, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VII
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE EXTENDED YALE B FACE DATABASE WITH RANDOM PIXEL CORRUPTIONS.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
10	21.02	34.58	8.88	56.05	33.64	27.58	45.47	30.79	27.09	40.55	**59.53**
15	24.20	46.17	8.50	60.04	25.13	17.97	45.37	28.68	18.93	24.65	**64.58**
20	26.74	53.25	4.90	62.48	10.92	5.69	37.87	28.62	5.74	7.52	**67.60**
25	29.13	58.01	10.94	63.41	26.84	20.14	39.39	28.66	19.95	27.73	**69.61**

TABLE VIII
CLASSIFICATION ACCURACY (%) VERSUS THE NUMBER OF TRAINING SAMPLES PER CLASS OF DIFFERENT METHODS ON THE AR FACE DATABASE WITH RANDOM PIXEL CORRUPTIONS.

No.	KNN	SVM	LDA	SLDA	OLDA	ULDA	DLA	MFA	SULDA	MPDA	RSLDA
4	47.92	57.79	6.91	83.42	74.33	55.72	74.89	65.83	57.58	72.27	**84.41**
6	54.91	69.07	6.68	89.48	74.85	48.65	62.14	66.25	48.04	82.26	**89.51**
8	60.20	77.01	5.78	92.76	66.62	31.99	49.55	60.14	33.56	88.93	**92.91**
12	66.74	86.73	4.09	95.71	53.85	16.56	23.37	48.28	17.59	94.04	**96.01**

the classification accuracy. In addition, we utilize a sparse error term to improve the robustness to the noise corruptions. The proposed method converges fast. Experimental results on six databases prove that compared with other competitive methods, the proposed method obtains the best performance. Experimental results also show that the proposed method can greatly improve the performance of image classification when these images are corrupted by noise.

REFERENCES

[1] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[2] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1209–1221, Jun. 2015.

[3] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, Jul.-Aug. 2012.

[4] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, no. Oct, pp. 1205–1224, Oct. 2004.

[5] Z. Zhang, L. Shao, Y. Xu, L. Liu, and J. Yang, "Marginal representation learning with graph structure self-adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, 2017, doi: 10.1109/TNNLS.2017.2772264.

[6] J. Ye, R. Janardan, and Q. Li, "Gpca: an efficient dimension reduction scheme for image compression and retrieval," in *International Conference on Knowledge Discovery and Data Mining*. ACM, 2004, pp. 354–363.

[7] J.-B. Yang and C.-J. Ong, "An effective feature selection method via mutual information estimation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 6, pp. 1550–1559, Mar. 2012.

[8] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 563–570, Aug. 2003.

[9] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-rank preserving projections," *IEEE Transactions on Cybernetics*, vol. 46, no. 8, pp. 1900–1913, Aug. 2016.

[10] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Transactions on Pattern analysis and Machine intelligence*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[11] J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, "Sparse representation classifier steered discriminative projection with applications to face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1023–1035, Jul. 2013.

[12] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2760–2771, Sep. 2015.

[13] Z. Zhang, Y. Xu, L. Shao, and J. Yang, "Discriminative block-diagonal representation learning for image recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2017, doi:10.1109/TNNLS.2017.2712801.

[14] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, 2004, pp. 153–160.

[15] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, Jan. 2010.

[16] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *IEEE International Conference on Computer Vision*. IEEE, Oct. 2005, pp. 1208–1213.

[17] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data," *Iaeng International Journal of Applied Mathematics*, vol. 39, no. 1, pp. 48–60, Jan. 2009.

[18] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, Feb. 2002.

[19] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.

[20] J. Ye and T. Xiong, "Null space versus orthogonal linear discriminant analysis," in *International Conference on Machine Learning*, Jun. 2006, pp. 1073–1080.

[21] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1312–1322, Oct. 2006.

[22] J. Yang, D. Zhang, X. Yong, and J.-y. Yang, "Two-dimensional discriminant transform for face recognition," *Pattern recognition*, vol. 38, no. 7, pp. 1125–1129, Jul. 2005.

[23] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and 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, Jan. 2007.

[24] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *European Conference on Computer Vision*, Oct. 2008, pp. 725–738.

[25] Y. Zhou and S. Sun, "Manifold partition discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 830–840, 2017.

[26] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear discriminant analysis using rotational invariant l1 norm," *Neurocomputing*, vol. 73, no. 13-15, pp. 2571–2579, 2010.

[27] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with l1-norm," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[28] M. Dorfer, R. Kelz, and G. Widmer, "Deep linear discriminant analysis," in *International Conference on Learning Representations*, 2015, pp. 1–13.

[29] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

[30] Z. Lai, Y. Xu, Z. Jin, and D. Zhang, "Human gait recognition via sparse discriminant projection learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1651–1662, Oct. 2014.

[31] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, Apr. 2011.

[32] X. Zhang, D. Chu, and R. C. Tan, "Sparse uncorrelated linear discriminant analysis for undersampled problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1469–1485, 2016.

[33] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[34] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 796–808, Apr. 2015.

[35] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley, New York, 1973.

[36] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[37] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[38] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. K. Wong, "Robust latent subspace learning for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2017, doi: 10.1109/TNNLS.2017.2693221.

[39] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, Feb. 2006.

[40] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[41] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11–48, May 2011.

[42] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition." in *International Conference on Machine Learning*, vol. 32, 2014, pp. 647–655.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[44] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *IEEE International Conference on Computer Vision*, 2010, pp. 221–228.

[45] S. Cai, L. Zhang, W. Zuo, and X. Feng, "A probabilistic collaborative representation based approach for pattern classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2950–2959.

[46] S. A. Nene, S. K. Nayar, H. Murase *et al.*, "Columbia object image library (coil-20)," 1996.

[47] A. M. Martinez, "The ar face database," *Cvc Technical Report*, vol. 24, 1998.

[48] A. S. Georghiades, P. N. Belhumeur, and 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.

[49] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 53–59.

[50] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," *California Institute of Technology*, 2007.

[51] N. Pinto, Z. Stone, T. Zickler, and D. Cox, "Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2011, pp. 35–42.

[52] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[53] T. Tommasi and T. Tuytelaars, "A testbed for cross-dataset analysis," in *European Conference on Computer Vision*, vol. 8927, 2014, pp. 18–31.

[54] B. Becker and E. Ortiz, "Evaluating open-universe face identification on the web," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 904–911.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

Jie Wen received the M.S. degree at Harbin Engineering University, China in 2015. He is currently pursuing the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His research interests include, image and video processing, pattern recognition and machine learning.

Xiaozhao Fang received his M.S. degree in 2008, and the Ph.D. degree in computer science and technology at Shenzhen Graduate School, HIT, Shenzhen (China) in 2016. He is currently with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition and machine learning.

Jinrong Cui received the Ph.D.degree in computer science and technology at Shenzhen Graduate School, HIT, Shenzhen (China) in 2015. She is currently with the College of Mathematics and Informatics, South China Agricultural University. Her current research interests include pattern recognition and machine learning.

Lunke Fei received the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. He is currently with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition and biometrics.

Ke Yan received the M.S. degree in computer science and technology from the Shenzhen Graduate School, Harbin Institute of Technology University, China, in 2015. He is currently pursuing the Ph.D. degree in computer science and technology with the Shenzhen Graduate School, Harbin Institute of Technology, China. His research interests include pattern recognition, machine learning and bioinformatics.

Yan Chen received her B.E. and M.E. degree in computer science from Northeastern University, China in 1997 and 2000 respectively, and her Ph.D. in 2010 from university of Technology, Sydney(UTS), Australian. She is a R&D member of Shenzhen Sunwin Intelligent Co. Ltd. Her research interests include computer vision and pattern recognition.

Yong Xu received his B.S. degree, M.S. degree in 1994 and 1997, respectively. He received the Ph.D. degree in Pattern Recognition and Intelligence system at NUST (China) in 2005. Now he works at Shenzhen Graduate School, Harbin Institute of Technology. His current interests include pattern recognition, biometrics, machine learning and video analysis. More information please refer to http://www.yongxu.org/lunwen.html.