Pattern Recognition 96 (2019) 106963

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse approximation to discriminant projection learning and application to image classification☆

Yu-Feng Yu[a], Chuan-Xian Ren[b], Min Jiang[c], Man-Yu Sun[b], Dao-Qing Dai[b], Guodong Guo[c,*]

[a] Department of Statistics and Institute of Intelligent Finance, Accounting & Taxation, Guangzhou University, Guangzhou 510006, China
[b] Intelligent Data Center and Department of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China
[c] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA

ARTICLE INFO

Article history:
Received 1 December 2018
Accepted 10 July 2019
Available online 12 July 2019

Keywords:
Image classification
Feature selection
Subspace learning
Discriminant analysis
Dimensionality reduction

ABSTRACT

Subspace learning for dimensionality reduction is an important topic in pattern analysis and machine learning, and it has extensive applications in feature representation and image classification. Linear discriminant analysis (LDA) is a well-known subspace learning approach for supervised dimensionality reduction due to its effectiveness and efficacy in discriminant analysis. However, LDA is not stable and suffers from the singularity problem when addressing small sample size and high-dimensional data. In this paper, we develop a novel subspace learning model, named sparse approximation to discriminant projection learning (SADPL), to learn the sparse projection matrix. Different from the traditional LDA-based methods, we learn the projection matrix based on a new objective function rather than the Fisher criterion, which avoids the matrix singularity problem. In order to distinguish which features play an important role in discriminant analysis, we embed a feature selection framework to the subspace learning model to select the informative features. Finally, we can attain a convex objective function which can be solved by an effective optimization algorithm, and theoretically prove the convergence of the proposed optimization algorithm. Extensive experiments on all sorts of image classification tasks, such as face recognition, palmprint recognition, object categorization and texture classification show that our SADPL achieves competitive performance compared to the state-of-the-art methods.

1. Introduction

In computer vision and machine learning, image classification is an important research topic whose goal is to classify an individual image into a specific category [1]. However, the high-dimensional images make the classification problem difficult. In order to deal with this problem, many subspace learning-based dimensionality reduction methods have been proposed [2–4]. They in general include unsupervised [5], semi-supervised [6] and supervised learning approaches [7]. Among them, the classical ones include principal component analysis (PCA) [8], linear discriminant analysis (LDA) [9–11], and so on. PCA is an unsupervised subspace learning approach, which is to find a projection matrix by maximizing the determinant of the total scatter matrix of the training images and project the original image space into a low-rank subspace. LDA has been shown better performance than PCA, and it has been used with success in a variety of specific applications such as face recognition [12–14] and image classification [15,16]. The basic idea of LDA is to find a projection matrix by maximizing the between-class variation and minimizing the within-class variation, and it needs to estimate the inverse of within-class covariance matrix, thus the performance will degenerate rapidly in the case of small sample size (SSS).

In the past few years, this problem has attracted a lot of attentions, and many methods-based LDA have been presented to improve the performance and efficiency. The straightforward method is PCA plus LDA which first adopts PCA to decrease the dimension of the image space, and then utilizes the classical LDA to decrease the dimension to $C-1$, where C is the number of sample classes. However, Dai et al. [17] have proven that LDA still fails even after a PCA procedure. They propose a regularized discriminant analysis (RDA) model to deal with the SSS problem. Penalized discriminant analysis (PDA) is also an improvement of LDA [18,19]. The objective of PDA is to deal with the SSS problem and improve the discriminative ability to smooth the coefficients of

☆ This work is supported in part by the National Science Foundation of China under Grants 11631015, and U1611265, and partially supported by a project of Center of Identification Technology Research (CITeR).

* Corresponding author.
E-mail addresses: yuyufeng220@163.com (Y.-F. Yu), rchuanx@mail.sysu.edu.cn (C.-X. Ren), minjiang.aca@gmail.com (M. Jiang), sunmy8@mail2.sysu.edu.cn (M.-Y. Sun), stsddq@mail.sysu.edu.cn (D.-Q. Dai), guodong.guo@mail.wvu.edu (G. Guo).

https://doi.org/10.1016/j.patcog.2019.106963
0031-3203/

discriminant vectors. However, the major shortcoming of PDA is that it has no good flexibility [20]. Considering that graph embedding technique has been used with success, Cai et al. [21] consider both spectral graph and regression analysis to present a novel method for improving the discriminative ability, called SRDA. Specifically, SRDA does not need to compute the eigenvector, and only deals with the regularized least square problems. Thus, it can reduce the computational complexity and storage cost. Zheng et al. [22] argue that the class empirical mean may not be equivalent to the expectation in practice and develop a perturbation LDA via utilizing the perturbation random vectors. Cai et al. [23] consider that the covariance matrix cannot be computed effectively if the training samples are not enough and present a semi-supervised discriminant analysis model (SDA), in which the labeled training samples are utilized to extract discriminant structure and both labeled and unlabeled training samples are used to extract the intrinsic geometrical structure. Lai et al. [24] argue that these methods which utilize the l_2-norm to depict the scatter matrix of the data are sensitive to the outliers, and adopt the $l_{2,1}$-norm to propose an unified rotational invariant LDA (RILDA) model for dimensionality reduction.

Recently, local structure has been proven that it is important to subspace learning for dimensionality reduction. Cai et al. [25] consider that local structure is more helpful for improving discriminative ability than global structure when the training samples are not enough, and propose a locality sensitive discriminant analysis (LSDA) model. The objective of LSDA is to learn a projection matrix and project the original sample space into a subspace. And in the subspace the samples of the same class should be close and the samples of the different class should be far apart. Moreover, locality preserving projection (LPP) [26] has also been embedded into subspace learning. Sugiyama [27] utilizes the locality preserving property of LPP to deal with multi-modal samples, and present a local FDA model. Fan et al. [28] argue that the local sample structure is more effectively than the global structure for discriminant subspace learning and propose an LLDA algorithm, in which the local linear discriminant vectors are learnt to construct the projection matrix.

It should be noted that these subspace learning methods have a common shortcoming that the learnt low-dimensional features are the combination of all original features. These features are not distinguished which ones play an important role in discriminant subspace learning [29]. In order to address this problem, some sparse and robust methods have been presented to extract the important information for subspace learning and dimensionality reduction. Typical ones like l_1-PCA [30], R_1-PCA [31] and l_1-LDA [32] adopt l_1-norm to replace l_2-norm for improving the sparsity and robustness. Kwak [30] proposes a robust PCA based on l_1-norm (l_1-PCA), which is robust to outliers and is also rotational invariant. The l_1-norm optimization can be solved by a simple and efficient algorithm to find a locally maximal solution. Similar to Kwak [30], Ding et al. [31] also present a rotational invariant l_1-norm PCA (R_1-PCA) which softens the effects of outliers. Different from l_1-PCA, R_1-PCA can find a unique global solution. To improve the robustness and sparsity of LDA, Zhong et al. [32] propose a l_1-norm LDA, which can also effectively overcome the singular problem. In addition, sparse PCA (SPCA) [33] extracts the sparse principle components by combining the least angle regression [34] and l_1-norm elastic net [35] regression. Sparse discriminant analysis (SDA) [36] imposes a sparseness criterion to linear discriminant analysis such that feature selection and classification can be implemented at the same time. Sparse locality-preserving embedding (SLPE) [37] incorporates l_1 penalty with conventional locality preserving projections to learn sparse projections. In addition, sparse LDA (SLDA) [38] learns the sparse projections via imposing the lasso constraint [39]. It can be used to address the data piling problem. However, as mentioned in [40], for SDA, SLPE, and SLDA, there may still be the matrix singularity and small sample size problems.

Recently, $l_{2,1}$-norm has been commonly used in feature selection. For instance, Liu et al. [41] impose the $l_{2,1}$-norm on the transformation matrix to implement feature selection. He et al. [42] present a $l_{2,1}$-norm regularized correntropy model to extract informative features. Then an effective alternate optimization algorithm is proposed to solve the non-convex correntropy objective function. In [43], a novel robust linear discriminant analysis (RLDA) via using the $l_{2,1}$-norm to replace l_2-norm has been presented, in which the $l_{2,1}$-norm can be embedded into the linear discriminant analysis to improve the robustness. As mentioned in [44], Yang et al. incorporate discriminative analysis and $l_{2,1}$-norm regularization term into a joint model to select the discriminative features. Nie et al. [45] adopt $l_{2,1}$-norm on both loss function and regularization term to improve the effectiveness of feature selection.

Motivated by recent process in subspace learning and feature selection, in this paper we propose a novel subspace learning and feature selection algorithm, called sparse approximation to discriminant projection learning (SADPL). The proposed SADPL has resemblance to some subspace learning-based LDA methods [13,27,28,46,47], but is different from those. Those subspace learning-based LDA methods consider the eigenvectors-matrix of $\mathbf{S_W}^{-1}\mathbf{S_B}$ corresponding to nonzero eigenvalues as the projection matrix (Here $\mathbf{S_B}$ denotes the between-class scatter matrix and $\mathbf{S_W}$ is the within-class scatter matrix.), while SADPL considers $(\mathbf{S_W}+\mathbf{S_B})^{-1}\mathbf{A}$ as projection matrix (Here $\mathbf{A}$ is a low-rank matrix and $\mathbf{S_B}=\mathbf{A}\mathbf{A}^T$.). In addition, SADPL differs from other subspace learning-based LDA methods in that SADPL estimates the projection matrix based on a new objective function rather than the traditional Fisher criterion, and thus without the matrix singularity problem caused by the eigenvalue decomposition on Fisher criterion to get the projection matrix. In addition, there is no feature selection in [13,27,28,46] to distinguish which features play an important role in discriminant subspace learning, while the proposed SADPL utilizes feature selection to extract the important information for subspace learning and dimensionality reduction. Although MGSDA [48] and L21FLDA [29] also consider feature selection, they only add $l_{2,1}$-norm penalty term to the objective function, different from the objective function of SADPL. Moreover, SADPL adopts the F-norm and $l_{2,1}$-norm penalty terms simultaneously, which makes the solution of the objective function more stable while achieving sparse. Finally, the derived objective function of SADPL, which is convex, can be solved by an effective optimization algorithm, and the convergence of the proposed optimization algorithm can be proved theoretically.

Our main contributions include:

- Developing a new estimation method of projection matrix. Different from the conventional subspace learning based LDA methods, SADPL computes $(\mathbf{S_W}+\mathbf{S_B})^{-1}\mathbf{A}$ as projection matrix, which can avoid the matrix singularity problem.
- Joint using of F-norm and $l_{2,1}$-norm to embed a feature selection framework into the subspace learning, which is effective to select informative features and lead a sparse subspace.
- Proposing a supervised sparse discriminant projection learning algorithm, which preforms subspace learning and feature selection simultaneously. It also guarantees a globally optimal solution.
- Proposing an efficient optimization algorithm to effectively solve the derived objective function, which can be theoretically proved for the convergence.

The remainder of this paper is organized as follows. Section 2 presents our subspace learning and feature selection algorithm, i.e., sparse approximation to discriminant projection

learning. In Section 3, experiments are conducted to validate the proposed method, and compared to the state-of-the-art methods on various image databases to show the competitive performance of our algorithm. Finally, we draw the conclusion in Section 4.

2. Sparse approximation to discriminant projection learning

In this section, we introduce the proposed method called sparse approximation to discriminant projection learning. The main content will be separated into the following several parts including theoretical background, the derived objective function, algorithm optimization and computational complexity analysis.

2.1. Notations

For a vector $\mathbf{x} \in R^d$, we define the l_2-norm as $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$. For a matrix $\mathbf{M} \in R^{d\times n}$, the ith row and jth column are defined as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. The F-norm and $l_{2,1}$-norm of the matrix $\mathbf{M}$ are defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{d} \|\mathbf{m}^i\|_2^2} = \sqrt{\sum_{j=1}^{n} \|\mathbf{m}_j\|_2^2}$ and $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{m}^i\|_2 = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{n} m_{ij}^2}$, respectively.

2.2. Theoretical background

We assume that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in R^{d\times n}$ is a data matrix which consists of n samples from C classes. For the cth class, it contains n_c samples, $c = 1, 2, \ldots, C$. Let $\mathbf{u}^c$ and $\mathbf{u}$ be the sample mean for class c and the overall sample mean, respectively. Then the between-class scatter matrix $\mathbf{S_B}$ and within-class scatter matrix $\mathbf{S_W}$ are denoted as [11]

$$\mathbf{S_B} = \sum_{c=1}^{C} n_c(\mathbf{u}^c - \mathbf{u})(\mathbf{u}^c - \mathbf{u})^T, \tag{1}$$

$$\mathbf{S_W} = \sum_{c=1}^{C} \sum_{i=1}^{n_c} (\mathbf{x}_i^c - \mathbf{u}^c)(\mathbf{x}_i^c - \mathbf{u}^c)^T, \tag{2}$$

where $\mathbf{x}_i^c$ is the ith sample from class c.

Our subspace learning algorithm is built on the properties of low-rank decomposition of matrix and representation of eigenvectors-matrix [48], which are shown in Propositions 1 and 2.

Proposition 1. *Define $\mathbf{S_B}$ as in (1). There exists a low-rank matrix $\mathbf{A}$ such that $\mathbf{S_B} = \mathbf{A}\mathbf{A}^T$, and the kth column of $\mathbf{A}$ is denoted as*

$$\mathbf{A}_k = \frac{\sqrt{n_{k+1}}\left(\sum_{r=1}^{k} n_r(\mathbf{u}^r - \mathbf{u}^{k+1})\right)}{\sqrt{n}\sqrt{\sum_{r=1}^{k} n_r \sum_{r=1}^{k+1} n_r}}, \tag{3}$$

where $k = 1, \ldots, C-1$.

Note that the columns of the matrix $\mathbf{A}$ can be seen as the differences among the class means. Moreover, they denote orthogonal contrasts among the means of C classes. The proof of this Proposition can be found in [48].

Based on the Proposition 1, we obtain the intuitive representation for the eigenvectors-matrix of $\mathbf{S_W}^{-1}\mathbf{S_B}$.

Proposition 2. *Assume $\mathbf{A}$ is defined as in (3), then existing an orthogonal matrix $\mathbf{Q}$ such that $(\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}\mathbf{Q}$ is the eigenvectors-matrix of $\mathbf{S_W}^{-1}\mathbf{S_B}$, corresponding to nonzero eigenvalues.*

Proof. According to the eigenvalue decomposition of matrix, existing an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{A}^T(\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, and $\mathbf{\Lambda}$ is a diagonal matrix, in which the ith diagonal element μ_i is the ith nonzero eigenvalue. Assume $\mathbf{\Phi} = (\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}\mathbf{Q}$, then we have

$$(\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{S_B}\mathbf{\Phi} = (\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{S_B}(\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}\mathbf{Q}.$$

Substituting $\mathbf{S_B}$ by $\mathbf{A}\mathbf{A}^\mathbf{T}$, we get

$$\begin{aligned}(\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{S_B}\mathbf{\Phi} &= (\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}\mathbf{Q}\mathbf{\Lambda} \\ &= \mathbf{\Phi}\mathbf{\Lambda}.\end{aligned} \tag{4}$$

Then let's multiply both sides of Eq. (4) by $\mathbf{S_W} + \mathbf{S_B}$, we have

$$\mathbf{S_B}\mathbf{\Phi} = \mathbf{S_W}\mathbf{\Phi}\mathbf{\Lambda} + \mathbf{S_B}\mathbf{\Phi}\mathbf{\Lambda}.$$

After performing a simple transformation, we have

$$\mathbf{S_B}\mathbf{\Phi}(\mathbf{I} - \mathbf{\Lambda}) = \mathbf{S_W}\mathbf{\Phi}\mathbf{\Lambda},$$

and

$$\mathbf{S_W}^{-1}\mathbf{S_B}\mathbf{\Phi} = \mathbf{\Phi}\mathbf{\Lambda}(\mathbf{I} - \mathbf{\Lambda})^{-1}. \tag{5}$$

□

In (5), according to the definition of eigenvalue decomposition, we can see that $\mathbf{\Phi} = (\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}\mathbf{Q}$ is the eigenvectors-matrix of $\mathbf{S_W}^{-1}\mathbf{S_B}$.

It should be noted that we assume the matrix $\mathbf{I} - \mathbf{\Lambda}$ is invertible in (5), that is, every nonzero eigenvalue in $\mathbf{\Lambda}$ is not equal to one. In most cases, this assumption always holds. However, if it is not invertible, we can set $\mathbf{I} - \mathbf{\Lambda}$ as $\mathbf{I} - \mathbf{\Lambda} + \epsilon\mathbf{I}$, here ϵ is a perturbation constant. It also holds to our method.

Similar to the LDA-based methods, a straightforward way is to denote $\mathbf{\Phi}$ as the projection matrix, but the orthogonal matrix $\mathbf{Q}$ is unknown. In many state-of-the-art image classification methods, an image is represented as a high-dimensional feature vector, and the high-dimensional feature vector can be projected into a low-dimensional feature vector by a projection matrix, then the similarity between a query image and a gallery image is evaluated on the basis of the similarity measure between the low-dimensional feature vectors. Here we will prove that the orthogonal transformation makes no difference to the classification criteria based on various similarity measures, such as Euclidean distance, inner product, correlation coefficient and Mahalanobis distance. Without loss of generality, we consider Euclidean distance as the similarity measure, and have the following proposition.

Proposition 3. *For any orthogonal matrix $\mathbf{Q}$, the Euclidean distance of any two samples based on the projection matrix $\mathbf{P}$ is the same as the Euclidean distance based on the projection matrix $\mathbf{PQ}$.*

Proof. Assume $\mathbf{y}$ is the new testing sample and $\mathbf{x}$ is the training sample, then we have the Euclidean distance based on the projection matrix $\mathbf{P}$ as $\|\mathbf{P}^T\mathbf{y} - \mathbf{P}^T\mathbf{x}\|_2^2 = (\mathbf{y} - \mathbf{x})^T\mathbf{P}\mathbf{P}^T(\mathbf{y} - \mathbf{x})$. The Euclidean distance based on the projection matrix $\mathbf{PQ}$ can be defined as $\|(\mathbf{PQ})^T\mathbf{y} - (\mathbf{PQ})^T\mathbf{x}\|_2^2 = (\mathbf{y} - \mathbf{x})^T\mathbf{P}\mathbf{Q}\mathbf{Q}^T\mathbf{P}^T(\mathbf{y} - \mathbf{x}) = (\mathbf{y} - \mathbf{x})^T\mathbf{P}\mathbf{P}^T(\mathbf{y} - \mathbf{x})$. Hence, it proves the correctness of Proposition 3. □

Note that the Proposition 3 also holds to other similarity measures, such as inner product, correlation coefficient and Mahalanobis distance.

2.3. Objective function

Propositions 2 and 3 show that we can consider $\tilde{\mathbf{P}} = (\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}$ as the projection matrix, then we can denote the objective function as:

$$\begin{aligned}J(\mathbf{P}) &= \min_{\mathbf{P}} \frac{1}{2}\|(\mathbf{S_W} + \mathbf{S_B})^{1/2}\mathbf{P} - (\mathbf{S_W} + \mathbf{S_B})^{-1/2}\mathbf{A}\|_F^2 \\ &= \min_{\mathbf{P}} \frac{1}{2} Tr\left(\mathbf{P}^T(\mathbf{S_W} + \mathbf{S_B})\mathbf{P} - 2\mathbf{A}^T\mathbf{P}\right).\end{aligned} \tag{6}$$

Obviously, by performing the derivative of (6) with respect to $\mathbf{P}$ as zero, we can get $\tilde{\mathbf{P}} = (\mathbf{S_W} + \mathbf{S_B})^{-1}\mathbf{A}$.

For simplification, we have the following objective function:

$$J(\mathbf{P}) = \min_{\mathbf{P}} \frac{1}{2} Tr(\mathbf{P}^T \mathbf{S_W} \mathbf{P}) + \frac{1}{2} Tr(\mathbf{P}^T \mathbf{S_B} \mathbf{P} - 2\mathbf{A}^T \mathbf{P}) = \min_{\mathbf{P}} \frac{1}{2} Tr(\mathbf{P}^T \mathbf{S_W} \mathbf{P}) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P} - \mathbf{I}\|_F^2. \quad (7)$$

Our next aim is to propose a regularizer and integrate it into (7) to construct a sparse feature selection model. Note that l_1-norm regularizer $\|\mathbf{P}\|_1$ is widely used to improve the sparsity of the elements. As indicated in [49], a common feature of the methods using l_1-norm is that they do not distinguish the difference of the two indices (row and column of $\mathbf{P}$). But in general they have different meaning. Here they are the spatial dimensions and the number of features, respectively. In addition, our objective function estimates the eigenvectors-matrix of $\mathbf{S_W}^{-1}\mathbf{S_B}$ up to an orthogonal transformation and the element-wise sparsity via l_1-norm cannot be maintained under the orthogonal transformation.

Instead of l_1-norm, we consider the $l_{2,1}$-norm regularizer as follows:

$$\|\mathbf{P}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{p}^i\|_2 = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{C-1} p_{ij}^2}, \quad (8)$$

where $\mathbf{p}^i$ denotes the ith row of matrix $\mathbf{P}$. The $l_{2,1}$-norm regularizer leads to the row sparsity, namely it promotes the rows of matrix to get zero elements and considers the correlations of all features. Moreover, we can see that the row sparsity via $l_{2,1}$-norm is maintained under orthogonal transformation. It should be noted that $l_{2,1}$-norm is also introduced in [50,51], but they use it for multi-task learning and semi-supervised learning rather than subspace learning.

In addition, we embed the F-norm regularizer $\|\mathbf{P}\|_F^2$ and equation (8) into (7), we have the objective function of the presented method as:

$$J(\mathbf{P}) = \min_{\mathbf{P}} \frac{1}{2} Tr(\mathbf{P}^T \mathbf{S_W} \mathbf{P}) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P} - \mathbf{I}\|_F^2 + \frac{\lambda_1}{2} \|\mathbf{P}\|_F^2 + \lambda_2 \|\mathbf{P}\|_{2,1}, \quad (9)$$

where both λ_1 and λ_2 are tuning parameters. The goal of the first two items of the objective function in (9) is to minimize the within-class variation and control the between-class variation. It can be seen that the third item helps to avoid the matrix singularity problem. The fourth item can eliminate noisy features in the process of feature selection, and reduce the model complexity. Moreover, as indicated in [52], the joint use of $\|\mathbf{P}\|_F^2$ and $\|\mathbf{P}\|_{2,1}$ is to select the informative features and make the solution of $J(\mathbf{P})$ more stable while being sparse.

2.4. Optimization algorithm

Notice that (9) is a convex optimization problem, and thus it does not suffer from the multiple local minima issue, and its global minimization can be solved efficiently. Some algorithms have been presented to solve a similar problem [52,53]. In this paper, we adopt an alternating optimization strategy to find the global solution.

Taking the derivative of $J(\mathbf{P})$ with respect to $\mathbf{P}$ as zero, we can get

$$\frac{\partial J(\mathbf{P})}{\partial \mathbf{P}} = \mathbf{S_W}\mathbf{P} + \mathbf{A}(\mathbf{A}^T\mathbf{P} - \mathbf{I}) + \lambda_1 \mathbf{P} + \lambda_2 \mathbf{B}\mathbf{P}, \quad (10)$$

where $\mathbf{B}$ is a diagonal matrix with the ith diagonal element

$$B_{i,i} = \frac{1}{2\|\mathbf{p}^i\|_2}. \quad (11)$$

Setting the derivative in (10) to zero, we get

$$\mathbf{P} = (\mathbf{S_W} + \mathbf{A}\mathbf{A}^T + \lambda_1 \mathbf{I} + \lambda_2 \mathbf{B})^{-1}\mathbf{A}. \quad (12)$$

By observing (11), we can see that the matrix $\mathbf{B}$ is not independent of the value of $\mathbf{P}$. Hence, we utilize an iterative strategy to optimize $\mathbf{P}$ and $\mathbf{B}$ alternately. Algorithm 1 summarizes the optimization procedure. At each iteration, one of the variables $\mathbf{P}$ and $\mathbf{B}$ is fixed, while the other is updated, and then the roles of $\mathbf{P}$ and $\mathbf{B}$ are exchanged. We will prove that the value of the objective function $J(\mathbf{P})$ monotonically decreases along with $\mathbf{P}$ and $\mathbf{B}$ being updated at each iteration, and iterations are repeated until convergence.

Algorithm 1 SADPL algorithm.

Input: $\mathbf{X} \in R^{d\times n}$, λ_1 and λ_2.
Output: $\mathbf{P} \in R^{d\times C-1}$.
1: Compute within-class scatter matrix $\mathbf{S_W}$ and between-class scatter matrix $\mathbf{S_B}$;
2: Compute the matrix $\mathbf{A}$ based on Proposition 1;
3: Initialize $t = 0$;
4: Initialize $\mathbf{B_0}$;
5: **while** Not convergent **do**
6: $\quad \mathbf{P}^{[t+1]} = (\mathbf{S_W} + \mathbf{A}\mathbf{A}^T + \lambda_1 \mathbf{I} + \lambda_2 \mathbf{B}^{[t]})^{-1}\mathbf{A}$;
7: $\quad$ Update $\mathbf{B}^{[t+1]}$, here the ith diagonal element $B_{i,i}^{[t+1]} = 1/2\|(\mathbf{p}^{[t+1]})^i\|_2$;
8: $\quad t = t + 1$;
9: **end while**

Theorem 1. *In Algorithm 1, the value of the objective function $J(\mathbf{P})$ monotonically decreases along with the iteration.*

Proof. Assume $\mathbf{P}^{[t+1]}$ is the result of the $t+1$th iteration, and according to the Algorithm 1, we can get

$$\mathbf{P}^{[\mathbf{t+1}]} \leftarrow \min_{\mathbf{P}} \frac{1}{2} Tr(\mathbf{P}^T \mathbf{S_W} \mathbf{P}) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P} - \mathbf{I}\|_F^2 + \frac{\lambda_1}{2} \|\mathbf{P}\|_F^2 + \lambda_2 Tr(\mathbf{P}^T \mathbf{B}^{[t]} \mathbf{P}), \quad (13)$$

then, we have

$$\begin{aligned} &\frac{1}{2} Tr\left((\mathbf{P}^{[t+1]})^T \mathbf{S_W} \mathbf{P}^{[t+1]}\right) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P}^{[t+1]} - \mathbf{I}\|_F^2 \\ &+ \frac{\lambda_1}{2} \|\mathbf{P}^{[t+1]}\|_F^2 + \lambda_2 Tr\left((\mathbf{P}^{[t+1]})^T \mathbf{B}^{[t]} \mathbf{P}^{[t+1]}\right) \\ \le\ &\frac{1}{2} Tr\left((\mathbf{P}^{[t]})^T \mathbf{S_W} \mathbf{P}^{[t]}\right) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P}^{[t]} - \mathbf{I}\|_F^2 \\ &+ \frac{\lambda_1}{2} \|\mathbf{P}^{[t]}\|_F^2 + \lambda_2 Tr\left((\mathbf{P}^{[t]})^T \mathbf{B}^{[t]} \mathbf{P}^{[t]}\right). \end{aligned} \quad (14)$$

From (14), we get

$$\begin{aligned} &\frac{1}{2} Tr\left((\mathbf{P}^{[t+1]})^T \mathbf{S_W} \mathbf{P}^{[t+1]}\right) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P}^{[t+1]} - \mathbf{I}\|_F^2 \\ &+ \frac{\lambda_1}{2} \|\mathbf{P}^{[t+1]}\|_F^2 + \lambda_2 \sum_{i=1}^{d} \frac{\|(\mathbf{p}^{[t+1]})^i\|_2^2}{2\|(\mathbf{p}^{[t]})^i\|_2} \\ \le\ &\frac{1}{2} Tr\left((\mathbf{P}^{[t]})^T \mathbf{S_W} \mathbf{P}^{[t]}\right) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P}^{[t]} - \mathbf{I}\|_F^2 \\ &+ \frac{\lambda_1}{2} \|\mathbf{P}^{[t]}\|_F^2 + \lambda_2 \sum_{i=1}^{d} \frac{\|(\mathbf{p}^{[t]})^i\|_2^2}{2\|(\mathbf{p}^{[t]})^i\|_2}. \end{aligned} \quad (15)$$

Taking a simple transformation, we obtain

$$\lambda_2 \sum_{i=1}^{d} \frac{\|(\mathbf{p}^{[t+1]})^i\|_2^2}{2\|(\mathbf{p}^{[t]})^i\|_2} = \lambda_2 \sum_{i=1}^{d} \|(\mathbf{p}^{[t+1]})^i\|_2 - \lambda_2 \left(\sum_{i=1}^{d} \|(\mathbf{p}^{[t+1]})^i\|_2 - \sum_{i=1}^{d} \frac{\|(\mathbf{p}^{[t+1]})^i\|_2^2}{2\|(\mathbf{p}^{[t]})^i\|_2} \right), \quad (16)$$

and

$$\lambda_2 \sum_{i=1}^{d} \frac{\|(\mathbf{p}^{[t]})^i\|_2^2}{2\|(\mathbf{p}^{[t]})^i\|_2} = \lambda_2 \sum_{i=1}^{d} \|(\mathbf{p}^{[t]})^i\|_2 - \lambda_2 \left(\sum_{i=1}^{d} \|(\mathbf{p}^{[t]})^i\|_2 - \sum_{i=1}^{d} \frac{\|(\mathbf{p}^{[t]})^i\|_2^2}{2\|(\mathbf{p}^{[t]})^i\|_2} \right). \quad (17)$$

Substitute $\|(\mathbf{p}^{[t+1]})^i\|_2$ and $\|(\mathbf{p}^{[t]})^i\|_2$ by b and a, respectively. Using the property $b - \frac{b^2}{2a} \le a - \frac{a^2}{2a}$ and combining (15)–(17) together, we get

$$\begin{aligned} &\frac{1}{2} Tr\left((\mathbf{P}^{[t+1]})^T \mathbf{S_W} \mathbf{P}^{[t+1]}\right) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P}^{[t+1]} - \mathbf{I}\|_F^2 \\ &+ \frac{\lambda_1}{2} \|\mathbf{P}^{[t+1]}\|_F^2 + \lambda_2 \sum_{i=1}^{d} \|(\mathbf{p}^{[t+1]})^i\|_2 \\ &\le \frac{1}{2} Tr\left((\mathbf{P}^{[t]})^T \mathbf{S_W} \mathbf{P}^{[t]}\right) + \frac{1}{2} \|\mathbf{A}^T \mathbf{P}^{[t]} - \mathbf{I}\|_F^2 \\ &+ \frac{\lambda_1}{2} \|\mathbf{P}^{[t]}\|_F^2 + \lambda_2 \sum_{i=1}^{d} \|(\mathbf{p}^{[t]})^i\|_2. \end{aligned} \quad (18)$$

Obviously, from (18), we can see that the objective function $J(\mathbf{P})$ monotonically decreases along with the iteration, i.e. $J(\mathbf{P}^{[t+1]}) \le J(\mathbf{P}^{[t]})$. □

Hence, based on Algorithm 1, the objective function $J(\mathbf{P})$ can get the optimum solution. In Section 3.3, we can find that the objective function converges very fast.

2.5. Complexity analysis

We now discuss the computational complexity of the presented SADPL algorithm. Similar to the other LDA-based approaches, such as LLDA, we first perform PCA to obtain p principal components, and then use SADPL algorithm to learn projection matrix, namely PCA+SADPL. Because the PCA stage of SADPL is the same as the other algorithms, we only consider the computational complexity of the SADPL stage. The SADPL algorithm contains two main parts: the computation of the scatter matrix and the optimization of the projection matrix. Suppose that the number of training samples and the maximum number of iterations are n and T, respectively, the time complexities of the scatter matrix and the projection matrix optimization are $O(np^2)$ and $O(Tp^3)$, respectively. Thus, the time complexity of SADPL stage is $O(np^2 + Tp^3)$.

3. Experiments

To evaluate performance of the presented SADPL model, we compare it with both the classical and the state-of-the-art approaches for image classification, including PCA [8], LDA [11], LPP [26], LSDA [25], SRDA [21], RILDA [24], LLDA [28] and SRRS [54]. We use MATLAB codes of these compared methods (except for PCA and LDA), which are released/provided by the corresponding authors. Five publicly available image databases, that is, the FRGC [55] face database, the KTH-TIPS[56] texture database, the 2D plus 3D palmprint database [57], the COIL-20 [58] object database and the CIFAR-10 [59] tiny images database are used to demonstrate the performance of different methods. In our experiments, we consider PCA as the baseline algorithm, and retain 600 principle components for all the compared algorithms on the five image databases. The 1-nearest neighbor classifier is applied to the projected samples for classification.

3.1. Databases

The descriptions of the four databases are as follows:

Fig. 1. Some face images of the FRGC database. First row: Uncontrolled lighting variations; Second row: Controlled lighting variations.

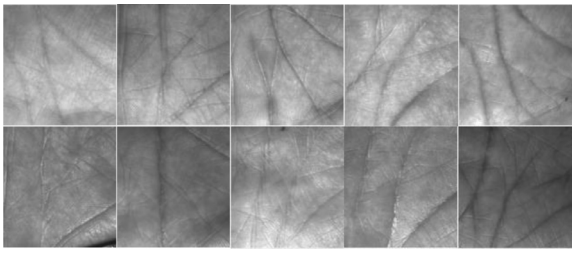

Fig. 2. Some palmprint images of the 2D plus 3D palmprint database. First row: Session 1; Second row: Session 2.

(1) FRGC face database: The FRGC database [55] includes 625 subjects and there are about 50,000 images. These images were collected at different time, under controlled and uncontrolled variations (occlusion, expression, illumination, etc.). In order to show the different evaluation purposes, we perform two groups of experiments. In the first group, the images with controlled illumination variations are used as a subset, which contains 1375 images of 275 subjects and 5 images for each. These images are cropped and resized to 100×100. For each classification, we randomly choose three images from each subject for training and the rest two for testing. In the second group, the images with uncontrolled illumination variations are used as the other subset, which were obtained in large illumination variations (outside, atriums, hallways, etc.), ageing and image blur. This subset also has 275 subjects and 5 images for each. Similar to the first group, these images are cropped and resized to 100×100 and we also choose three images from each subject for training and the rest two for testing. Some cropped images in this database are shown in Fig. 1.

(2) 2D plus 3D palmprint database: The 2D plus 3D palmprint database [57] includes 400 different palms, and there are 8000 samples, that is, each palm has two separated sessions and each session has ten palmprint samples. The interval of time between the two sessions is about 30 days. Each sample includes a 2D ROI (region of interest) and its corresponding 3D ROI. All samples are cropped and sized to 128×128. Some 2D palmprint images in this database are shown in Fig. 2. In this paper, the 2D ROI images are utilized to evaluate the performance. In order to show the different evaluation purposes, we divide the experiments into two groups. In the first group, that is, each palm has ten palmprint samples, and two of which are randomly selected for training and the other eight images for testing. In the second group, all images in session 2 are used as the other subset. Similar to the first group, we also choose 2 samples from each palm for training and the other eight for testing.

(3) KTH-TIPS texture database: KTH-TIPS[56] is a texture database for material categorization. It contains images of ten materials, e.g., Sponge, Corduroy, Brown bread. These

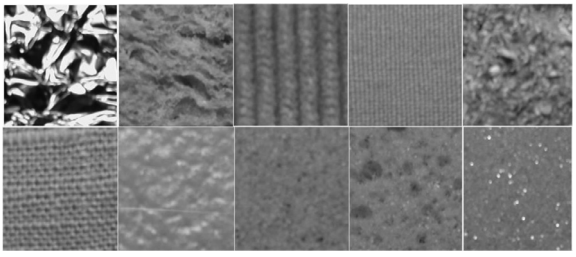

Fig. 3. Some cropped texture images of the KTH-TIPS database.

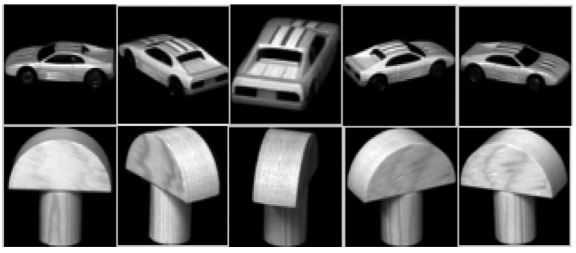

Fig. 4. Some cropped object images of the COIL-20 object database.

images were collected at nine different scales crossing 2 octaves. There are nine images in a combination of 3 lighting and 3 poses conditions for each scale. All these variations on lighting, pose, and scale make it large challenging. Most of these images are cropped to 200×200 size. Some cropped images in this database are shown in Fig. 3. In our experiments, we randomly choose T_s (T_s = 20, 30, 40) images of each subject for training, and the remaining are used for testing.

(4) COIL-20 object database: COIL-20 [58] is an object database. There are 20 different objects, and each object has 72 images. These images were collected at pose intervals of five degrees (i.e., 72 different poses per object). All images are cropped and resized to 64×64. Some cropped images are shown in Fig. 4. In our experiments, we randomly choose T_s (T_s=10, 20, 30) images of each subject for training, and the remaining are used for testing.

(5) CIFAR-10 tiny images database: CIFAR-10 [59] is a tiny images database. It contains images of ten classes, i.e., Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, and Truck. There are 6000 images per class and totally 60,000 images in 10 classes. Each image is cropped to 32×32 size. In our experiments, we use 50,000 images for training, and the rest 10,000 images for testing.

3.2. Parameters evaluation

In this subsection, we discuss the effect of the different parameters to the proposed SADPL model. We take the experiments on the five databases to test different settings of parameters λ_1 and λ_2 in the presented SADPL. In particular, we take the values of λ_1 and λ_2 changing by {1e+01, 1e+02, 1e+03, 1e+04, 1e+05, 1e+06} and {0.001, 0.01, 0.1, 1, 10, 100}, respectively. The classification results of the presented SADPL model are shown in Fig. 5. According to the experimental results, we can see that the performance of the presented SADPL is not changeless to the variations of parameters. It could achieve the best performance when the values of parameter pair (λ_1, λ_2) are set as (100,0.001), (10000,0.01), (1000,0.01), (100000,0.1) and (10,0.01) for FRGC, 2D plus 3D, KTH-TIPS, COIL-20 and CIFAR-10 databases, respectively.

Table 1
Classification accuracies for different approaches on the FRGC face database (Mean ± STD (%)).

Method	Controlled	Uncontrolled
PCA	90.78 ± 5.79	57.14 ± 3.26
LDA	91.96 ± 2.91	87.74 ± 1.28
LPP	99.27 ± 0.59	86.73 ± 1.31
LSDA	99.29 ± 0.59	86.89 ± 1.25
SRDA	99.09 ± 0.64	86.65 ± 1.27
RILDA	98.53 ± 0.91	82.20 ± 1.18
SRRS	94.91 ± 0.47	60.93 ± 2.32
LLDA	92.91 ± 0.59	72.98 ± 2.18
SADPL	**99.33 ± 0.54**	**88.87 ± 1.15**

3.3. Convergence analysis

As mentioned in Section 2.4, the value of the objective function $J(\mathbf{P})$ monotonically decreases along with the iteration, and it will converge to the global optimum. Here we report the convergence rate of our method with respect to the objective function value. Fig. 6(a) and (b) show the convergence rate of our SADPL on the FRGC and COIL-20 databases. It can be seen that our SADPL converges very fast. The objective function value achieves stable after about 8–10 iterations.

3.4. Comparisons between SADPL and other approaches

In this subsection, comparisons between SADPL and the related approaches on the five databases are given.

(1) Comparisons on the FRGC face database: The experiments are independently repeated 10 times and the average classification accuracies and the standard deviations are calculated and reported. The classification results are summarized in Table 1. For the controlled illumination variation subset, we can see that this is a relatively easy classification task, four algorithms can achieve high accuracies, over 99%. But the unsupervised PCA only can correctly classify 90.78% testing samples. Note that the accuracies of LPP, LSDA and SRDA are 99.27%, 99.29% and 99.09%, respectively. However, the proposed SADPL algorithm obtains accuracy of 99.33%, which outperforms the other methods. For uncontrolled subset, all these algorithms decrease their accuracies by different degrees due to the more complex illumination variations. But the proposed SADPL still achieves the best performance compared with other algorithms.
In addition, the Friedman's test [60] is used to further demonstrate the significant difference of the classification performance between the presented SADPL and the compared approaches. It should be mentioned that the Friedman's test is performed by evaluating the hypothesis that the column impacts are all the same vs they are not all the same. We combine the results under illumination controlled subset and uncontrolled subset together to perform the Friedman's test. Eight columns of the data matrix which are input to the Friedman's test procedure, are constituted from the eight approaches shown in Table 1. After performing the Friedman's test, we obtain that the p-value is 2.35e−21, which indicates the significant difference among these approaches. Besides the p-value, we utilize the result output by the Friedman's test and construct an interactive graph with multiple comparison intervals to further show the significance, which is displayed in Fig. 7. Note that the performance of two methods are not significantly different if their intervals overlap. Conversely, they are significantly different if their intervals are non-overlap. From the interactive

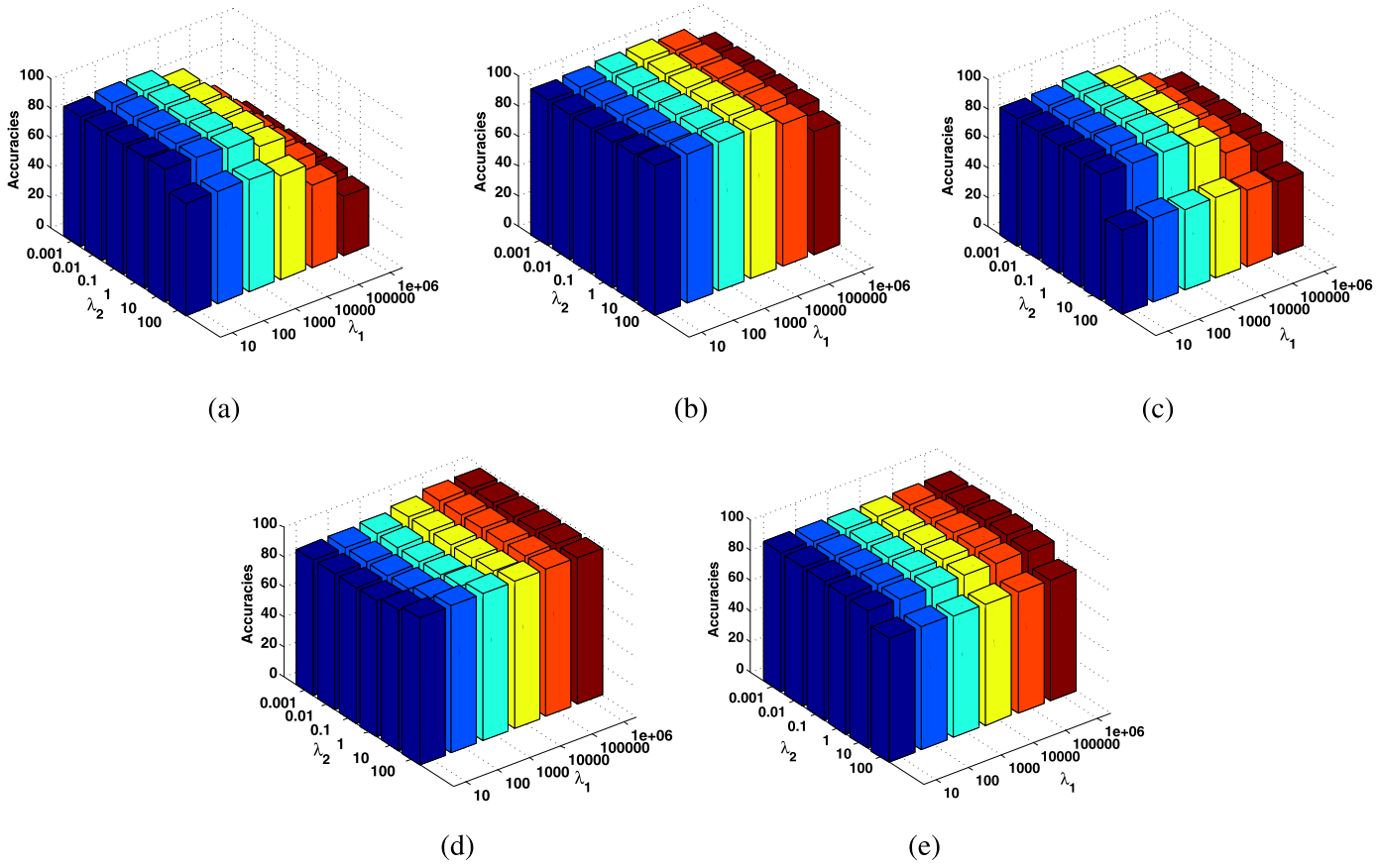

Fig. 5. Classification results of the proposed SADPL with different parameter settings on the different databases. (a) FRGC face database. (b) 2D plus 3D palmprint database. (c) KTH-TIPS texture database. (d) COIL-20 object database. (e) CIFAR-10 tiny images database.

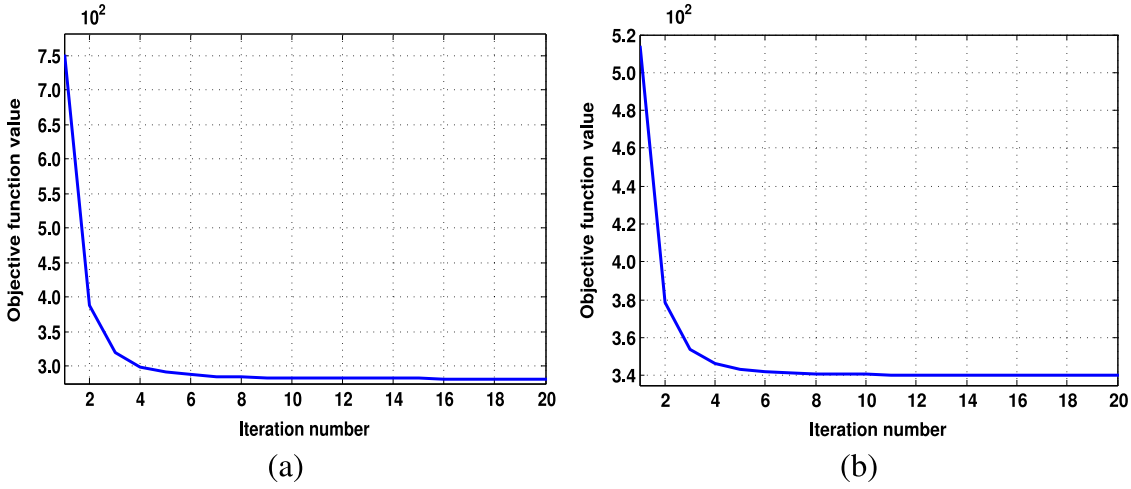

Fig. 6. Convergence rate of our SADPL with respect to the objective function value on the different databases. (a) FRGC face database. (b) COIL-20 object database.

graph shown in Fig. 7, we can see that six approaches have mean column ranks significantly different from our SADPL.

(2) Comparisons on the 2D plus 3D palmprint database: The experiments are independently repeated 20 times and the average classification accuracies and the standard deviations are calculated and reported. The results are shown in Table 2. As we can see, except for LLDA and RILDA, each of the other algorithms can obtain the accuracies higher than 96%. The best performance is achieved by our SADPL and the classification accuracies are 99.28% and 99.66%, respectively. It indicates that our feature selection model is effective to select the informative features.

In addition, we also perform the Friedman's test to further evaluate the significant difference between the presented SADPL and other algorithms on the palmprint database. Being similar to the test protocol of face image database, eight columns of the data matrix which are input to the Friedman's test procedure, are constituted from the eight approaches shown in Table 2. After performing the Friedman's test, we get that the p-value is 4.15e−62. It demonstrates the significant difference among the classification results obtained by these algorithms. The multicomparison results are shown in Fig. 8, in which seven approaches have mean column ranks significantly different from SADPL.

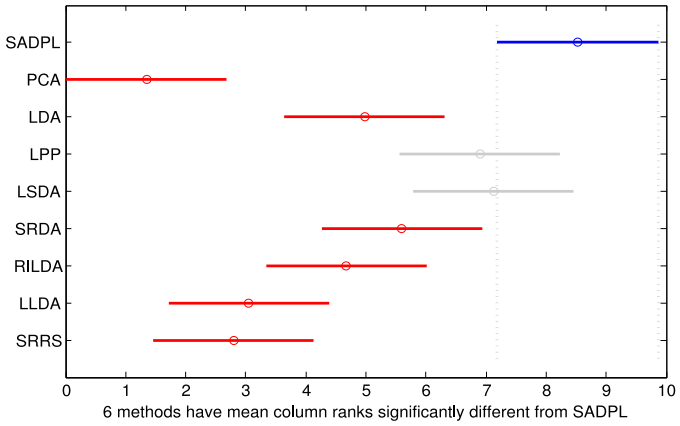

Fig. 7. Significant difference between the presented SADPL and the compared approaches on the FRGC face database.

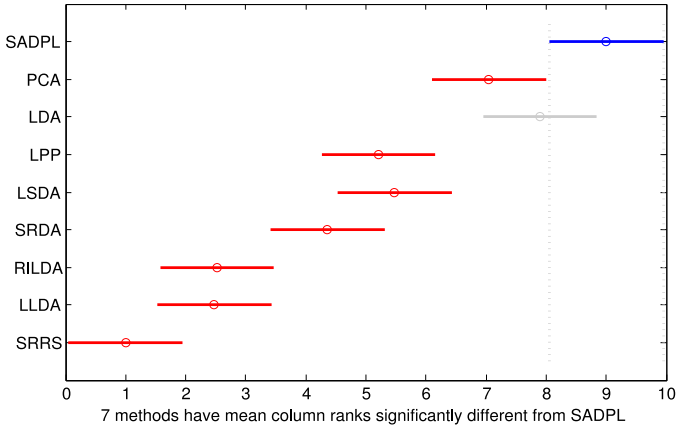

Fig. 8. Significant difference between the presented SADPL and the compared approaches on the 2D plus 3D palmprint database.

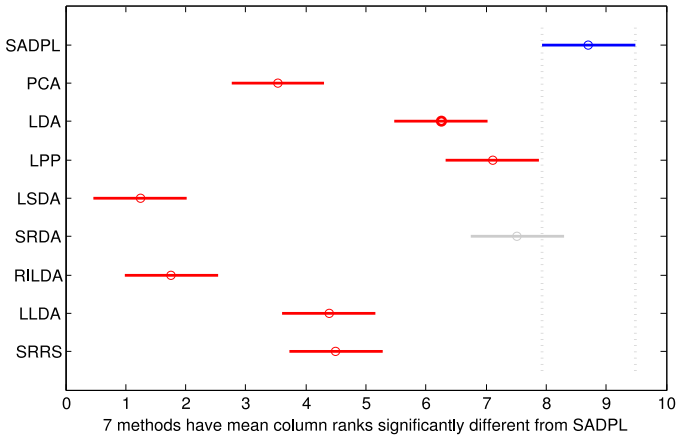

Fig. 9. Significant difference between the presented SADPL and the compared approaches on the KTH-TIPS texture database.

(3) Comparisons on the KTH-TIPS texture database: As mentioned above, the KTH-TIPS is an extremely challenging texture database. For all these compared methods, considering that LBP [61] is effective to texture classification, we first extract LBP features on this database, and then project the corresponding subspaces. The experiments are independently repeated 20 times and the average classification accuracies are calculated and reported. The classification results are presented in Table 3. When twenty images are randomly selected from each subject for training, the unsupervised PCA only can correctly classify 75.33% testing samples. The supervised algorithms such as LDA, LPP and SRDA attain better results than PCA, and their accuracies are 80.83%, 81.92% and 82.78%, respectively. Almost all these algorithms can obtain better results as the training size of each class increases. However, the performance of LSDA is gradually declining. The main reason may be that the local geometrical information could not be extracted correctly. We can see that the presented SADPL obtains the best performance, which is reflected by its accuracies, and the detailed results for the three different training settings are 84.37%, 88.62% and 90.41%, respectively.

We perform the Friedman's test and get the p-value of 9.97e−88. The multiple comparison results are shown in Fig. 9, in which seven groups have mean column ranks significantly different from SADPL. Therefore, the statistical significance of the differences between SADPL and the compared algorithms are shown.

(4) Comparisons on the COIL-20 object database: Following the recognition protocol in Section 3.1, the experiments are independently repeated 20 times and the average classification accuracies are calculated and reported. The results are shown in Table 4. We can see that the performance of the SADPL model are better than the compared approaches, and the accuracies are 94.77%, 98.48% and 99.32% respectively.

Table 2
Classification accuracies for different approaches on the 2D plus 3D palmprint database (Mean ± STD (%)).

Method	Session 1	Session 2
PCA	97.92 ± 0.67	98.65 ± 0.48
LDA	98.05 ± 0.64	99.14 ± 0.28
LPP	96.66 ± 0.70	97.59 ± 0.49
LSDA	96.67 ± 0.73	97.93 ± 0.48
SRDA	96.60 ± 0.73	97.75 ± 0.50
RILDA	94.02 ± 1.14	95.34 ± 0.85
SRRS	89.29 ± 2.66	89.71 ± 2.54
LLDA	94.04 ± 2.06	94.25 ± 2.12
SADPL	**99.28 ± 0.29**	**99.66 ± 0.17**

Table 3
Classification accuracies for different approaches on the KTH-TIPS texture database (Mean ± STD (%)).

Method	$T_s = 20$	$T_s = 30$	$T_s = 40$
PCA	75.33 ± 3.27	79.92 ± 2.62	82.88 ± 2.71
LDA	80.83 ± 1.70	84.83 ± 2.69	87.47 ± 2.59
LPP	81.92 ± 2.24	86.08 ± 2.32	88.08 ± 2.11
LSDA	45.67 ± 7.59	42.12 ± 2.12	33.08 ± 11.82
SRDA	82.78 ± 1.67	86.25 ± 2.43	88.14 ± 2.36
RILDA	48.43 ± 14.33	57.33 ± 13.31	58.75 ± 11.14
SRRS	75.91 ± 2.98	80.92 ± 2.74	84.57 ± 2.66
LLDA	75.80 ± 3.15	79.87 ± 2.57	85.61 ± 2.71
SADPL	**84.37 ± 2.36**	**88.62 ± 2.15**	**90.41 ± 2.26**

Table 4
Classificationaccuracies for different approaches onthe COIL-20 object database (Mean ± STD (%)).

Method	$T_s = 10$	$T_s = 20$	$T_s = 30$
PCA	91.70 ± 3.16	96.03 ± 1.68	97.88 ± 1.21
LDA	90.79 ± 3.24	94.29 ± 1.50	95.76 ± 1.24
LPP	87.28 ± 3.31	39.82 ± 43.76	27.38 ± 39.78
LSDA	79.75 ± 3.70	65.76 ± 16.13	64.83 ± 15.06
SRDA	86.96 ± 3.46	91.21 ± 1.74	93.13 ± 1.50
RILDA	70.53 ± 6.73	76.47 ± 10.31	78.80 ± 11.13
SRRS	93.64 ± 3.22	97.70 ± 1.70	99.03 ± 0.93
LLDA	87.63 ± 2.88	92.38 ± 1.82	93.88 ± 1.08
SADPL	**94.77 ± 3.34**	**98.48 ± 1.17**	**99.32 ± 0.66**

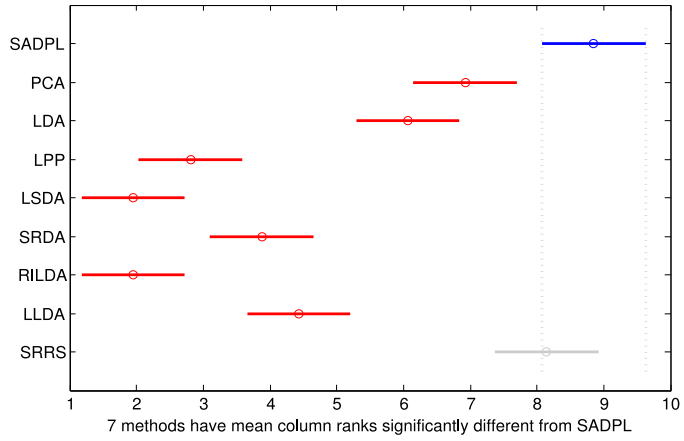

Fig. 10. Significant difference between the presented SADPL and the compared approaches on the COIL-20 object database.

Table 5
Classification accuracies for different approaches on the CIFAR-10 database (%).

PCA	LDA	LPP	LSDA	SRDA	RILDA	SRRS	LLDA	SADPL
91.22	90.63	88.65	90.59	90.57	89.78	90.23	89.58	91.25

Table 6
Classification accuracies of SADPL under different regularization terms (%).

Database	FRGC	2D+3D	KTH-TIPS	COIL-20	CIFAR-10
Adding F-norm	99.33	99.28	88.62	99.32	91.25
Without F-norm	99.07	96.57	86.37	93.14	91.08
Using l_1-norm	99.27	99.28	88.59	99.31	91.20

In addition, experimental results of three subsets are combined together to perform the Friedman's test and get the p-value of 2.98e−89. Hence, it validates a significant difference among the classification results obtained by the compared subspace learning algorithms. We also use the result output by the Friedman's test to conduct a multiple comparison test, and show the multicomparison results in Fig. 10. One can see that the performance of SADPL is significantly different from other algorithms.

(5) Comparisons on the CIFAR-10 database: For all these compared methods, considering that the CIFAR-10 is an extremely challenging tiny images database, we utilize training set to fine-tune all layers of the pre-trained AlexNet [62] model by continuing the back propagation. Then we remove the last fully connected layer and treat the rest of the network as fixed feature extractor. The dimension of the extracted features is 4096. Finally, we project the features into the corresponding subspaces. The classification accuracies are reported in Table 5. One can see that all these compared methods obtain the high accuracies around 90%, and our SADPL gets the best performance.

3.5. Discussion

In this section, we discuss the effect of the different regularization terms to our SADPL algorithm.

(1) Impact analysis by adding F-norm vs. without F-norm. To analyze the effect of SADPL by adding F-norm in model (9), we conduct the compared experiments on the five databases. The performance comparison results are shown in Table 6. Here *without F-norm* means we do not add F-norm in model (9). From the second and third rows in Table 6, we can see that the performance of SADPL is better by adding the F-norm than without F-norm, especially on the 2D+3D palmprint database, KTH-TIPS texture database and COIL-20 object database. SADPL adding F-norm obtains the accuracies of 99.28%, 88.62% and 99.32% on the three databases, whereas the accuracies of SADPL without F-norm are 96.57%, 86.37% and 93.14%, respectively. It indicates that the joint using of F-norm and $l_{2,1}$-norm is effective to select the informative features.

(2) $l_{2,1}$-norm vs. l_1-norm. We further discuss the effect of our SADPL by using $l_{2,1}$-norm to replace l_1-norm. The corresponding results are shown in Table 6. Here the results of the second row in Table 6 are obtained by the joint using of F-norm and $l_{2,1}$-norm, and the results of the fourth row are obtained by the joint using of F-norm and l_1-norm. One can see that they have the similar performance. However, the computational cost of SADPL using l_1-norm is much higher than using $l_{2,1}$-norm. For instance, on our computer with i7-6700K CPU and 16.0G memory, the running time of SADPL using $l_{2,1}$-norm is 18.5 s and 32.1 s on the FRGC and 2D+3D databases, while the corresponding running time of SADPL using l_1-norm reaches up to 286.5 s and 610.6 s, respectively.

4. Conclusion

We have presented a supervised sparse discriminant projection learning algorithm which preforms subspace learning and feature selection simultaneously. The proposed method learns the projection matrix based on a new objective function rather than the traditional Fisher criterion, avoiding the matrix singularity problem and also can selecting the informative features. Furthermore, we have presented an effective optimization approach to deal with the new objective function and proved that the presented optimization algorithm is convergent. Extensive experiments have shown that the proposed approach could achieve competitive performance on various image classification tasks. It is clear that the proposed method is a vector-based SADPL. In the future, we will extend the vector-based SADPL to a matrix-based formulation.

References

[1] W.C. Lin, C.F. Tsai, Z.Y. Chen, S.W. Ke, Keypoint selection for efficient bag-of-words feature generation and effective image classification, Inf. Sci. 329 (2016) (2016) 33–51.
[2] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.
[3] J. Ye, R. Janardan, Q. Li, H. Park, Feature reduction via generalized uncorrelated linear discriminant analysis, IEEE Trans. Knowl. Data Eng. 18 (10) (2006) 1312–1322.
[4] N. Zhou, Y. Xu, H. Cheng, J. Fang, W. Pedrycz, Global and local structure preserving sparse subspace learning: an iterative approach to unsupervised feature selection, Pattern Recognit. 53 (5) (2016) 87–101.
[5] J.L. Tang, H. Liu, An unsupervised feature selection framework for social media data, IEEE Trans. Knowl. Data Eng. 26 (12) (2014) 2914–2927.
[6] M. Belkin, P. Niyogi, Semi-supervised learning on riemannian manifolds: theoretical advances in data clustering, Mach. Learn. 56 (1–3) (2004) 209–239.
[7] C.H. Nguyen, H. Mamitsuka, Discriminative graph embedding for label propagation., IEEE Trans. Neural Netw. 22 (9) (2011) 1395–1405.
[8] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.
[9] R.A. Fisher, The statistical utilization of multiple measurements, Ann. Eugen. 8 (4) (1938) 376–386.
[10] C.R. Rao, The utilization of multiple measurements in problems of biological classification, J. R. Stat. Soc. Ser. B 10 (2) (1948) 159–203.
[11] P.N. Belhumeur, J. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[12] L. Zhang, M. Yang, Z. Feng, D. Zhang, On the dimensionality reduction for sparse representation based face recognition, in: International Conference on Pattern Recognition, 2010, pp. 1237–1240.
[13] W.H. Yang, D.Q. Dai, Two-dimensional maximum margin feature extraction for face recognition, IEEE Trans. Syst. Man Cybern. Part B 39 (4) (2009) 1002–1012.
[14] Q. Gao, L. Zhang, D. Zhang, Sequential row–column independent component analysis for face recognition, Neurocomputing 72 (4) (2009) 1152–1159.

[15] 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.
[16] H. Wang, X. Lu, Z. Hu, W. Zheng, Fisher discriminant analysis with l_1-norm, IEEE Trans. Cybern. 44 (6) (2014) 828–842.
[17] D.Q. Dai, P.C. Yuen, Face recognition by regularized discriminant analysis, IEEE Trans. Syst. Man Cybern. Part B 37 (4) (2007) 1080–1085.
[18] T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, Ann. Stat. 23 (1) (1995) 73–102.
[19] T. Hastie, R. Tibshirani, A. Buja, Flexible discriminant analysis by optimal scoring, J. Am. Stat. Assoc. 89 (428) (1994) 1255–1270.
[20] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Netw. 17 (1) (2006) 157–165.
[21] D. Cai, X.F. He, J.W. Han, Srda: an efficient algorithm for large-scale discriminant analysis, IEEE Trans. Knowl. Data Eng. 20 (1) (2008) 1–12.
[22] W.S. Zheng, J.H. Lai, P.C. Yuen, S.Z. Li, Perturbation lda: learning the difference between the class empirical mean and its expectation, Pattern Recognit. 42 (5) (2009) 764–779.
[23] D. Cai, X.F. He, J.W. Han, Semi-supervised discriminant analysis, in: IEEE International Conference on Computer Vision, 2007, pp. 1–7.
[24] Z. Lai, Y. Xu, J. Yang, L. Shen, D. Zhang, Rotational invariant dimensionality reduction algorithms, IEEE Trans. Cybern. 47 (11) (2017) 3733–3746.
[25] D. Cai, X.F. He, K. Zhou, J. Han, H. Bao, Locality sensitive discriminant analysis., in: International Joint Conference on Artificial Intelligence, 2007, pp. 708–713.
[26] X.F. He, X. Niyogi, Locality preserving projections, in: International Conference on Neural Information Processing Systems, 2003, pp. 153–160.
[27] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, J. Mach. Learn. Res. 8 (5) (2007) 1027–1061.
[28] Z. Fan, Y. Xu, D. Zhang, Local linear discriminant analysis framework using sample neighbors, IEEE Trans. Neural Netw. 22 (7) (2011) 1119–1132.
[29] X. Shi, Y. Yang, Z. Guo, Z. Lai, Face recognition by sparse discriminant analysis via joint $l_{2,1}$-norm minimization, Pattern Recognit. 47 (7) (2014) 2447–2453.
[30] N. Kwak, Principal component analysis based on l_1-norm maximization, IEEE Trans. Pattern Anal. Mach. Intell. 30 (9) (2008) 1672–1680.
[31] C. Ding, D. Zhou, X. He, H. Zha, R 1-pca: rotational invariant l_1-norm principal component analysis for robust subspace factorization, in: International Conference on Machine learning, ACM, 2006, pp. 281–288.
[32] F. Zhong, J. Zhang, Linear discriminant analysis based on l_1-norm maximization, IEEE Trans. Image Process. 22 (8) (2013) 3018–3027.
[33] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.
[34] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[35] H. Zou, T. Hastie, Regression shrinkage and selection via the elastic net, with applications to microarrays, J. R. Stat. Soc. Ser. B 67 (2) (2003) 301–320.
[36] L. Clemmensen, T. Hastie, D. Witten, B. Ersbøll, Sparse discriminant analysis, Technometrics 53 (4) (2012) 406–413.
[37] Z.L. Zheng, Sparse locality preserving embedding, in: International Congress on Image and Signal Processing, 2009, pp. 1–5.
[38] Z. Qiao, L. Zhou, J.Z. Huang, Sparse linear discriminant analysis with applications to high dimensional low sample size data, Int. J. Appl. Math. 39 (1) (2009) 48–60.
[39] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1) (1996) 267–288.
[40] Z. Lai, W.K. Wong, Z. Jin, J. Yang, Y. Xu, Sparse approximation to the eigensubspace for discrimination, IEEE Trans. Neural Netw. Learn. Syst. 23 (12) (2012) 1948–1960.
[41] X. Liu, L. Wang, J. Zhang, J. Yin, H. Liu, Global and local structure preservation for feature selection, IEEE Trans. Neural Netw. Learn. Syst. 25 (6) (2014) 1083–1095.
[42] R. He, T. Tan, L. Wang, W.-S. Zheng, $l_{2,1}$ regularized correntropy for robust feature selection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2504–2511.
[43] H. Zhao, Z. Wang, F. Nie, A new formulation of linear discriminant analysis for robust dimensionality reduction, IEEE Trans. Knowl. Data Eng. (2018), doi:10.1109/TKDE.2018.2842023.
[44] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning, in: International Joint Conference on Artificial Intelligence, 2011, pp. 1589–1594.
[45] F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, in: Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.
[46] J.P. Ye, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, J. Mach. Learn. Res. 6 (4) (2005) 483–502.
[47] X.S. Zhuang, D.Q. Dai, Inverse fisher discriminate criteria for small sample size problem and its application to face recognition, Pattern Recognit. 38 (11) (2005) 2192–2194.
[48] I. Gaynanova, J.G. Booth, M.T. Wells, Simultaneous sparse estimation of canonical vectors in the $p \gg n$ setting, J. Am. Stat. Assoc. 111 (514) (2016) 696–706.
[49] J. Gui, Z. Sun, S. Ji, D. Tao, T. Tan, Feature selection based on structured sparsity: a comprehensive study, IEEE Trans. Neural Netw. Learn. Syst. 28 (7) (2017) 1490–1507.
[50] A. Evgeniou, M. Pontil, Multi-task feature learning, in: Advances in Neural Information Processing Systems, 2007, pp. 41–48.
[51] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, X. Zhou, Semisupervised feature selection via spline regression for video semantic recognition, IEEE Trans. Neural Netw. Learn. Syst. 26 (2) (2015) 252–264.
[52] M. Yang, L. Zhang, X. Feng, D. Zhang, Sparse representation based fisher discrimination dictionary learning for image classification, Int. J. Comput. Vis. 109 (3) (2014) 209–232.
[53] Q. Gu, J. Han, Towards feature selection in network, in: International Conference on Information and Knowledge Management, 2011, pp. 1175–1184.
[54] S. Li, Y. Fu, Learning robust and discriminative subspace with low-rank constraints, IEEE Trans. Neural Netw. Learn. Syst. 27 (11) (2016) 2160–2173.
[55] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2005, pp. 947–954.
[56] B. Caputo, E. Hayman, P. Mallikarjuna, Class-specific material categorisation, in: IEEE International Conference on Computer Vision, 2005, pp. 1597–1604.
[57] Hk-polyu 2d+3d palmprint database, http://www.comp.polyu.edu.hk/~biometrics/2D_3D_Palmprint.htm.
[58] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, in: International Conference on Neural Information Processing Systems, 2009, pp. 2223–2231.
[59] A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features from Tiny Images, Technical Report, University of Toronto, 2009.
[60] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30.
[61] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[62] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

Yu-Feng Yu received the Ph.D. degree in statistics from Sun Yat-Sen University, Guangzhou, China, in 2017. He is currently an assistant professor in the Department of statistics, Guangzhou University. From 2016 to 2017, he was a visiting scholar in the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA. From 2017 to 2018, he was a senior research associate in the Department of Electronic Engineering, City University of Hong Kong. His research interests include image processing, statistical optimization, pattern recognition and machine learning.

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

Min Jiang received the B.S. degree and M.S. degree in electrical engineering from China University of Mining and Technology, Beijing, China. She is currently pursuing the Ph.D. degree with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, WV, USA. Her research area includes computer vision, machine learning and signal processing, in particular, BMI analysis from human visual appearance and astronomical signal denoising.

Man-Yu Sun received the B.Sc. degree in mathematics from Wuhan University, Wuhan, China, in 2014. He is currently pursuing the Ph.D. degree in Sun Yat-Sen University, Guangzhou, China. His research interests include image processing and computer vision.

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. From 1998 to 1999, he was an Alexander von Humboldt Research Fellow with Free University, Berlin, Germany. He is currently a Professor with the School of Mathematics, Sun Yat-sen University. He has authored or coauthored over 100 refereed technical papers. His current research interests include image processing, wavelet analysis, face recognition, and bioinformatics. Dr. Dai was a recipient of the Outstanding Research Achievements in Mathematics Award from the International Society for Analysis, Applications, and Computation, Fukuoka, Japan, in 1999.

Guodong Guo received the B.E. degree in automation from Tsinghua University, Beijing, China, the Ph.D. degree in computer science from the University of Wisconsin, Madison, WI, USA. He is currently the Deputy Head of the Institute of Deep Learning, Baidu Research, and also an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), USA. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; and Microsoft Research, Beijing, China. He authored a book, “Face, Expression, and Iris Recognition Using Learning-based Approaches” (2008), co-edited two books, “Support Vector Machines Applications” (2014) and “Mobile Biometrics” (2017), and published over 100 technical papers. He is an Associate Editor of IEEE Transactions on Affective Computing, and Journal of Visual Communication and Image Representation, and serves on the editorial board of IET Biometrics. His research interests include computer vision, biometrics, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, Outstanding Researcher (2017–2018, 2013–2014) at CEMR, WVU, and New Researcher of the Year (2010–2011) at CEMR, WVU. He was selected the “People's Hero of the Week” by BSJB under Minority Media and Telecommunications Council (MMTC) in 2013. Two of his papers were selected as “The Best of FG'13” and “The Best of FG'15”, respectively.