Two Dimensional Slow Feature Discriminant Analysis via $L_{2,1}$ Norm Minimization for Feature Extraction

Xingjian Gu[1,3], Xiangbo Shu[2], Shougang Ren[1] and Huanliang Xu[1]
[1] College of Information Science and Technology, Nanjing Agricultural University
Nanjiang, 210095 China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, 210094, China
[3]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University)
Fuzhou, 350121, China
rensg@njau.edu.cn
*Corresponding author: Shougang Ren

Received September 12, 2017; revised January 30, 2018; accepted February 6, 2018; published July 31, 2018

Abstract

Slow Feature Discriminant Analysis (SFDA) is a supervised feature extraction method inspired by biological mechanism. In this paper, a novel method called Two Dimensional Slow Feature Discriminant Analysis via $L_{2,1}$ norm minimization (2DSFDA-$L_{2,1}$) is proposed. 2DSFDA-$L_{2,1}$ integrates $L_{2,1}$ norm regularization and 2D statically uncorrelated constraint to extract discriminant feature. First, $L_{2,1}$ norm regularization can promote the projection matrix row-sparsity, which makes the feature selection and subspace learning simultaneously. Second, uncorrelated features of minimum redundancy are effective for classification. We define 2D statistically uncorrelated model that each row (or column) are independent. Third, we provide a feasible solution by transforming the proposed $L_{2,1}$ nonlinear model into a linear regression type. Additionally, 2DSFDA-$L_{2,1}$ is extended to a bilateral projection version called BSFDA-$L_{2,1}$. The advantage of BSFDA-$L_{2,1}$ is that an image can be represented with much less coefficients. Experimental results on three face databases demonstrate that the proposed 2DSFDA-$L_{2,1}$/BSFDA-$L_{2,1}$ can obtain competitive performance.

Keywords: sparse projection, slow feature discriminant analysis, feature extraction, $L_{2,1}$ norm

http://doi.org/10.3837/tiis.2018.07.012 **ISSN : 1976-7277**

1. Introduction

Feature extraction has been playing an important role in the domains of image recognition[1], image retrieval [2,3,4,5] and image classification[6,7]. The goal of feature extraction is to seeks a set of meaningful low-dimensional representations of high dimensional data to simplify data analysis problems, such that the intrinsic structures of original high-dimensional data are revealed. Principal Component Analysis (PCA) [8] and Linear Discriminant Analysis (LDA) [9] are the most well-known classical linear techniques. However, linear feature extraction methods fail to discover the nonlinear structure of data. To reveal the underlying nonlinear structure, a family of manifold-based learning methods have been developed. The represented manifold-based learning methods include locality preserving projections (LPP) [10], neighborhood preserving embedding (NPE) [11], linear local tangent space alignment (LLTSA) [12] and so on. LPP, NPE and LLTSA emphasize the locality property and are suitable for feature extraction on nonlinear manifold. To enhance the performance of classification, there emerged several nonlinear manifold learning methods to extract discriminant feature [13,14,15]. Yu et al. [13] extended LPP to discriminant locality preserving projections (DLPP) method to improve the classification performance. Yan et al. [14] provided marginal fisher analysis (MFA) that combines locality and class label information to minimize the within-class compactness and maximize between-class separability. Chen et al. [15] proposed the local discriminant embedding (LDE) to maximize the margin of between classes.

Recently, the idea inspired by biological mechanism is used to design some feature extraction methods. Wiskott et al. [16] proposed slow feature analysis (SFA) to extract slowly varying features and invariance from vectorial temporal signals based on temporal slowness principle. SFA has been performed for many applications in the field of computational neuroscience [17, 18, 19, 20]. For example SFA was initially developed for learning complex-cell receptive fields [19] and place cells in the hippocampus [20]. SFA turns out to be useful in pattern recognition and classification. Many researchers have successfully introduced the slowness principle into the applications of pattern recognition [21]. Zhang et al. [21] proposed a SFA framework to deal with the problem of human action recognition and obtained a well performance. However, in real applications, there are also numerous discrete data sets. In face recognition, the sample images have no obvious temporal structure as action images in video sequence. To deal with the discrete scenario, it is necessary to construct time series before implementation of SFA. The quality of constructed time series is key to the performance of SFA. In the literature [22], the authors utilized *k* nearest neighbor (KNN) criterion to construct time series and proposed a new framework of SFA to characterize the underlying structure of manifold for nonlinear dimensionality reduction. To get more accuracy of time series, the authors [23] proposed a supervised slow feature analysis based on consensus matrix to construct time series for face recognition. In order to get discriminant slow feature, they also proposed slow feature discriminant analysis (SFDA) for digital handwrite recognition [24], which minimizes withinclass temporal variation and maximizes between-class temporal variation simultaneously. Gu et al. [25] proposed an adaptive criterion to generate time series and derived an optimal slow discriminant feature subspace for classification.

Further more, previous works[26,27,28] have also demonstrated that reducing the correlation of extracted feature should contribute to improve the recognition performance. Because statistically correlated features contain redundancy, which will distort the distribution of the features. Jin et al. [26] proposed an uncorrelated linear discrimination analysis (ULDA) approach which maximizes Fisher criterion and simultaneously eliminates correlation between extracted features. To explore uncorrelated local information, Jing et al. [27] proposed a feature extraction approach named local uncorrelated discriminant transform (LUDT) for face recognition by adding a local uncorrelated constraints and calculated the optimal discriminant vectors. The drawback of the above mentioned uncorrelated methods (ULDA and LUDT) is that they all take an iterative manner, which will cost a long time to complete the iterative process.

Recently, two dimension based methods have been successfully used in feature extraction since it can exploit the spatial information. Yang et al. [29] proposed the two dimensional principal component analysis (2DPCA) which is based on 2D image matrix rather than 1D vector. 2DPCA and its variants [30, 31] have been widely applied in face recognition. Motivated by 2DPCA, two-dimensional linear discriminant analysis (2DLDA) [32] and two-dimensional locality preserving projection (2DLPP) [33, 34, 35] was proposed for extraction 2D-feature. 2DLPP can approximate the original images more accurately than LPP, and the learned subspace has smaller dimensionality. Zhang et al. [36] used 2D images rather than 1D vectors as an input feature, and they proposed a two-dimensional neighborhood preserving projection (2DNPP) for face recognition. However, these 2D matrix-based methods always use the Euclidean distance as a metric to classify data points.

Because the L_2 norm and Frobenius norm (F-norm) of a matrix are sensitive to noise in data and they take all the original features (i.e. pixels in image) equally and ignore the differences among them, the performance of the methods based on the two norms would be degraded. As the L_1 norm is robust to noise or outliers in data, many dimensionality reduction methods based on the L_1 norm have been proposed [37-38]. Zhao et al. [37] proposed a L_1-norm-based two-dimensional locality preserving projections (2DLPP-L_1) method to improve the robustness of 2DLPP against outliers and corruptions. Tang et al. [38] proposed a two-dimensional discriminant LPP method based on the L_1 norm (2D-DLPP-L_1). 2D-DLPP-L_1 preserves the spatial topology structures of images effectively.

In recent years, $L_{2,1}$-norm has caused wide research interests [39,40,41]. Nie et al. [39] proposed an effective and robust feature selection method via joint $L_{2,1}$ norms minimization, which can interpret features play an important role in discriminant analysis. Gu et al. [40] proposed a framework to select relevant features and learn transformation matrix simultaneously, which imposes $L_{2,1}$-norm on loss term. Lai et al. [41] proposed a sparse version of the 2D local discriminant projection (S2DLDP), which provides an intuitive, semantic and interpretable feature subspace for classification.

The above mentioned 2D methods only operate on image rows while ignoring the information behind the image columns. By joining both of rows and columns of image information, bilateral projection based 2DPCA (B2DPCA) [42,43] and 2DBPP two-dimensional bilinear preserving projections [44] are developed. They seek two projection matrices to extract row information and column information simultaneously.

In this paper, we propose a novel method called two dimensional slow feature discriminant analysis via $L_{2,1}$ norm minimization for feature extraction (2DSFDA-$L_{2,1}$). 2DSFDA-$L_{2,1}$ integrates the $L_{2,1}$ norm regularization term and extracts uncorrelated discriminant features for classification. Inspired by the works [31, 42, 45], the 2DSFDA-$L_{2,1}$ is further extended to bilateral version BSFDA-$L_{2,1}$. The main contributions of this paper are listed as follow:

(1) We propose a novel framework of sparse 2D slow feature discriminant analysis, which integrates the $L_{2,1}$ norm regularized term and 2D uncorrelated constraint. The proposed methods remove the redundancy of 2D features and learn a row-sparsity projective matrix to enhance the discriminant ability.

(2) As an extension of 2DSFDA-$L_{2,1}$, BSFDA-$L_{2,1}$ is developed, where left and right projection directions are calculated simultaneously. BSFDA-$L_{2,1}$ can represent an image with much less coefficients than 2DSFDA-$L_{2,1}$.

(3) By analyzing the proposed $L_{2,1}$ norm based model, we transform the nonlinear optimization problem into linear optimization problem and provide a feasible manner to get the row-sparsity projective matrix.

The rest of the paper is organized as follows. In section 2, we briefly review SFDA and 2DLDA. In section 3, we give the motivations of two dimensional sparse slow feature analysis (2DSSFDA) and describe it in detail. In section 4 , experiments with three standard face image databases are carried out to demonstrate the effectiveness of the proposed method. Finally, the conclusions are made in section 5.

2. Related Work

In this subsection, we review the details of slow feature discriminant analysis (SFDA) [24] used for discrete data that does not have an obvious temporal structure. We have a vector-based training sample set $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}\in R^{D\times n}$ belonging to c classes, where n is the number of training sample and D is the dimension of training sample. For each training sample $\mathbf{x}_i$, SFDA constructs within-class time series t_w^i and between-class time series t_b^i using neighboring information as follows.

$$t_w^i=\left\{\left(\mathbf{x}_i,\mathbf{x}_i^l\right)\right\},l=1,2,\cdots,k_1 \tag{1}$$

where $\mathbf{x}_i^l\in N_k\left(\mathbf{x}_i\right)$, $N_k\left(\mathbf{x}_i\right)$ is the k nearest neighbors of $\mathbf{x}_i$ and $(\mathbf{x}_i,\mathbf{x}_i^l)$ share the same class label.

$$t_b^i=\left\{\left(\mathbf{x}_i,\ \tilde{\mathbf{x}}_i^l\right)\right\},l=1,2,\cdots,k_2 \tag{2}$$

where $\mathbf{x}_i^l\in\tilde{N}_k\left(\mathbf{x}_i\right)$, $\tilde{N}_k\left(\mathbf{x}_i\right)$ is the k nearest neighbors of $\mathbf{x}_i$ and $(\mathbf{x}_i,\tilde{\mathbf{x}}_i^l)$ have different class labels.

Based on the constructed time series t_w^i and t_b^i , $i=1,2,\cdots,n$, the model of SFDA can be written as follows:

$$\min_{\mathbf{W}} \frac{\mathbf{W}^T \sum_{i=1}^{n} \sum_{l=1}^{k} \left(\mathbf{x}_i - \mathbf{x}_i^l\right)\left(\mathbf{x}_i - \mathbf{x}_i^l\right)^T \mathbf{W}}{\mathbf{W}^T \sum_{i=1}^{n} \sum_{l=1}^{k} \left(\mathbf{x}_i - \tilde{\mathbf{x}}_i^l\right)\left(\mathbf{x}_i - \tilde{\mathbf{x}}_i^l\right)^T \mathbf{W}} \tag{3}$$

where $W \in R^{D \times d}$ is the transform matrix, d is the dimension of low-dimensional feature subspace, $(\mathbf{x}_i, \mathbf{x}_i^l) \in t_w^i$ and $(\mathbf{x}_i, \tilde{\mathbf{x}}_i^l) \in t_b^i$. The solution to the optimization of model (3) can be solved by generalized eigenvalue problem.

3. Two-dimensional Slow Feature Discriminant Analysis via $L_{2,1}$Norm Minimization (2DSFDA-$L_{2,1}$)

3.1 Motivation

The motivation for 2DSFDA-$L_{2,1}$ arises from the following three main considerable aspects. First, images often contain much redundant information, and the discriminant information is not decided by all the pixels. Selecting useful pixels plays an important role in feature extraction. Recent study indicates that introducing the $L_{2,1}$ norm regularized term to the projects matrix can produce row-sparsity matrix and enhance the discriminant ability of extracted feature. Second, previous works have demonstrated that statically uncorrelated feature is of great importance for classification task. Here, we propose a new uncorrelated model in 2D case. Third, we offer a feasible solution by transforming the proposed $L_{2,1}$ norm based nonlinear model into a linear regression model.

Based on the aforementioned analysis, we propose a novel $L_{2,1}$ norm based 2D slow feature discriminant analysis framework, which appends the sparseness and uncorrelated constraint into the 2D slow feature discriminant analysis to enhance the performance of classification task.

3.2 Two-dimensional Slow Feature Discriminant Analysis (2DSFDA)

Inspired by 2DLDA [32], we first extend SFDA into a 2D version. The goal of 2DSFDA is to learn a projective matrix $\mathbf{W}$ to minimize the 2D image within-class slowness scatter and between-class fastness scatter. Similar to SFDA, for each training image matrix $\mathbf{X}_i$, the within-class time series t_w^i and between-class time series t_b^i of 2DSFDA can be defined as follows.

$$t_w^i = \left\{\left(\mathbf{X}_i, \mathbf{X}_i^l\right)\right\}, l = 1, 2, \cdots, k_1 \tag{4}$$

where $\mathbf{X}_i^l \in N_k\left(\mathbf{X}_i\right)$, $N_k\left(\mathbf{x}_i\right)$ is the k nearest neighbors of $\mathbf{X}_i$ and $(\mathbf{X}_i, \tilde{\mathbf{X}}_i^l)$ share the same class label.

$$t_b^i = \left\{\left(\mathbf{X}_i, \tilde{\mathbf{X}}_j^l\right)\right\}, l = 1, 2, \cdots, k_2 \tag{5}$$

where $\tilde{\mathbf{X}}_i^l \in \tilde{N}_k\left(\mathbf{X}_i\right)$, $\tilde{N}_k\left(\mathbf{X}_i\right)$ is the k nearest neighbors of $\mathbf{X}_i$ and $(\mathbf{X}_i, \tilde{\mathbf{X}}_i^l)$ have the different class labels.

Thus, the matrix-based within-class temporal variation J_w and matrix-based between-class temporal variation J_b can be defined as follows:

$$\begin{aligned} J_w(\mathbf{W}) &= \sum_{i=1}^{n}\sum_{l=1}^{k_1}\left\|\mathbf{W}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\right\|_2^2 \\ &= \sum_{i=1}^{n}\sum_{l=1}^{k_1} tr\left(\mathbf{W}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\left(\mathbf{X}_i-\mathbf{X}_i^l\right)^T\mathbf{W}\right) \\ &= tr\left(\mathbf{W}^T\sum_{i=1}^{n}\sum_{l=1}^{k_1}\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\left(\mathbf{X}_i-\mathbf{X}_i^l\right)^T\mathbf{W}\right) \end{aligned} \tag{6}$$

$$\begin{aligned} J_b(\mathbf{W}) &= \sum_{i=1}^{n}\sum_{l=1}^{k_2}\left\|\mathbf{W}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\right\|_2^2 \\ &= \sum_{i=1}^{n}\sum_{l=1}^{k_2} tr\left(\mathbf{W}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)^T\mathbf{W}\right) \\ &= tr\left(\mathbf{W}^T\sum_{i=1}^{n}\sum_{l=1}^{k_2}\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)^T\mathbf{W}\right) \end{aligned} \tag{7}$$

where $(\mathbf{X}_i,\mathbf{X}_i^l)\in t_w^i$ and $(\mathbf{X}_i,\tilde{\mathbf{X}}_i^l)\in t_b^i$. Thus, the objective of 2DSFDA can be written as follows:

$$\min_{\mathbf{W}}\frac{\mathbf{W}^T\sum_{i=1}^{n}\sum_{l=1}^{k_1}\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\left(\mathbf{X}_i-\mathbf{X}_i^l\right)^T\mathbf{W}}{\mathbf{W}^T\sum_{i=1}^{n}\sum_{l=1}^{k_2}\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)^T\mathbf{W}} \tag{8}$$

3.3 Two-dimensional statistical uncorrelated constrain

In order to remove the redundancy for 2D feature matrix $\mathbf{Y}_i\in R^{m_1\times d}$, $i=1,2,\cdots,n$, we consider the following uncorrelated model:

$$\sum_{i=1}^{n}\left(\mathbf{y}_i^p-\tilde{\mathbf{y}}^p\right)^T\left(\mathbf{y}_i^q-\tilde{\mathbf{y}}^q\right)=0,\ p\neq q, q=1,2,\cdots,d \tag{9}$$

where $\mathbf{y}_i^p\in R^{m_1}$ and $\mathbf{y}_i^q\in R^{m_1}$ are considered as two random column vector variables. Furthermore, we could normalize the projective matrix $\mathbf{W}$ to satisfy:

$$\sum_{i=1}^{n}\left(\mathbf{y}_i^p-\tilde{\mathbf{y}}^p\right)^T\left(\mathbf{y}_i^p-\tilde{\mathbf{y}}^p\right)=1,\ p=1,2,\cdots,d \tag{10}$$

Thus, the matrix form of Equation (9) can be written as follows:

$$
\begin{aligned}
&\sum_{i=1}^{n}\begin{pmatrix}\left(\mathbf{y}_i^1-\tilde{\mathbf{y}}^1\right)^T\\ \left(\mathbf{y}_i^2-\tilde{\mathbf{y}}^2\right)^T\\ \vdots\\ \left(\mathbf{y}_i^d-\tilde{\mathbf{y}}^d\right)^T\end{pmatrix}\begin{pmatrix}\left(\mathbf{y}_i^1-\tilde{\mathbf{y}}^1\right) & \left(\mathbf{y}_i^2-\tilde{\mathbf{y}}^2\right) & \cdots & \left(\mathbf{y}_i^d-\tilde{\mathbf{y}}^d\right)\end{pmatrix}\\
&=\begin{pmatrix}\sum_{i=1}^{n}\left(\mathbf{y}_i^1-\tilde{\mathbf{y}}^1\right)^T\left(\mathbf{y}_i^1-\tilde{\mathbf{y}}^1\right) & \cdots & \sum_{i=1}^{n}\left(\mathbf{y}_i^1-\tilde{\mathbf{y}}^1\right)^T\left(\mathbf{y}_i^d-\tilde{\mathbf{y}}^d\right)\\ \vdots & \ddots & \vdots\\ \sum_{i=1}^{n}\left(\mathbf{y}_i^d-\tilde{\mathbf{y}}^d\right)^T\left(\mathbf{y}_i^1-\tilde{\mathbf{y}}^1\right) & \cdots & \sum_{i=1}^{n}\left(\mathbf{y}_i^d-\tilde{\mathbf{y}}^d\right)^T\left(\mathbf{y}_i^d-\tilde{\mathbf{y}}^d\right)\end{pmatrix}\\
&=\sum_{i=1}^{n}\left(\mathbf{Y}_i-\tilde{\mathbf{Y}}\right)^T\left(\mathbf{Y}_i-\tilde{\mathbf{Y}}\right)\\
&=\sum_{i=1}^{n}\mathbf{W}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{W}\\
&=\mathbf{W}^T\mathbf{S}_t\mathbf{W}\\
&=\mathbf{I}_d
\end{aligned}
\tag{11}
$$

where $\mathbf{S}_t=\sum_{i=1}^{n}\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)$ is the total scatter matrix, $I_d\in R^{d\times d}$ is an identity matrix, $\mathbf{Y}=\left[\mathbf{y}^1,\cdots,\mathbf{y}^d\right]$ and $\tilde{\mathbf{Y}}=\left[\tilde{\mathbf{y}}^1,\cdots,\tilde{\mathbf{y}}^d\right]$.

3.4 Objective function and Solution

With the preparation above, the model of 2DSFDA-$L_{2,1}$ can be written as follows:

$$
\begin{aligned}
&\min_{\mathbf{W}} tr\left(\mathbf{W}^T\left(T_w-\alpha\mathbf{T}_b\right)\mathbf{W}\right)+\lambda\left\|\mathbf{W}\right\|_{2,1}\\
&s.t.\ \ \mathbf{W}^T\mathbf{S}_t\mathbf{W}=\mathbf{I}_d
\end{aligned}
\tag{12}
$$

where $\mathbf{T}_w=\sum_{i=1}^{n}\sum_{l}^{k_1}\left(\mathbf{x}_i-\mathbf{x}_i^l\right)\left(\mathbf{x}_i-\mathbf{x}_i^l\right)^T$, $\mathbf{T}_b=\sum_{i=1}^{n}\sum_{l}^{k_2}\left(\mathbf{x}_i-\tilde{\mathbf{X}}_i^l\right)\left(x_i-\tilde{\mathbf{X}}_i^l\right)^T$, $\alpha\geq 0$ is a tradeoff parameter that balances temporal variation of within-class and between-class and $\mathbf{I}_d$ is the identity matrix. According to our previous work[37], total scatter matrix $\mathbf{S}_t$ can be written as:

$$
\mathbf{S}_t=\mathbf{H}\mathbf{H}^T \tag{13}
$$

where $\mathbf{H}=\mathbf{P}\Lambda^{\frac{1}{2}}$ and singular value decomposition of $\mathbf{S}_t$ is $\mathbf{S}_t=\mathbf{P}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}P^T$. Thus, model [12] can be reformulated as:

$$\begin{aligned} &\min_{W} tr\left(\mathbf{Q}^T\mathbf{H}^{-1}\left(\mathbf{T}_w-\alpha\mathbf{T}_b\right)\left(\mathbf{H}^{-1}\right)^T\mathbf{Q}\right)+\lambda\left\|\mathbf{W}\right\|_{2,1} \\ &s.t.\ \mathbf{Q}^T\mathbf{Q}=\mathbf{I},\mathbf{W}=\mathbf{H}^T\mathbf{Q} \end{aligned} \tag{14}$$

where $tr\left(\mathbf{Q}^T\mathbf{H}^{-1}\left(\mathbf{T}_w-\alpha\mathbf{T}_b\right)\left(\mathbf{H}^{-1}\right)^T\mathbf{Q}\right)$ is the difference between within-class temporal variation and between-class temporal variation and $\left\|\mathbf{W}\right\|_{2,1}$ is the regular term.

Since $\mathbf{Q}$ and $\mathbf{W}$ are dependent on each other, problem (14) cannot be solved directly. Inspired by success in [34,38], we decompose nonlinear model into two sub-problems.

◆Solving $\mathbf{Q}$ by removing $\mathbf{W}$, the objective function (14) is deducted to

$$\begin{aligned} &\min_{\mathbf{Q}} tr\left(\mathbf{Q}^T\mathbf{H}^{-1}\left(\mathbf{T}_w-\alpha\mathbf{T}_b\right)\left(\mathbf{H}^{-1}\right)^T\mathbf{Q}\right) \\ &s.t.\ \ \mathbf{Q}^T\mathbf{Q}=\mathbf{I} \end{aligned} \tag{15}$$

Because matrix $\mathbf{H}^{-1}\left(\mathbf{T}_w-\alpha\mathbf{T}_b\right)\left(\mathbf{H}^{-1}\right)^T$ is symmetric, the subproblem (15) can be easily solved by strand eigenvalue decomposition.

◆While the matrix $\mathbf{Q}$ is fixed, the projection matrix $\mathbf{W}$ can be solved using a regression model

$$\begin{aligned} &\min_{\mathbf{W}} \left\|\mathbf{W}\right\|_{2,1} \\ &s.t.\ \left(\mathbf{H}^{-1}\right)^T\mathbf{W}=\mathbf{Q} \end{aligned} \tag{16}$$

When the linear equation $\left(\mathbf{H}^{-1}\right)^T\mathbf{W}=\mathbf{Q}$ has a single unique solution, and it does not have the row sparsity property. We turn to the constrained problem (16) as the following regularized problem:

$$\min_{\mathbf{W}} \left\|\mathbf{W}\right\|_{2,1}+\mu\left\|\left(\mathbf{H}^{-1}\right)^T\mathbf{W}-\mathbf{Q}\right\|_F^2 \tag{17}$$

where $\mu>0$ is regularize parameter. Inspired by Gu[34] and Nie[33], we take an iterative algorithm to optimize problem (17).

In summary, we present the algorithm for optimizing problem (14) in Algorithm 1 The convergence of this algorithm was proved in [46]. In our paper, we set a maximum number of iterations.

Algorithm 1 Solving problem for model (14)

Require: Matrixes $\mathbf{T}_w$, $\mathbf{T}_b$ and $\mathbf{S}_t$;

Ensure: $\mathbf{W}\in R^{m_2\times d_2}$

1:Perform Singular Value Decomposition to $\mathbf{S}_t=\mathbf{P}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{P}^T$ and compute $\mathbf{H}=\mathbf{P}\Lambda^{\frac{1}{2}}$;

2:Solve eigenvalue decomposition problem $\mathbf{Q}^T\mathbf{H}^{-1}\left(\mathbf{T}_w-\alpha\mathbf{T}_b\right)\left(\mathbf{H}^{-1}\right)^T\mathbf{Q}$ and

select eigenvector $\mathbf{Q}=[\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_d]$ corresponding to the first smallest eigenvalues $\lambda_1,\lambda_2,\cdots,\lambda_d$;

3:Set $t=0$, initialize $\mathbf{G}_t=\mathbf{I}$

4: **while** not convergence

5: Calculate $\mathbf{W}_{t+1}=\mathbf{G}^{-1}\mathbf{H}^{-1}\left(\left(\mathbf{H}^{-1}\right)^T\mathbf{G}^{-1}\mathbf{H}^{-1}+\frac{1}{2\mu}I\right)^{-1}\mathbf{Q}$

6: Calculate the diagonal matrix $\mathbf{G}_{t+1}$,where the *ith* diagonal element is $\frac{1}{2\left\|\mathbf{w}_{t+1}^i\right\|_2}$

7: $t=t+1$;

8: **end while**

3.5 Bilateral SFDA via $L_{2,1}$ Norm Minimization (BSFDA-$L_{2,1}$)

2DSFDA-$L_{2,1}$ adopts a unilateral projection (right multiplication) scheme, which needs more coefficients for representing an image than vector-based methods (i.e. SFDA). As an extension of 2DSFDA-$L_{2,1}$, bilateral projections method called Bilateral SFDA via $L_{2,1}$ Norm Minimization (BSFDA-$L_{2,1}$) is developed, where left and right projection directions are calculated simultaneously. BSFDA- $L_{2,1}$ takes much less coefficients than 2DSFDA-$L_{2,1}$ to represents an image. The goal of BSFDA- $L_{2,1}$ is to find two projection matrixes $U\in R^{m_1\times d_1}$ and $V\in R^{m_2\times d_2}$ to minimize within-class temporal variation and maximize between-class temporal variation simultaneously. Thus, the model of BSFDA-$L_{2,1}$ can be written as follows:

$$\begin{aligned}&\min_{\mathbf{U},\mathbf{V}} J_w(\mathbf{U},\mathbf{V})-\alpha J_b(\mathbf{U},\mathbf{V})+\lambda_1\|\mathbf{U}\|_{2,1}+\lambda_2\|\mathbf{V}\|_{2,1}\\ &s.t.\sum_{i=1}^{n}\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{V}\mathbf{V}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\mathbf{U}=\mathbf{I}_{d_1}\\ &\sum_{i=1}^{n}\mathbf{V}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\mathbf{U}\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{V}=\mathbf{I}_{d_2}\end{aligned}\tag{18}$$

where $J_w(\mathbf{U},\mathbf{V})$ is the matrix based on within-class temporal variation and $J_b(\mathbf{U},\mathbf{V})$ is between-class temporal variation. Within-class temporal variation $J_w(\mathbf{U},\mathbf{V})$ defined as follows:

$$\begin{aligned}J_w(\mathbf{U},\mathbf{V}) &=\sum_{i=1}^{n}\sum_{l=1}^{k_1}\left\|\mathbf{U}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\mathbf{V}\right\|_2^2\\ &=\sum_{i=1}^{n}\sum_{l=1}^{k_1}tr\left(\mathbf{U}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\mathbf{V}\mathbf{V}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)^T\mathbf{U}\right)\\ &=\sum_{i=1}^{n}\sum_{l=1}^{k_1}tr\left(\mathbf{V}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)^T\mathbf{U}\mathbf{U}^T\left(\mathbf{X}_i-\mathbf{X}_i^l\right)\mathbf{V}\right)\end{aligned}\tag{19}$$

and between-class temporal variation $J_b\left(\mathbf{U},\mathbf{V}\right)$ is defined as:

$$\begin{aligned} J_b\left(\mathbf{U},\mathbf{V}\right) &= \sum_{i=1}^{n}\sum_{l=1}^{k_2}\left\|\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\mathbf{V}\right\|_2^2 \\ &= \sum_{i=1}^{n}\sum_{l=1}^{k_2} tr\left(\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\mathbf{V}\mathbf{V}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)^T\mathbf{U}\right) \\ &= \sum_{i=1}^{n}\sum_{l=1}^{k_2} tr\left(\mathbf{V}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)^T\mathbf{U}\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}_i^l\right)\mathbf{V}\right) \end{aligned} \tag{20}$$

We solve the variables U and V alternatively because there is no closed-form solution for the problem 18. The proposed iterative method includes two sub-steps.

◆Given $\mathbf{U}=\mathbf{U}_t$, update $\mathbf{V}_{t+1}$ by:

$$\begin{aligned} &\min_{\mathbf{V}} J_w\left(\mathbf{U}_t,\mathbf{V}\right)-\alpha J_b\left(\mathbf{U}_t,\mathbf{V}\right)+\lambda_2\left\|\mathbf{V}\right\|_{2,1} \\ &s.t.\sum_{i=1}^{n}\mathbf{V}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\mathbf{U}_t\mathbf{U}_t^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{V}=\mathbf{I}_{d_2} \end{aligned} \tag{21}$$

◆ Given $\mathbf{V}=\mathbf{V}_{t+1}$, update $\mathbf{U}_{t+1}$ by:

$$\begin{aligned} &\min_{\mathbf{U}} J_w\left(\mathbf{U},\mathbf{V}_{t+1}\right)-\alpha J_b\left(\mathbf{U},\mathbf{V}_{t+1}\right)+\lambda_1\left\|\mathbf{U}\right\|_{2,1} \\ &s.t.\sum_{i=1}^{n}\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{V}_{t+1}\mathbf{V}_{t+1}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\mathbf{U}=\mathbf{I}_{d_1} \end{aligned} \tag{22}$$

Similar to the way of solving 2DSFDA-$L_{2,1}$, we use algorithm 1 to solve problem 3. Thus, the process of algorithm for BSFDA-$L_{2,1}$ is summarized in Algorithm 2.

Algorithm 2 Alternatively iterative method for model (18)

Ensure: Two projection matrixes $\mathbf{U}\in R^{m_1\times d_1}$ and $\mathbf{V}\in R^{m_2\times d_2}$

1: Initialize $t=0$ and $\mathbf{U}_0$ as a random matrix

while not convergence

3: Given $\mathbf{U}=\mathbf{U}_t$,update $\mathbf{V}_{t+1}$ by:

$$\begin{aligned} &\min_{\mathbf{V}} J_w\left(\mathbf{U}_t,\mathbf{V}\right)-\alpha J_b\left(\mathbf{U}_t,\mathbf{V}\right)+\lambda_2\left\|\mathbf{V}\right\|_{2,1} \\ &s.t.\sum_{i=1}^{n}\mathbf{V}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\mathbf{U}_t\mathbf{U}_t^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{V}=\mathbf{I}_{d_2} \end{aligned}$$

Given $\mathbf{V}=\mathbf{V}_{t+1}$, update $\mathbf{U}_{t+1}$ by:

$$\begin{aligned} &\min_{\mathbf{U}} J_w\left(\mathbf{U},\mathbf{V}_{t+1}\right)-\alpha J_b\left(\mathbf{U},\mathbf{V}_{t+1}\right)+\lambda_1\left\|\mathbf{U}\right\|_{2,1} \\ &s.t.\sum_{i=1}^{n}\mathbf{U}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)\mathbf{V}_{t+1}\mathbf{V}_{t+1}^T\left(\mathbf{X}_i-\tilde{\mathbf{X}}\right)^T\mathbf{U}=\mathbf{I}_{d_1} \end{aligned}$$

4: $t = t + 1$
5: **end while**

4. Experimental Classification Results and Analysis

To evaluate the proposed 2DSSFDA-$L_{2,1}$ and BSFDA-$L_{2,1}$, we compared it with other feature extraction methods including 2D based methods and bilateral methods on three well-known face image databases (Extended YaleB, CMU PIE and COIL-20). The 2D based methods include 2DPCA[29], 2DLPP[33], 2DLDA[32], S2DLDP[42], 2DNPP[36], N2DNPP[47] and 2DSFDA. According to the literatures[41,42],the authors take the optimal α to guarantee that the power of discriminant arrives maximum using an iterative method. In our experiment, we set $\alpha = \dfrac{trace\left(\mathbf{Q}^T\mathbf{H}^{-1}\mathbf{T}_w\left(\mathbf{H}^{-1}\right)^T\mathbf{Q}\right)}{trace\left(\mathbf{Q}^T\mathbf{H}^{-1}\mathbf{T}_b\left(\mathbf{H}^{-1}\right)^T\mathbf{Q}\right)}$ for maximizing the power of discriminant. To indicate the effectiveness of bilateral version, we compare BSFDA-$L_{2,1}$ with two other bilateral 2D methods include BLDA and BSFDA. For BSFDA-$L_{2,1}$, we also take the same manner to determine the value of tradeoff parameter α. After projecting a testing sample into a learned subspace, the nearest neighbor (NN) classifier is employed for classification task.

4.1 Experiment on Extended YaleB face database

The Extended YaleB Face Database contains 16128 images under 9 poses and 64 illumination conditions. In this experiment, we use the same experiment setting in [48]. We choose the frontal pose and use all the images under different illumination, thus we get a subset contains 2431 images of 38 individuals. Before implementing our experiment, all of the face images are resized into the resolution of 96×84. **Fig. 1** shows some sample images.

Fig. 1. Sample images in Extended YaleB face database

Specially, we randomly select $l(=5,6)$ samples from each individual for training, and the rest of samples are used for testing. For each given l, we repeat each experiment 20 times and calculate the average recognition accuracy. **Fig. 2** shows the recognition accuracy of 2D based methods with varying projections. The maximal average recognition accuracy of each 2D method and the corresponding number of projections as well as standard deviations are given in **Table 1**. **Table 2** lists average recognition accuracy at varying feature number of bilateral 2D methods. **Fig. 4** gives the recognition curves of three bilateral 2D methods at different number of features. **Fig. 3** shows the maximal recognition curves of eight different methods with different number of training samples per class.

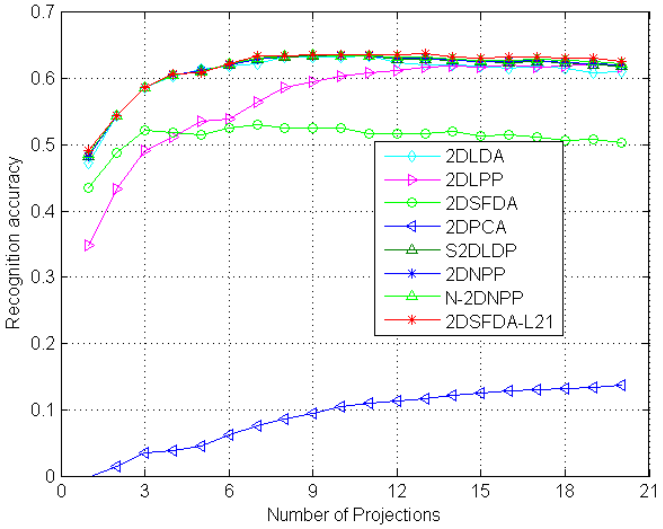


Fig. 2. Recognition accuracy of 2D based methods with varying projections number on the Extended YaleB face database

Table 1. The maximal average recognition accuracy (%) and their corresponding standard deviations, optimal number of projections of across 20 runs on Extended YaleB

Method	2DPCA	2DLPP	2DLDA	2DSFDA	S2DLDP	2DNPP	N-2DNPP	2DSFDA-$L_{2,1}$
$l = 5$	21.10 ±1.70 96×20	57.87 ±3.17 96×18	59.57 ±3.15 96×12	47.40 ±4.62 96×5	60.56 ±2.25 96×10	59.98 ±2.25 96×11	60.44 ±2.25 96×10	61.33 ±2.56 96×10
$l = 6$	24.70 ±1.37 96×20	62.73 ±4.14 96×18	64.30 ±2.22 96×17	53.43 ±2.82 96×10	64.94 ±1.96 96×13	63.53 ±2.25 96×11	64.29 ±2.25 96×12	65.30 ±2.28 96×12

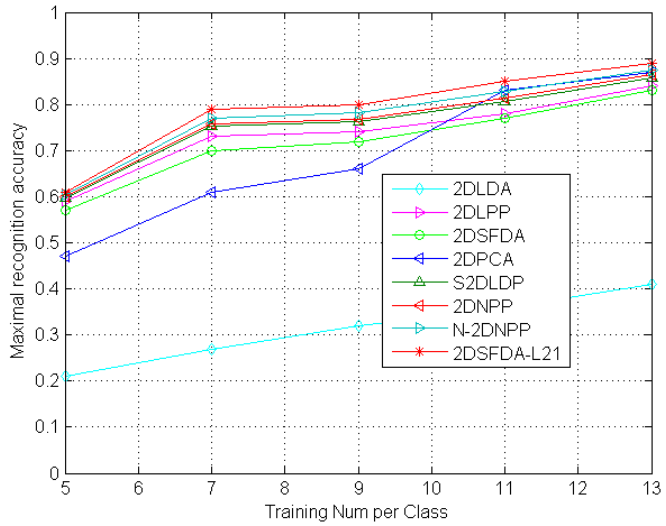


Fig. 3. Maximal recognition accuracy of 2D based methods with varying training number per class on the Extended YaleB face database

Fig. 2 and **Fig. 3** show the variations of the classification accuracy with different subspace dimensionality and different number of training samples. From **Fig. 2** and **Table 1**, we can see that our proposed method 2DSFDA-$L_{2,1}$ obtains better recognition accuracy than other 2D based methods including 2DPCA, 2DLPP, 2DLDA, 2DSFDA, 2DNDP and N2DNDP. 2DSFDA-$L_{2,1}$ obtains the best classification accuracy on the 10th projection. From **Fig. 3**, we can see that increasing training number of each class will improve the recognition accuracy. The reason is that more samples are favor to explore the potential structure of data. The experimental results show that 2DSFDA-$L_{2,1}$can obtain the best classification accuracy compared with the other methods.

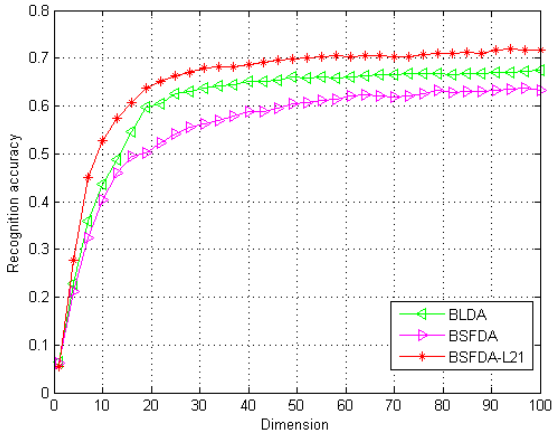


Fig. 4. Recognition accuracy of Bilateral based methods with dimensions 10×10 on the Extended YaleB face database

Table 2. Recognition Accuracy (%) of Bilateral 2D based Methods on Extended YaleB face database

Right / Left	5			8			10		
	BLDA	BSFDA	BSFDA-$L_{2,1}$	BLDA	BSFDA	BSFDA-$L_{2,1}$	BLDA	BSFDA	BSFDA-$L_{2,1}$
5	51.47	53.40	61.83	57.20	58.83	62.57	56.20	59.53	65.93
8	63.07	59.97	67.70	63.43	62.20	68.40	63.70	63.20	69.73
10	67.03	60.67	68.67	66.57	63.73	69.43	67.47	63.38	71.87

As shown in **Fig. 4** and **Table 2**, $L_{2,1}$norm based bilateral method BSFDA-$L_{2,1}$ also outperforms other bilateral methods. The above experimental results demonstrate the effectiveness and efficiency of $L_{2,1}$ norm penalty term. The reason is that using $L_{2,1}$ norm term can take feature selection and subspace learning simultaneously, which can improve subspace learning and encourage row-sparsity. In addition, the uncorrelated feature can provide substantial complementary information for face recognition. Comparing **Table 1** and **Table 2**, the best recognition accuracy of 2DSFDA-$L_{2,1}$ is 65.30% when the dimension arrives 96*12=1152 , and the best recognition accuracy of 2DSFDA-$L_{2,1}$ is 71.87% when the dimension arrives 10*10=100 . It demonstrates that the feature extracted by bilateral projections is with much less coefficients and better favor for classification.

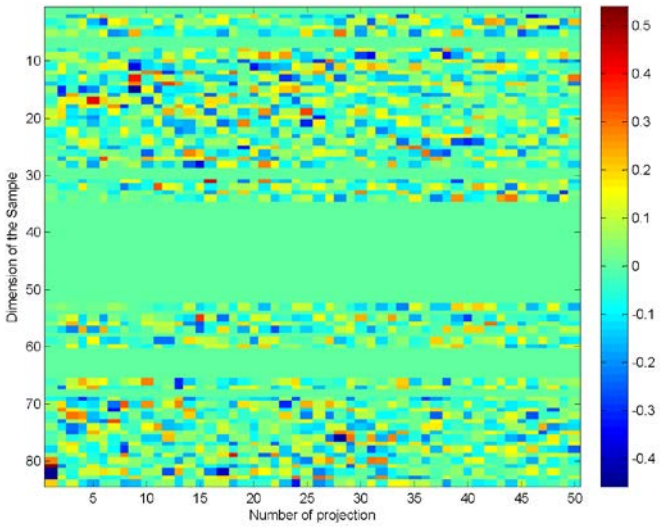
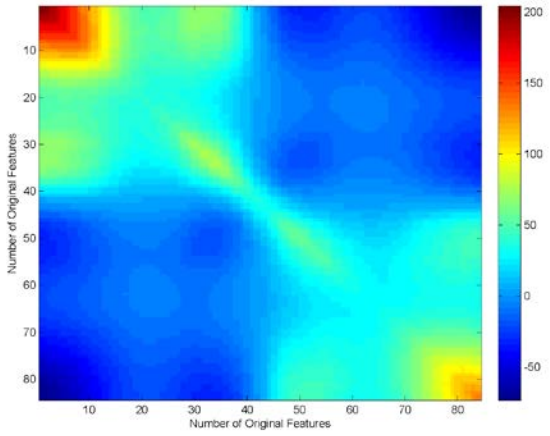
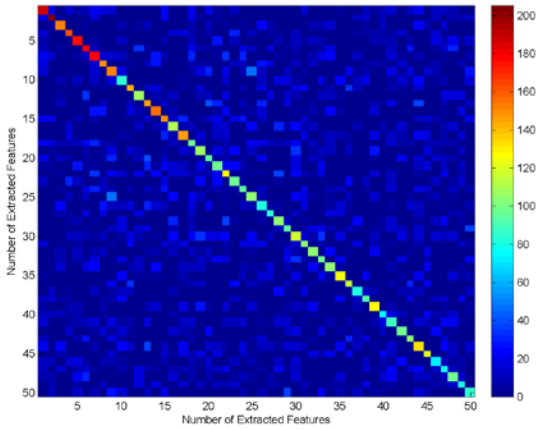


Fig. 5. Graphic of Projection Matrix

In the second experiment, we demonstrate the functionalities of the proposed model 12 by the graphics of projective matrix W. **Fig. 5** shows the graphic of projective matrix W, we can see that W has the property of row-sparsity. It demonstrates the effectiveness of $L_{2,1}$ regularized term imposed on W, which makes the feature selection and subspace learning simultaneously. When W is multiplied on both of sides of S_t, we could get the total scatter matrix $S_L = W^T S_t W$ in the subspace. **Fig. 6(a)** shows the graphic of total matrix S_t in the original feature space and **Fig. 6(b)** shows the graphic of matrix S_L in the learned subspace. The non diagonal elements of **Fig. 6(a)** reflect the correlation between different features. From **Fig. 6(b)**, all the "energies" are assembled in the diagonal of the matrix. Values of non diagonal elements are close to zero. Thus, the graphic reveals that the extracted features in the low-dimensional subspace are highly uncorrelated, which means that the majority of the redundant information has been reduced.



(a) The total scatter matrix in the original feature space



(b) The total scatter matrix in the subspace

Fig. 6. Graphic of total scatter matrix in original feature space and extracted feature space

4.2 Experiment on CMU PIE face database

The CMU PIE Face Database [39] contains 41,368 images from 68 individual. These images of each individual were taken under 13 different poses, 43 different illumination conditions, and with 4 different expressions. Same as the experiment setting [44], We choose the five near frontal poses (C05, C07, C09, C27, C29) and use all the images under different illuminations and expressions, thus we get a subset contains 11,554 images of 68 individuals. Before implement our experiment, we crop the face portion of the image into the resolution of 64×64. Some sample images are shown in **Fig. 7**.

Fig. 7. Sample images in PIE database

First, we test the performance of proposed 2DSFDA-$L_{2,1}$ and BSFDA-$L_{2,1}$ compared with some other 2D methods. We randomly choose $l=(5,6)$ images of each person for training and the rest images for testing. Each experiment is repeated 20 times to get the average recognition accuracy and standard deviations. **Table 3** gives the maximal average recognition accuracy obtained by different unilateral 2D methods as well as standard deviations and the corresponding projection numbers. **Fig. 8** shows the average recognition accuracy of different unilateral 2D methods with varying projections. The recognition curves of three bilateral 2D methods at different number of feature are show in **Fig. 9**. **Table 4** gives the average recognition accuracy at different features of bilateral 2D methods.

According to the experimental results includeing **Fig. 8**, **Fig. 9**, **Table 3** and **Table 4**, we can see that the $L_{2,1}$ norm minimization based methods 2DSFDA-$L_{2,1}$ and BSFDA- $L_{2,1}$ perform better than other 2D based methods in most cases. It also demonstrates that the row-sparsity projection performs as a filter and can reduce the negative effect as a result of the facial expression and illumination variation. Thus, the row-sparsity projections can obtain

good performance when using low dimensional features for classification. Another mechanism is that uncorrelated feature is helpful to improve the performance of classification task, because removing the correlation between features can enhance the discriminant ability.

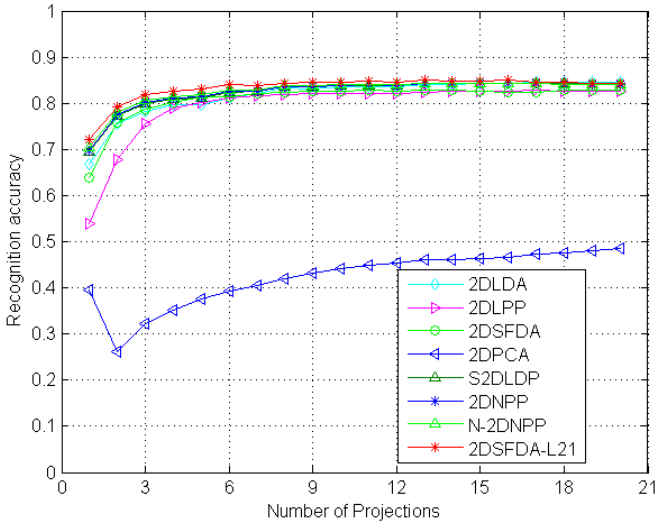


Fig. 8. Recognition accuracy of 2D based methods with varying number of projections on CMU PIE face database

Table 3. The maximal average recognition accuracy (%) and their corresponding standard deviations, optimal number of projections of across 20 runs on CMU PIE face database

Method	2DPCA	2DLPP	2DLDA	2DSFDA	S2DLDP	2DNPP	N-2DNPP	2DSFDA-$L_{2,1}$
l = 5	53.23 ±3.30 64×20	79.77 ±1.99 64×20	81.33 ±2.14 64×20	80.77 ±2.39 64×15	81.43 ±2.39 64×15	80.95 ±2.23 64×15	81.59 ±2.34 64×16	82.37 ±2.10 64×12
l = 6	57.47 ±2.53 64×20	82.77 ±1.64 64×17	84.17 ±1.72 64×20	82.93 ±1.26 64×18	83.84 ±1.48 64×13	82.84 ±1.60 64×13	84.34 ±1.75 64×14	85.33 ±1.60 64×10

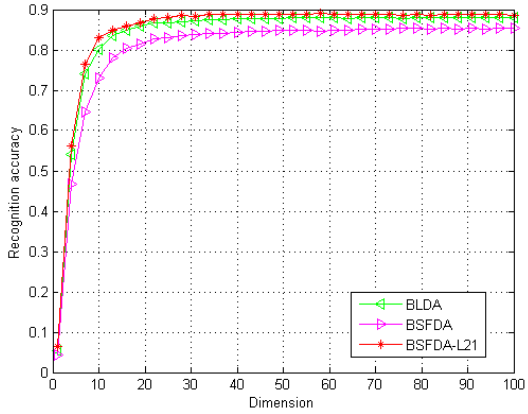


Fig. 9. Recognition accuracy of bilateral 2D based methods with dimensions 10×10 on the CMU PIE face database

Table 4. Recognition Accuracy (%) of Bilateral 2D based method on CMU PIE face database

Right / Left	5			8			10		
	BLDA	BSFDA	BSFDA-$L_{2,1}$	BLDA	BSFDA	BSFDA-$L_{2,1}$	BLDA	BSFDA	BSFDA-$L_{2,1}$
5	85.43	82.60	87.43	87.60	84.70	87.93	86.17	83.87	86.67
8	87.00	82.93	88.00	87.77	84.20	88.57	88.23	84.83	88.57
10	88.20	84.27	89.57	87.53	84.43	88.87	88.10	85.47	89.07

In the second experiment, we study the impact of parameter α to the performance of 2DSFDA-$L_{2,1}$. The value of parameter α is set to be 0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50 and 100 respectively. We randomly select l (l = 5,6) images from each class for training, while the remaining images are used for testing. **Fig. 10** shows the performance of 2DSFDA-$L_{2,1}$ when using different values of parameter α and different numbers of training samples for each class. From the experimental results, we find that 2DSFDA-$L_{2,1}$ can always obtain the best recognition rate in most cases. It also demonstrates that our method is robust to the choice of parameter α.

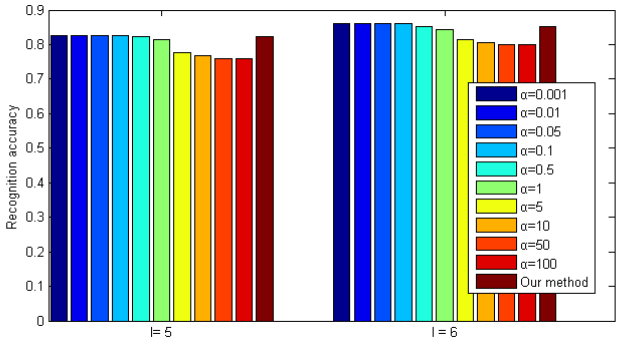


Fig. 10. Maximal average recognition accuracy of 2DSFDA- $L_{2,1}$with different parameter α using l (l = 5,6) training samples per class on CMU PIE face database

4.3 Experiment on COIL-20 database

The COIL-20 database consists of 1440 images from 20 objects. The those objects were placed on a motorized turntable against a black background. Images of each objects were taken at pose intervals 5 degrees, corresponding to 72 image per object. All the images are cropped and resized to 32× 32 pixels. Some samples are shown in **Fig. 11**.

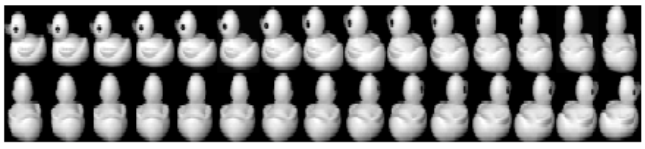


Fig. 11. Sample images in COIL20 database

The same setting as previous experiments, for each subject, l(=5,6) images are randomly selected used for training and the remaining images are used for testing. To compute average recognition accuracy, each experiment is randomly repeated 20 times. The maximum recognition accuracy and the corresponding projection number of five unilateral 2D methods

are shown in **Table 5**. **Table 6** lists the recognition accuracy at varying features number of bilateral 2D based method. **Fig. 12** shows the variations of the projections number versus recognition accuracy with six different unilateral 2D methods. **Fig. 13** shows the recognition accuracy versus different number of feature with three different bilateral 2D based methods. The experimental results also support that our methods performances better than others.

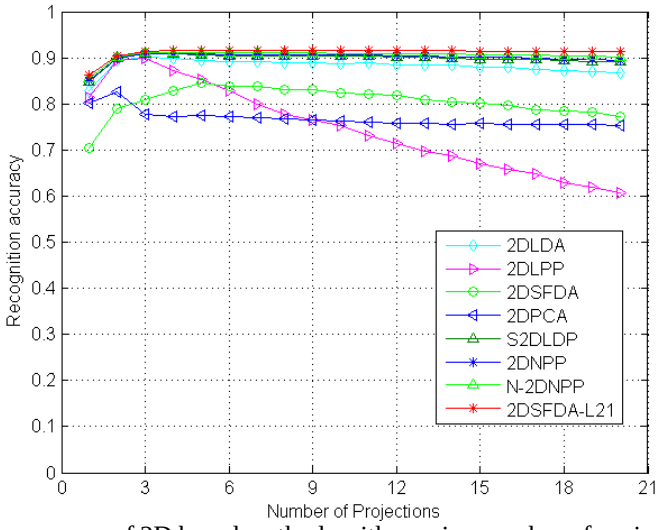


Fig. 12. Recognition accuracy of 2D based methods with varying number of projections on the COIL20 database

Table 5. The maximal average recognition accuracy (%) and their corresponding standard deviations, optimal number of projections of across 20 runs on COIL20 database

Method	2DPCA	2DLPP	2DLDA	2DSFDA	S2DLDP	2DNPP	N-2DNPP	2DSFDA-$L_{2,1}$
$l = 5$	82.53 ±2.12 32×2	89.93 ±2.37 32×3	90.27 ±2.23 32×3	84.67 ±3.49 32×5	90.89 ±1.29 32×4	89.79 ±2.13 32×4	90.35 ±1.57 32×5	91.67 ±2.52 32×4
$l = 6$	83.92 ±1.94 32×2	91.83 ±1.64 32×3	93.42 ±1.79 32×3	87.67 ±3.88 32×5	92.24 ±1.50 32×11	92.07 ±1.55 32×11	92.67 ±1.85 32×11	92.75 ±2.63 32×17

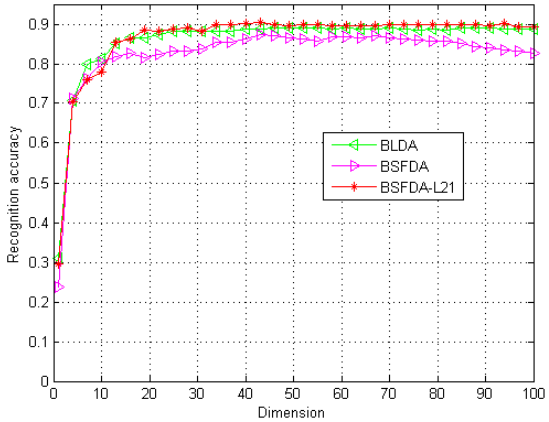
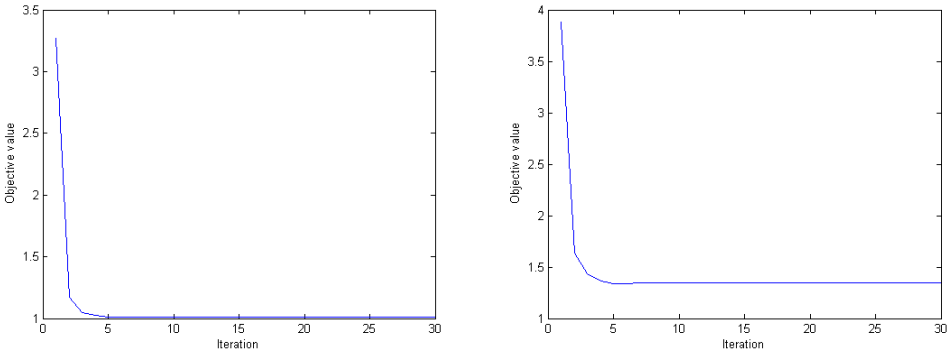


Fig. 13. Recognition accuracy of Bilateral based methods with dimensions 10×10 on the COIL20 database

Table 6. Recognition Accuracy (%) of Bilateral 2D based method on COIL20 database

Right / Left	5			8			10		
	BLDA	BSFDA	BSFDA-$L_{2,1}$	BLDA	BSFDA	BSFDA-$L_{2,1}$	BLDA	BSFDA	BSFDA-$L_{2,1}$
5	85.50	79.92	86.42	85.53	80.40	88.13	86.75	78.83	88.13
8	87.83	82.00	89.58	88.33	82.33	90.33	90.25	82.50	90.67
10	87.50	84.33	88.67	87.08	81.83	88.75	89.67	84.00	90.08

In the second, we investigate the convergence of BSFDA-L21responding to the objective value of model 18. We randomly choose 6 images per class for training. **Fig. 14** show convergence curves of BSFDA-L21 in terms of objective value with different dimensions. From the experimental results, we can see that BSFDA-L21 always converges very fast, usually than 5 iterations.



(a)$U \in R^{32\times8}, V \in R^{32\times8}$ (b)$U \in R^{32\times10}, V \in R^{32\times10}$

Fig. 14. The objective value of BSFDA-L_{21} of each iteration on COIL20 database

5. Conclusion

This paper presents a Two dimensional based sparse slow feature discriminant analysis model including 2DSFDA-$L_{2,1}$ and BSFDA-$L_{2,1}$ for feature extraction and face recognition. The key of our model is to combine the $L_{2,1}$ norm regression and statistically uncorrelated constraint into the 2D slow feature discriminant analysis framework. We presented a feasible solution by transforming $L_{2,1}$ norm based nonlinear model into a linear regression type. The learned row-sparsity projection can make feature selection and subspace learning simultaneously. Experiments on four benchmark databases demonstrate the effectiveness of our proposed methods. Although promising results have been obtained by our algorithms, two future efforts are still worth making. The first, we will enhance to interpret "slow feature" from the view of Biological mechanism. The second , we will extend our work to large-scale image retrieval by using slow features extraction.

Acknowledgments

This work is supported by Fundamental Research Funds for the Central Universities (Grant No. KYZ201666, KYZ201753), the National Natural Science Foundation of China (Grant No. 61702265, 61373062, 61373063,61203247), Natural Science Foundation of Jiangsu Province(Grant No. BK20170856), CCF-Tencent Open Research, National Key Technology Research and Development Program of the Ministry of Science and Technology of China

(Grant No. 2015BAK36B05), and Open Fund project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Mingjiang University)(No.MJUKF201730).

References

[1] Z. Zhang, F. Li, M. Zhao, et al. "Robust Neighborhood Preserving Projection by Nuclear/L2,1-Norm Regularization for Image Feature Extraction," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 26, no. 4, pp. 1607-1622, 2017. Article (CrossRef Link)

[2] N. Ali, B.Bajwa, R.Sablatnig and Z. Mehmood. "Image retrieval by addition of spatial information based on histograms of triangular regions," *Computers & Electrical Engineering*, vol. 54, pp. 539-550, August , 2016, Article (CrossRef Link)

[3] N.Ali, K.Bajwa et al., "A Novel Image Retrieval Based on Visual Words Integration of SIFT and SURF," *Plos One*, vol.11, no.6, pp.e0157428, Jun, 2016, Article (CrossRef Link)

[4] Jinhui Tang, Zechao Li, Meng Wang, Ruizhen Zhao. "Neighborhood Discriminant Hashing for Large-Scale Image Retrieval," *IEEE Transactions on Image Processing*, vol. 24, no. 9, 2015. Article (CrossRef Link)

[5] Haojie Li, Xiaohui Wang, Jinhui Tang, Chunxia Zhao, "Combining global and local matching of multiple features for precise item image retrieval," *Multimedia Syst.*, vol. 19, no. 1, pp, 37-49, February, 2013, Article (CrossRef Link)

[6] H. Tan,Y. Gao,Z. Ma, "Regularized constraint subspace based method for image set classification," *Pattern Recognition*, vol.76, PP. 434-448, April, 2017. Article (CrossRef Link)

[7] X. Zhu, X. Li, S. Zhang, "Block-Row Sparse Multiview Multilabel Learning for Image Classification," *IEEE Transactions on Cybernetics*, vol.46, no.2, pp.450-461, February, 2015. Article (CrossRef Link)

[8] M. Turk, A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*,vol. 3, no.1, pp. 71–86, 1991. Article (CrossRef Link)

[9] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, July,1997. Article (CrossRef Link)

[10] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang,"Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328-340, March, 2005, Article (CrossRef Link)

[11] X. He, D. Cai, S. Yan, H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. of IEEE International Conference on Computer Vision*, Vol. 2, pp. 1208-1213 , October 17-20, 2005, Article (CrossRef Link)

[12] T. Zhang, J. Yang, D. Zhao, X. Ge, "Linear local tangent space alignment and application to face recognition," *Neurocomputing*,vol.70, no.7, pp. 1547-1553, March, 2007. Article (CrossRef Link) .

[13] W. Yu, X. Teng, C. Liu, "Face recognition using discriminant locality preserving projections," *Image and Vision computing*, vol. 24, no. 3, pp.239-248, March, 2006. Article (CrossRef Link)

[14] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no.1, pp.40-51, November, 2006. Article (CrossRef Link)

[15] H.-W. Chang, T.-L. Liu, "Local discriminant embedding and its variants," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2005, pp. 846-853, June 20-25, 2005. Article (CrossRef Link)

[16] L. Wiskott, T. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp.715-770, April, 2002. Article (CrossRef Link)

[17] R. Legenstein, N. Wilbert, L. Wiskott, "Reinforcement learning on slow features of high dimensional input streams," *PLoS computational biology* vol. 6, no.8, pp.833-835, August, 2010. Article (CrossRef Link)

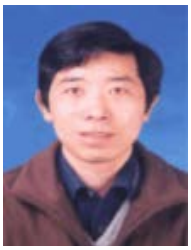
[18] M. Franzius, N. Wilbert, L. Wiskott, "Invariant object recognition and pose estimation with slow feature analysis," *Neural computation* vol. 23, no.9, pp. 2289-2323, September, 2011. Article (CrossRef Link)
[19] P. Berkes, L. Wiskott, "Slow feature analysis yields a rich repertoire of complex cell properties," *Journal of Vision*, vol. 5, no.6, pp. 579-602, 2005. Article (CrossRef Link)
[20] M. Franzius, H. Sprekeler, L. Wiskott, "Slowness and sparseness lead to place, head-direction, and spatial-view cells," *PLoS Computational Biology*, vol.3, no.8, pp1605-1622, August, 2007. Article (CrossRef Link)
[21] Z. Zhang, D. Tao, "Slow feature analysis for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol.34, no.3, pp.436-450, March 2012. Article (CrossRef Link)
[22] Y. Huang, J. Zhao, Y. Liu, S. Luo, Q. Zou, M. Tian, "Nonlinear dimensionality reduction using a temporal coherence principle," *Information Sciences*, vol. 181, no.16, pp.3284-3307, August, 2011. Article (CrossRef Link)
[23] X. Gu, C. Liu, S. Wang, "Supervised slow feature analysis for face recognition," *Biometric Recognition*, pp. 178–184, November 16-17, 2013. Article (CrossRef Link)
[24] Y. Huang, J. Zhao, M. Tian, Q. Zou, S. Luo, "Slow Feature Discriminant Analysis and its application on handwritten digit recognition," in *Proc. of International Symposium on Neural Networks*, pp. 1294-1297, June 14-19, 2009. Article (CrossRef Link)
[25] X. Gu, C. Liu, S. Wang, C. Zhao, "Feature extraction using adaptive slow feature discriminant analysis," *Neurocomputing*,vol.154, pp. 139-148, Aprial, 2015, Article (CrossRef Link)
[26] Z. Jin, J.-Y. Yang, Z.-M. Tang, Z.-S. Hu, "A theorem on the uncorrelated optimal discriminant vectors," *Pattern Recognition*, vol. 34, no.10, pp. 2041-2047, October, 2001. Article (CrossRef Link)
[27] X. Jing, S. Li, D. Zhang, J. Yang, "Face recognition based on local uncorrelated and weighted global uncorrelated discriminant transforms," in *Proc. of IEEE International Conference on Image Processing*, pp. 3049-3052, September 11-14, 2011. Article (CrossRef Link)
[28] C. Zhao, D. Miao, Z. Lai, C. Gao, C. Liu, J. Yang, "Two-dimensional color uncorrelated discriminant analysis for face recognition," *Neurocomputing*, vol. 113, pp. 251–261, August, 2013, Article (CrossRef Link)
[29] J. Yang, D. Zhang, A. F. Frangi, J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no.1, pp.131-137, January, 2004, Article (CrossRef Link)
[30] X. Li, Y. Pang, Y. Yuan, "L1-norm-based 2dpca," *IEEE transactions on systems, man, and cybernetics. Part B*, vol. 40, no.4, pp. 1170–1175, August 2010, Article (CrossRef Link)
[31] F. Zhang, J. Yang, J. Qian, Y. Xu, "Nuclear norm-based 2-dpca for extracting features from images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no.10, October, 2015. Article (CrossRef Link)
[32] M. Li, B. Yuan, "2d-lda: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527-532, Aprial, 2005. Article (CrossRef Link)
[33] S. Chen, H. Zhao, M. Kong, B. Luo, "2d-lpp: a two-dimensional extension of locality preserving projections," *Neurocomputing*, vol.70, no.4, pp. 912-921, January, 2007. Article (CrossRef Link)
[34] D. Hu, G. Feng, Z. Zhou, "Two-dimensional locality preserving projections (2dlpp) with its application to palmprint recognition," *Pattern recognition*, vol. 40, no.1, pp. 339-342, January, 2007. Article (CrossRef Link)
[35] B. Niu, Q. Yang, S. C. K. Shiu, S. K. Pal, "Two-dimensional laplacianfaces method for face recognition," *Pattern Recognition*, vol. 41, no.10, pp.3237-3243, October, 2008. Article (CrossRef Link)
[36] H. Zhang, Q. M. Wu, Tommy W. S. Chow, and M. Zhao, "A two-dimensional Neighborhood Preserving Projection for appearance-based face recognition," *Pattern Recognition*, vol. 45, no. 5, pp. 1866-1876, May, 2012. Article (CrossRef Link)
[37] H. Zhao, H. Xing, X. Wang et al., "L_1-Norm-Based 2DLPP," in *Proc. of Control and Decision Conference*, vol. 1-6, pp.1259-1264, May 23-25, 2011. Article (CrossRef Link)

[38] Y. Tang, Z. Zhang,Y. Zhang et al., "Robust L_1-norm matrixed locality preserving projection for discriminative subspace learning," in *Proc. of International Joint Conference on Neural Networks*, pp.4199-4204, July 24-29, 2016. Article (CrossRef Link)
[39] F. Nie, H. Huang, X. Cai, C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," *Advances in Neural Information Processing Systems*, pp. 1813-1821, December 6-9 2010. Article (CrossRef Link)
[40] Q. Gu, Z. Li, J. Han, "Joint feature selection and subspace learning," in *Proc. of International Joint Conference on Artificial Intelligence*, pp. 1294–1299, July 16-22, 2011. Article (CrossRef Link)
[41] Z. Lai, M. Wan, Z. Jin, J. Yang, "Sparse two-dimensional local discriminant projections for feature extraction," *Neurocomputing*, vol.74, no.4, pp. 629–637, January, 2011. Article (CrossRef Link)
[42] H. Kong, L. Wang, E. K. Teoh, X. Li, J.-G. Wang, R. Venkateswarlu, "Generalized 2d principal component analysis for face image representation and recognition," *Neural Networks*, vol.18, no.5, pp. 585–594, June, 2005. Article (CrossRef Link)
[43] D. Zhang, Z.-H. Zhou, "(2d) 2pca: Two-directional two-dimensional pca for efficient face representation and recognition," *Neurocomputing*, vol.69, no.1, pp.224–231, September 17-19, 2005. Article (CrossRef Link)
[44] Y. Li, Z. Tan, Y. Zhan, "Two-dimensional bilinear preserving projections for image feature extraction and classification," *Neural Computing and Applications*, vol.24, no.3-4, pp.901–909, March, 2014. Article (CrossRef Link)
[45] J. Yang, C. Liu, "Horizontal and vertical 2dpca-based discriminant analysis for face verification on a large-scale database," *IEEE Transactions on Information Forensics and Security* vol.2, no.4, pp.781–792, December, 2017. Article (CrossRef Link)
[46] X. Gu, C. Liu, S. Wang, C. Zhao, S. Wu, "Uncorrelated slow feature discriminant analysis using globality preserving projections for feature extraction," *Neurocomputing*, vol.168, pp.488–499, November, 2015. Article (CrossRef Link)
[47] Z. Zhang, F.Li, M.Zhao et al., "Robust Neighborhood Preserving Projection by Nuclear/L2,1-Norm Regularization for Image Feature Extraction," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 26(4):1607-1622, Aprial, 2017. Article (CrossRef Link)
[48] D. Cai, X. He, J. Han, "Spectral regression for efficient regularized subspace learning," in *Proc. of International Conference on Computer Vision*, pp. 214-221, October 14-21, 2007. Article (CrossRef Link)

Xingjian Gu is currently a lecturer in the College of Information Science and Technology at Nanjing Agricultural University. He received the Ph.D. degree from Nanjing University of Science and Technology and B.S. degree from Nanjing University of Information Science and Technology in 2015 and 2009 respectively. His research interests include pattern recognition, computer vision and machine learning.

Xiangbo Shu is an Assistant Professor in School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received his Ph.D. degree in July 2016 from Nanjing University of Science and Technology. From 2014 to 2015, he worked as a visiting scholar in the Department of Electrical and Computer Engineering at National University of Singapore. His research interests include computer vision, and machine learning. He has received the Best Student Paper Award in MMM 2016 and the Best Paper Runner-up in ACM MM 2015.

Shougang Ren is currently an assistant professor in the College of Information Science and Technology at Nanjing Agricultural University. He received his Ph.D. degree and M.S. degree from Nanjing University of Aeronautics and Astronautics in 2005. His research interests focus on the research of artificial intelligence, machine vision, computer agriculture and so on.



Huanliang Xu is a Full Professor in the College of Information Science and Technology at Nanjing Agricultural University. He received his Ph.D. degree from Nanjing University of Aeronautics and Astronautics. His areas of research include machine learning, image processing and the Internet of things.