Locally Joint Sparse Marginal Embedding for Feature Extraction

Dongmei Mo, Zhihui Lai, and Waikeung Wong

***Abstract*—Classical linear discriminant analysis (LDA) has the limitation that it requires the within-class scatter matrix to be nonsingular so that it can perform eigen-decomposition to obtain optimal solutions. To break through this limitation, many methods based on LDA have been proposed. However, these methods are either sensitive to outliers or lack joint sparsity for effective feature extraction. To release these problems, this paper proposes a locally joint sparse marginal embedding (LJSME) method. LJSME reconstructs the scatter matrices and utilizes the locality graph to weigh each pair of data, such that it is robust to outliers and able to preserve the neighborhood relationship of the data. Moreover, LJSME can easily avoid the small sample-size problem by a maximum margin criterion and obtain joint sparsity for effective feature extraction by using joint sparse regularization. The comprehensive analysis between the proposed LJSME and the related methods is presented, which indicates the advantages of the proposed method. A series of experiments was conducted to evaluate the performance of LJSME when compared with the state-of-the-art methods. The MATLAB code of LJSME can be downloaded from https://github.com/TungmeeMo/LJSME.git.**

***Index Terms*—Feature extraction, classification, discriminant analysis, joint sparsity, robustness.**

I. Introduction

With the rapid development of technology and the Internet, the research of pattern recognition is being more and more accessible and available [1]. Many methods have been proposed to improve the performance of tasks on pattern recognition [2]–[4]. However, there are still many difficulties remain to be solved. In pattern recognition, the two major tasks are to reduce the dimensionality of the data and to obtain important features for effective classification. Dimensionality reduction is important because a matrix with $i \times j$ dimensions (an image with $i \times j$ pixels) is too large for computation while conducting recognition tasks. Also, it is hard to guarantee good performance since redundant features can even have negative effect on pattern recognition [5].

To solve this problem, many approaches have been proposed [6]–[8]. One of the most well-known methods is the principle component analysis (PCA) which obtains projections that maximize the total scatters of the face images [9]. However, PCA tends to maintain some unnecessary information due to the interference from the variation of illumination or facial expression. From this perspective, PCA projections fit for reconstruction in low dimensional space but not discriminative analysis. Linear discriminant analysis (LDA) is a classical statistical method which aims to obtain the optimal projections that maximize the between-class scatter and at the same time minimize the within-class scatter. The optimal solutions of LDA can be obtained by solving an eigen-decomposition problem that related to both of the between-class scatter and the within-class scatter [10]. Compared with PCA, LDA takes the label information into consideration so that it can obtain more discriminative information for feature selection. Although LDA is quite simple and effective, it still has some drawbacks: the classical LDA requires one of the scatter matric to be nonsingular. Unfortunately, the dimensionality of data is usually larger than the number of the samples in practical applications (i.e., undersampled problem), which leads to the singularity to one of the scatter matrices [11]. To release the undersampled problem, null-space linear discriminant analysis (NLDA) [12] and direct linear discriminant analysis (DLDA) [13] were proposed, respectively [14]. While NLDA aims to extract those discriminant features from the null space of the within-class scatter, DLDA first discards the null space of the between-class scatter and then extracts the discriminant information from the null space of the within-class scatter. Nevertheless, both of DLDA and NLDA tend to lose some discriminant information that may be useful for pattern recognition or classification. Another method called PCA+LDA [15], [16] is also proposed to deal with the undersampled problem by conducting LDA after using PCA as preprocessing. Unfortunately, PCA+LDA may also miss some important discriminant information in the first stage because only $n - C$ principle components can be kept (note that PCA always needs to keep $n - 1$ principle components to avoid losing information, where n is the number of samples and C is the number of class). To fully use all of the discriminant information, Wang *et al*. proposed dual-space LDA [17].

Manuscript received July 8, 2018; revised November 4, 2018 and January 8, 2019; accepted April 28, 2019. Date of publication May 10, 2019; date of current version November 19, 2019. This work was supported by the Hong Kong Polytechnic University under Project RHR1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zixiang Xiong. *(Corresponding author: Waikeung Wong.)*

D. Mo is with the Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: dongmei_mo@qq.com).

Z. Lai is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: lai_zhi_hui@163.com).

W. Wong is with the Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, and also with The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518057, China (e-mail: calvin.wong@polyu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2916093

In addition, [18] Li *et al*. proposed 2-dimensional LDA (2DLDA) to extract discriminative information from image matrix by using the technic of Fisher's linear discriminant analysis. [19] Ye *et al*. proposed another method called 2DLDA to release singularity problem in classical LDA and improve the efficiency. Besides, Ye *et al*. in [19] also proposed 2DLDA+LDA to extend this work to further reduce the dimensionality of the data before LDA. The tensor-based LDA called TensorLDA conducts discriminative analysis by using matrix representation instead of higher order tensor. Note that 2DLDA and TensorLDA have the same objective function with different solutions. While 2DLDA obtains its solutions from an iterative algorithm, TensorLDA obtains the solutions by computing two independent subspaces and then combining both of them. Another method called discriminant analysis with tensor representation (DATER) was also proposed by Yan *et al*. to generalize classical LDA to tensor case [20]. In conclusion, methods from three levels of data representation are proposed to deal with dimensionality reduction problem, i.e., vector, matrix and tensor [21].

All of the methods mentioned above have an intrinsic limitation that they may distort the optimal directions of the projections and improve the sensitivity to outliers because they use the Frobenius norm as the basic measurement [22], [23]. To release this limitation, many methods based on L_1-norm were proposed because L_1-norm can enhance the robustness to outliers. The most classical method is sparse principle component analysis (SPCA) [24] and the sparse discriminant analysis (SDA) [25] which use L_1-norm on the sparsity penalty to obtain sparse projections. Inspired by the effectiveness of the SPCA and SDA, many other methods were also proposed [26], [28]. Besides, a method called rotational invariant L_1-norm based discriminant criterion (DCL_1) was proposed by Li *et al*. to efficiently enhance the separability of the between-class scatter and the compactness of the within-class scatter. DCL_1 utilizes the rotational invariant L_1-norm instead of the Frobenius norm [14] as the basic measurement. Wang *et al*. proposed Fisher discriminant analysis with L_1-norm to remodel the classical LDA by using L_1-norm with small sample size (SSS) problem and/or rank limitation circumvented [29]. Also, Zhong *et al*. proposed a method called linear discriminant analysis based on L_1-norm maximization with the purpose of maximizing the ration of the L_1-norm-based between-class scatter and the L_1-norm-based within-class scatter [30]. Although the idea in [29] and [30] is to obtain a set of optimal projections that can maximize the L_1-norm-based objective function, the optimization processes of these two methods are totally different [4].

Even though the above methods are able to reduce the sensitivity to outliers [31], they cannot obtain the joint sparsity for discriminant analysis [32]. Joint sparsity is derived from the new measurement technic called rotational invariance L_1-norm or $L_{2,1}-$ norm, which means that some rows of the projection matrix are all zeros [33], [34]. The non-zero elements represent the important features while the zero ones indicate the unimportant features, such that the important features are emphasized and the redundant features are filtered out. Recently, lots of work have presented the good performance of $L_{2,1}-$ norm in joint feature selection, such as [35]–[40]. In [41], Lai *et al*. proposed rotational invariant LDA (RILDA) to replace the Frobenius norm with $L_{2,1}-$ norm as the basic measurement on the scatter matrix so that the influence of outliers can be reduced.

Motivated by RILDA, in this paper, we propose a novel method called Locally Joint Sparse Marginal Embedding (LJSME) for feature extraction. LJSME is able to obtain the joint discriminant projections and avoid the SSS problem. Moreover, it can improve the robustness to outliers. The main contributions of LJSME lies in the following three folds:

1) The between-class scatter and within-class scatter are reconstructed by using the $L_{2,1}-$ norm instead of L_2- norm as the basic measurement. In this way, the sensitivity to outliers can be reduced. What is more, two weighted graphs are constructed and used so that the neighborhood relationship of the data can be preserved.
2) The $L_{2,1}-$ norm is used on the regularization term so that the jointly sparse projections can be obtained to improve the effectiveness of feature extraction.
3) The theoretic analysis including computational complexity and the convergence of the proposed method is discussed. Experimental results demonstrate the superiority of the proposed method.

The rest of this paper is organized as follows. The related works are reviewed in Section II while the proposed method and its optimal solution are presented in Section III. The theoretic analysis is presented in Section IV. Experiment results are shown in Section V and the conclusion of the paper is drawn in Section VI.

II. Related Works

In this section, we firstly give some notations and then briefly review several related works of the proposed method.

A. The Notations

In this paper, we denote scalars as lowercase or uppercase italic letters, i.e., i, j, n, C, etc. and vectors as bold italic letters, i.e., $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{v}$, etc., while bold uppercase italic letters are denoted the matrices, i.e., $\boldsymbol{X}$, $\boldsymbol{W}$, $\boldsymbol{S}$, etc. $tr(\cdot)$ denotes the trace of a matrix and $||\cdot||_p$ defines the $L_{\mathrm{p}}-$ norm of a matrix.

B. The Definition of $L_{2,1}$-Norm

Feature selection aims to obtain a small amount of features from the whole. For this sake, joint sparsity is a suitable way to achieve the purpose. Suppose each element in the projection matrix $\boldsymbol{Q}$ represents the weight of each feature. Then most of the elements in $\boldsymbol{Q}$ should be approximately equal to zero so that the non-zero elements can be highlighted. Using $L_{2,1}-$ norm on the projection matrix on the objective function can guarantee that the projection matrix is jointly sparse, which means some rows of the projection matrix are all zeros. The definition of $L_{2,1}-$ norm is as below

$$\|\boldsymbol{Q}\|_{2,1}=\sum_{i=1}^{n}\sqrt{\sum\nolimits_{j=1}^{m}\boldsymbol{Q}_{ij}^{2}}=\sum_{i=1}^{n}\left\|\boldsymbol{q}^{i}\right\|_{2}, \tag{1}$$

where $\boldsymbol{Q}_{ij}$ represents the element in $i-th$ row and $j-th$ column of $\boldsymbol{Q}$, $\boldsymbol{q}^i$ and $\boldsymbol{q}_j$ denote the $i-th$ row and j-th column of $\boldsymbol{Q}$, respectively. The $L_{2,1}-$norm holds rotational invariant property: $\|\boldsymbol{QR}\|_{2,1}=\|\boldsymbol{Q}\|_{2,1}$, where $\boldsymbol{R}$ is a rotational matrix [42].

C. The Classical LDA

Classical linear discriminant analysis [43], known as Fisher's linear discriminant, is a popular technique for linear dimensionality reduction and discriminant analysis. The purpose of LDA is to find a projection matrix $\boldsymbol{P}\in R^{d\times k}$ which can maximize the so-called Fisher criterion

$$J_F(\boldsymbol{P})=tr\left(\frac{\boldsymbol{P}^T\boldsymbol{S}_b\boldsymbol{P}}{\boldsymbol{P}^T\boldsymbol{S}_w\boldsymbol{P}}\right), \tag{2}$$

where $\boldsymbol{S}_b=\sum_{i=1}^{C}\rho_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T$ and $\boldsymbol{S}_w=\sum_{i=1}^{C}\rho_i\boldsymbol{S}_i$ are between-class scatter matrix and within-class scatter matrix, respectively. C denotes the class number, ρ_i is a priori probability of class i, $\boldsymbol{m}_i$ and $\boldsymbol{m}$ are mean vector of class i and mean vector over all the data, respectively. $\boldsymbol{S}_i$ represents the within-class scatter matrix of class i. The covariance matrix of the data is $\boldsymbol{S}_t=E[(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T]=\boldsymbol{S}_w+\boldsymbol{S}_b$ [44], where E denotes the expectation.

The optimal solution of classical LDA can be obtained by solving the standard eigen-decomposition problem which is related to $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$. The limitation of LDA is that $\boldsymbol{S}_w$ is required to be nonsingular because of the inverse operation, that is, the classical LDA has the SSS problem [44].

III. The Proposed Method

In this section, we first present the motivation of the proposed method. After that, we give our optimization problem and the corresponding optimal solution. In addition, we compare the proposed method with existing related methods and discuss their advantages and disadvantages.

A. Motivation

In order to solve SSS problem in LDA, many methods have been proposed and some of them can obtain good performance.

However, these methods still have several drawbacks. Firstly, even though some Frobenius norm based extensions of LDA are able to solve the SSS problem, they are sensitive to outliers due to the square operation in constructing the scatter matrix. Secondly, although some methods reconstruct the scatter matrix using L_1-norm instead of Frobenius norm to reduce the sensitivity to outliers, they do not consider the joint sparsity for feature extraction. That is, they use L_1-norm penalty (or have no sparsity penalty) on the regularization term, which cannot guarantee jointly sparse feature extraction. Last but not least, even though some methods consider joint sparsity, they ignore the robustness in designing the scatter matrices. For example, zhang *et al*. utilizes the $L_{2,1}-$norm regularization on the classical LDA [45] to obtain the joint sparsity, but they still use the Frobenius norm to construct scatter matrix.

Therefore, it is necessary to develop a new method that can guarantee joint sparsity for feature extraction and at the same time improve the robustness to outliners. In addition, since local structure is very important to discover the manifold structure for feature extraction and recognition [46]–[50], in this paper, we also incorporate locality graph of the data to redesign scatter matrices using $L_{2,1}-$norm as basic measurement.

B. The Objective Function

Motivated by the intuition that the data points nearby usually have the similar properties [36], we construct two weighted graphs $\boldsymbol{W}_b\in R^{n\times n}$, $\boldsymbol{W}_w\in R^{n\times n}$ for the between-class scatter and the within-class scatter, respectively. Suppose the data matrix is $\boldsymbol{X}\in R^{d\times n}$, where each column of $\boldsymbol{X}$ represents a sample. Let $\boldsymbol{y}$ be the sample label set, the between-class graph $\boldsymbol{W}_b$ is defined as

$$\boldsymbol{W}_{b,ij}=\begin{cases}1, & \text{if } [\boldsymbol{x}_i\in N_K(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j\in N_K(\boldsymbol{x}_i)] \text{ and } \boldsymbol{y}_i\neq\boldsymbol{y}_j\\ 0, & \text{otherwise}\end{cases},$$

and within-class graph $\boldsymbol{W}_w$ is denoted as

$$\boldsymbol{W}_{w,ij}=\begin{cases}1, & \text{if } [\boldsymbol{x}_i\in N_K(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j\in N_K(\boldsymbol{x}_i)] \text{ and } \boldsymbol{y}_i=\boldsymbol{y}_j\\ 0, & \text{otherwise}\end{cases},$$

where K is a constant, and $N_K(\boldsymbol{x}_i)$ represents K-nearest neighbor of $\boldsymbol{x}_i$. The elements in the weighted graphs reveal the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

We first propose Proposition 1 to present the generalization of constructing scatter with similarity between pair data by using $L_{2,1}-$norm as basic measurement, then give the objective function of the proposed method that based on Proposition 1.

Proposition 1: Suppose $\boldsymbol{X}\in R^{d\times n}$ is a data matrix where $\boldsymbol{x}_i,\boldsymbol{x}_j$ represents i-th and j-th sample in $\boldsymbol{X}$, and $\boldsymbol{U}\in R^{d\times k}$ is a projection matrix. Let $\boldsymbol{W}_w\in R^{n\times n}$ and $\boldsymbol{W}_b\in R^{n\times n}$ be the within-class and between-class graph of the data, respectively. Then, the sum of the weighted $L_{2,1}-$norm distance can be calculated via the trace of a matrix, i.e.,

a) $\sum_{i=1}^{n}\sum_{j=1}^{n}\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\|_2\boldsymbol{W}_{w,ij}=tr(\boldsymbol{U}^T\boldsymbol{X}_w^T\boldsymbol{D}_w\boldsymbol{X}_w\boldsymbol{U})$,

b) $\sum_{i=1}^{n}\sum_{j=1}^{n}\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\|_2\boldsymbol{W}_{b,ij}=tr(\boldsymbol{U}^T\boldsymbol{X}_b^T\boldsymbol{D}_b\boldsymbol{X}_b\boldsymbol{U})$,

where $\boldsymbol{X}_w\in R^{n^2\times d}$, $\boldsymbol{D}_w\in R^{n^2\times n^2}$, $\boldsymbol{X}_b\in R^{n^2\times d}$, $\boldsymbol{D}_b\in R^{n^2\times n^2}$.

Proof: The proof is in the appendix.

The equality (a) and (b) in Proposition 1 use the $L_{2,1}-$norm as the basic measurement to redesign the scatter matrices of data. They are finally converted to be the trace of matrix to define a new scatter of the data.

Based on Proposition 1, we propose the following objective function

$$\min_{\boldsymbol{U}}\ tr(\boldsymbol{U}^T\boldsymbol{X}_w^T\boldsymbol{D}_w\boldsymbol{X}_w\boldsymbol{U})-\alpha tr(\boldsymbol{U}^T\boldsymbol{X}_b^T\boldsymbol{D}_b\boldsymbol{X}_b\boldsymbol{U})+\beta\|\boldsymbol{U}\|_{2,1}$$
$$\text{s.t. } \boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}. \tag{3}$$

That is,

$$\min_{\boldsymbol{U}}\quad tr(\boldsymbol{U}^T\boldsymbol{S}_{2,1}^w\boldsymbol{U})-\alpha tr(\boldsymbol{U}^T\boldsymbol{S}_{2,1}^b\boldsymbol{U})+\beta\|\boldsymbol{U}\|_{2,1}$$
$$\text{s.t.}\quad \boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}, \tag{4}$$

where $\boldsymbol{S}_{2,1}^{b} = \boldsymbol{X}_b^T \boldsymbol{D}_b \boldsymbol{X}_b \in R^{d\times d}$and $\boldsymbol{S}_{2,1}^{w} = \boldsymbol{X}_w^T \boldsymbol{D}_w \boldsymbol{X}_w \in R^{d\times d}$ are the $L_{2,1}-$ norm based between-class scatter and the $L_{2,1}-$ norm based within-class scatter, respectively. $\boldsymbol{U} \in R^{d\times k}$ is the projection matrix which is supposed to be column-orthonormal, d is the dimensionality of the data, k is the objective number of the projections. α and β are parameters to balance the three terms.

Similar to [51], [52], the graph embedding concept is also applied in our proposed method. The first and second term in (4) is the maximum margin criterion (MMC) [53] based discriminant analysis. The third term is the regularization term which is able to guarantee the joint sparsity for the projections. By requiring $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$, i.e., the projections are orthonormal, the proposed method can preserve the shape of the distribution of data [54]. Compared with the classical LDA, the proposed method based on MMC does not need to compute the inverse of the within-class scatter so that it can avoid the SSS problem.

C. The Optimal Solution

According to (3) and (4), we have

$$\begin{aligned} &tr(\boldsymbol{U}^T \boldsymbol{S}_{2,1}^{w} \boldsymbol{U}) - \alpha tr(\boldsymbol{U}^T \boldsymbol{S}_{2,1}^{b} \boldsymbol{U}) + \beta \|\boldsymbol{U}\|_{2,1} \\ &\quad = tr(\boldsymbol{U}^T (\boldsymbol{S}_{2,1}^{w} - \alpha \boldsymbol{S}_{2,1}^{b} + \beta \boldsymbol{D}) \boldsymbol{U}), \end{aligned} \tag{5}$$

and

$$\begin{aligned} &tr(\boldsymbol{U}^T \boldsymbol{S}_{2,1}^{w} \boldsymbol{U}) - \alpha tr(\boldsymbol{U}^T \boldsymbol{S}_{2,1}^{b} \boldsymbol{U}) + \beta \|\boldsymbol{U}\|_{2,1} \\ &\quad = tr(\boldsymbol{U}^T \boldsymbol{X}_w^T \boldsymbol{D}_w \boldsymbol{X}_w \boldsymbol{U}) \\ &\qquad - \alpha tr(\boldsymbol{U}^T \boldsymbol{X}_b^T \boldsymbol{D}_b \boldsymbol{X}_b \boldsymbol{U}) + \beta tr(\boldsymbol{U}^T \boldsymbol{D} \boldsymbol{U}), \end{aligned} \tag{6}$$

where $\boldsymbol{D} \in R^{d\times d}$ is a diagonal matrix with its diagonal elements [38]

$$\boldsymbol{D}_{ii} = \frac{1}{2\|\boldsymbol{u}^i\|_2}. \tag{7}$$

The following Lemma 1 is presented to help to obtain the optimal solution of the objective function.

Lemma 1: Suppose $\boldsymbol{G} \in R^{d\times d}$ and $\boldsymbol{H} \in R^{d\times d}$ are symmetric matrices, the optimal solution of the following problem

$$\min_{\boldsymbol{Z}} \frac{tr(\boldsymbol{Z}^T \boldsymbol{G} \boldsymbol{Z})}{tr(\boldsymbol{Z}^T \boldsymbol{H} \boldsymbol{Z})} \tag{8}$$

comes from the matrix $\boldsymbol{Z}^* \in R^{d\times k}$ with columns are the eigenvectors corresponding to the k smallest eigenvalues of the standard eigen-decomposition $\boldsymbol{G}\boldsymbol{z}_i = \lambda \boldsymbol{H}\boldsymbol{z}_i, i = 1, 2, \ldots, k$.

Proof: The proof is omitted due to limited space, which is similar to [55].

By integrating the constraint condition, the optimization problem in (4) can be converted to [55]

$$\min_{\boldsymbol{U}} \frac{Tr(\boldsymbol{U}^T (\boldsymbol{S}_{2,1}^{w} - \alpha \boldsymbol{S}_{2,1}^{b} + \beta \boldsymbol{D}) \boldsymbol{U})}{\mathrm{tr}(\boldsymbol{U}^T \boldsymbol{U})}. \tag{9}$$

From Lemma 1, the optimal solution is easy to be obtained by solving the standard eigen-decomposition:

$$(\boldsymbol{S}_{2,1}^{w} - \alpha \boldsymbol{S}_{2,1}^{b} + \beta \boldsymbol{D}) \boldsymbol{U}_i = \lambda \boldsymbol{U}_i. \tag{10}$$

We can see that the optimal solution $\boldsymbol{U}^* = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k]$ is composed of first k eigenvectors corresponding to the smallest $k$eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ from the eigen-decomposition of $(\boldsymbol{S}_{2,1}^{b} - \alpha \boldsymbol{S}_{2,1}^{w} + \beta \boldsymbol{D})$. Since the optimal solution cannot be obtained by only singular step of eigen-decomposition due to the need of updating the variable $\boldsymbol{D}$, we propose an iterative algorithm to obtain the optimal solution. The detail of the algorithm is shown in Table I.

D. The Comparison and Discussion

In this subsection, we present comparison and discussion between the proposed LJSME and the most related $L_{2,1}$-norm methods that based on discriminant analysis and regression, respectively. The objectives of all these methods are summarized in Table II.

1) Comparison With $L_{2,1}$-Norm Methods Based on Discriminant Analysis: It is known that the classical LDA aims to seek a series of projections by which the between-class scatter can be maximized and the within-class scatter can be minimized. The optimal solution comes from the egien-decomposition problem. But this is only fit for the situation when the within-class scatter matrix is not singular, that is, it has the intrinsic limitation (SSS problem). As mentioned in Section I, many methods based on discriminant analysis have been proposed to overcome the limitation in the LDA and to improve the performance of feature selection and extraction [12], [13], [16]–[56].

FLDA via joint $L_{2,1}-$ norm (L21FLDA) [57] avoids the SSS problem by first transforming the within-class scatter matrix into a nonsingular matrix by PCA and then conducting singular value decomposition as FisherFace [56]. It also obtains the joint sparsity to improve the performance of feature selection by adding $L_{2,1}-$ norm regularization term on the objective function. Since the scatter matrices of L21FLDA are constructed by using traditional L_2- norm as basic measurement, it is not robust to outliers. To enhance the robustness, another method called rotational invariant LDA (RILDA) [41] was proposed. RILDA reconstructs the scatter matrices with $L_{2,1}-$ norm instead of L_2- norm to reduce the sensitivity to outliers.

Unfortunately, the above discriminant analysis methods still have two main drawbacks. First, they cannot simultaneously solve SSS problem and enhance the robustness to outliers as well as obtain joint sparsity to improve the performance of feature selection and extraction. Second, they ignore the neighborhood structure of the data which is important for recognition task as shown in [46]–[48]. Different from these methods, the proposed LJSME do not have such drawbacks. Compared with L21FLDA, RILDA and other discriminant analysis methods whose objective functions are based on the form of Fisher discriminative criterion (FDC), the objective function of LJSME is based on the form of MMC, by which the SSS problem can be avoided. In addition, LJSME incorporates the idea of locality graph to weight each data pair, so that the neighborhood relationship of the data is preserved.

2) Comparison With $L_{2,1}$-Norm Methods Based on Regression: Recently, $L_{2,1}-$ norm is widely used in regression based methods to improve performance of classification. As we can

TABLE I
THE LJSME ALGORITHM

Input: The sample matrix $\boldsymbol{X} \in R^{d\times n}$, the label set $\boldsymbol{y}$, the parameter α, β, the objective number of projections k, the iteration number T
Output: The feature selection matrix $\boldsymbol{U}$

Step 1: Construct similarity graph $\boldsymbol{W}_b$, $\boldsymbol{W}_w$, $\boldsymbol{X}_w$, $\boldsymbol{X}_b$
Step 2: Initialize $\boldsymbol{U}$ as column-orthogonal matrix with size $d\times k$, compute $\boldsymbol{D}$ using (7)
Step 3: For $i=1:T$ do
- Construct matrix, $\boldsymbol{D}_w$, $\boldsymbol{D}_b$, compute $\boldsymbol{S}_{2,1}^w$, $\boldsymbol{S}_{2,1}^b$
- Solve the eigen-decomposition problem of (10) to obtain eigenvectors $\boldsymbol{U}=(\boldsymbol{u}_1, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_k)$
- Update $\boldsymbol{D}$ using (7)
Step 4: Output the feature selection matrix $\boldsymbol{U}$ for further classification

TABLE II
THE SUMMARY OF OBJECTIVES OF THE PROPOSED METHOD AND THE MOST RELATED METHODS

Methods	Objectives	Variables	Remarks
RFS [38]	$\min_{\boldsymbol{W}} \frac{1}{\gamma}\|\boldsymbol{X}^T\boldsymbol{W}-\boldsymbol{Y}\|_{2,1}+\|\boldsymbol{W}\|_{2,1}$	$\boldsymbol{W}\in R^{d\times c}$	$L_{2,1}$-norm minimization on loss function and regularization term
UDFS [40]	$\min_{\boldsymbol{W}^T\boldsymbol{W}=I_{d\times d}} tr(\boldsymbol{W}^T\boldsymbol{M}\boldsymbol{W})+\gamma\|\boldsymbol{W}\|_{2,1}$	$\boldsymbol{W}\in R^{d\times c}$	$L_{2,1}$-norm used on regularization term
CRFS [61]	$\min_{\boldsymbol{W}}\{1-\sum_{k=1}^{n}\exp(-\frac{\|(\boldsymbol{X}^T\boldsymbol{W}-\boldsymbol{Y})^k\|_2^2}{\sigma^2})+\|\boldsymbol{W}\|_{2,1}\}$	$\boldsymbol{W}\in R^{d\times c}$	$L_{2,1}$-norm used on regularization term
JELSR [59]	$\min_{\boldsymbol{U},\boldsymbol{Y}\boldsymbol{Y}^T=I_{k\times k}} tr(\boldsymbol{Y}\boldsymbol{L}\boldsymbol{Y}^T)+\beta(\|\boldsymbol{U}^T\boldsymbol{X}-\boldsymbol{Y}\|_2^2+\alpha\|\boldsymbol{U}\|_{r,p}^p)$	$\boldsymbol{Y}\in R^{k\times n}$ $\boldsymbol{U}\in R^{d\times k}$	$L_{2,1}$-norm used on regularization term, when r=2, p=1
RILDA [41]	$\min_{\boldsymbol{U}^T\boldsymbol{U}=I}\frac{tr(\boldsymbol{U}^T\boldsymbol{X}_{Rw}\boldsymbol{D}_{Rw}\boldsymbol{X}_{Rw}^T\boldsymbol{U})}{tr(\boldsymbol{U}^T\boldsymbol{X}_{Rb}\boldsymbol{D}_{Rb}\boldsymbol{X}_{Rb}^T\boldsymbol{U})}$	$\boldsymbol{U}\in R^{d\times k}$	$L_{2,1}$-norm based construction on scatter matrice
GRR [58]	$\min_{\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I},\boldsymbol{U},\boldsymbol{h}} \sum_i\sum_j\|\boldsymbol{x}_i^T\boldsymbol{U}\boldsymbol{A}^T-\boldsymbol{x}_j^T\boldsymbol{U}\boldsymbol{A}^T\|_2\boldsymbol{W}_{ij}+\beta\|\boldsymbol{U}\|_{2,1}$ $+\gamma\|\boldsymbol{X}^T\boldsymbol{U}\boldsymbol{A}^T+\boldsymbol{1}\boldsymbol{h}^T-\boldsymbol{Y}\|_{2,1}+\lambda\|\boldsymbol{h}\|_2^2$	$\boldsymbol{U}\in R^{d\times k}$ $\boldsymbol{A}\in R^{c\times k}$ $\boldsymbol{h}\in R^{c\times 1}$	$L_{2,1}$-norm minimization on loss function and regularization term
LJSME	$\min_{\boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}} tr(\boldsymbol{U}^T\boldsymbol{S}_{2,1}^w\boldsymbol{U})-\alpha tr(\boldsymbol{U}^T\boldsymbol{S}_{2,1}^b\boldsymbol{U})+\beta\|\boldsymbol{U}\|_{2,1}$	$\boldsymbol{U}\in R^{d\times k}$	$L_{2,1}$-norm based reconstruction on scatter matrices with locality graph; $L_{2,1}$-norm used on regularization term

see from Table II, robust feature selection (RFS) [38] and generalized robust regression (GRR) [58] use $L_{2,1}-$norm as the basic measurement on the loss function and the regularization term, while unsupervised discriminative feature selection (UDFS) [40] and $L_{2,1}-$norm regularized correntropy for robust feature selection (CRFS) only apply $L_{2,1}-$norm penalty on the objective functions. For joint embedding learning and sparse regression (JELSR) [59], we can say that it also utilizes $L_{2,1}-$norm on the regularization term when the variables r and p are set as 2 and 1, respectively. Based on the objectives in Table II, we can find that the common property among RFS, GRR, UDFS, CRFS, JELSR and the proposed LJSME is that they all apply $L_{2,1}-$norm penalty on the objective function, by which the jointly sparse projections can be obtained.

However, there are substantial difference between the proposed LJSME and these $L_{2,1}-$norm based regression methods. For example, even though GRR is also proposed for feature selection and extraction, the motivation and focus of GRR and LJSME are totally different. GRR is an extension of ridge regression (RR) and aims to solve the intrinsic small-class problem in RR. However, LJSME is different. The basic idea of LJSME is for robust jointly sparse discriminant analysis instead of ridge regression learning as in GRR. LJSME does not focus on solving the intrinsic drawbacks in RR. Instead, it introduces the locality graph to redesign the within-class scatter matrix and between-class scatter matrix in classical LDA so as to obtain discriminant information to improve the performance of feature selection and extraction. Moreover, the $L_{2,1}-$norm based reconstruction on the scatter matrices makes LJSME more robust to outliers.

For short, most of current $L_{2,1}-$norm methods either do not consider the compactness of the same classes and the separability of different classes, or do not consider the robustness as well as neighborhood relationship of data. In contrast, LJSME based on the concept of discriminant analysis can maximize the between-class scatter and minimize the within-class scatter. In addition, it can enhance the robustness to outliers and preserve

neighborhood relationship of data by introducing the locality graph to redesign the scatter matrices with robust $L_{2,1}-$ norm instead of traditional L_2- norm. All these properties make the proposed LJSME different from current $L_{2,1}-$ norm based methods. Such superiority of LJSME can be further verified from the experimental results in Section V.

IV. Theoretic Analysis

In this section, we will first discuss the computational complexity of the proposed iterative algorithm in Table I, and then present the convergence of the proposed method.

A. Computational Complexity

Since the main computational complexity of the proposed algorithm is to solve the eigen-decomposition problem in (10), we mainly analyze it and present it as final complexity. Suppose the algorithm needs T steps. Solving the eigen-decomposition problem in (10) needs $O(d^3)$ and it is easy to know that the final computational complexity is$O(Td^3)$, where d is the dimensionality of the sample. Even though the computational cost of the proposed method is larger than the classical LDA, experimental results demonstrate that the iteration number of the algorithm is usually small. Therefore, we can still say that the proposed algorithm is efficient.

B. The Convergence

Since the proposed optimization problem is solved by the iterative algorithm proposed in Table I, it is necessary to prove the convergence. For this sake, we first have the following Lemma.

Lemma 2: For any non-zero vectors $\boldsymbol{a}, \boldsymbol{b} \in R^d$, the following inequality holds

$$\|\boldsymbol{a}\|_2 - \frac{\|\boldsymbol{a}\|_2^2}{2\|\boldsymbol{b}\|_2} \leq \|\boldsymbol{b}\|_2 - \frac{\|\boldsymbol{b}\|_2^2}{2\|\boldsymbol{b}\|_2}. \tag{11}$$

The convergence of the proposed LJSME can be summarized in the theorem below.

Theorem 1: The proposed algorithm will monotonically decrease the objective function in (3) or (4) in each iteration and finally obtain local optimal solution.

Proof: For simplicity, we present the objective function in (3) as $F(\boldsymbol{D}_w, \boldsymbol{D}_b, \boldsymbol{U}, \boldsymbol{D})$. Supposed in the $(t-1)$-th iteration, all the variable are obtained and in the $t-th$ iteration, since $\boldsymbol{U}$ is updated from (10), it goes

$$\begin{aligned} &F((\boldsymbol{D}_w)_{t-1}, (\boldsymbol{D}_b)_{t-1}, \boldsymbol{U}_t, \boldsymbol{D}_{t-1}) \\ &\quad \leq F((\boldsymbol{D}_w)_{t-1}, (\boldsymbol{D}_b)_{t-1}, \boldsymbol{U}_{t-1}, \boldsymbol{D}_{t-1}). \end{aligned} \tag{12}$$

That is,

$$\begin{aligned} &tr(\boldsymbol{U}_t^T \boldsymbol{X}_w^T (\boldsymbol{D}_w)_{t-1} \boldsymbol{X}_w \boldsymbol{U}_t) \\ &\quad - \alpha tr(\boldsymbol{U}_t^T \boldsymbol{X}_b^T (\boldsymbol{D}_b)_{t-1} \boldsymbol{X}_b \boldsymbol{U}_t) + \beta tr(\boldsymbol{U}_t^T \boldsymbol{D}_{t-1} \boldsymbol{U}_t) \\ &\quad \leq tr(\boldsymbol{U}_{t-1}^T \boldsymbol{X}_w^T (\boldsymbol{D}_w)_{t-1} \boldsymbol{X}_w \boldsymbol{U}_{t-1}) \\ &\quad - \alpha tr(\boldsymbol{U}_{t-1}^T \boldsymbol{X}_b^T (\boldsymbol{D}_b)_{t-1} \boldsymbol{X}_b \boldsymbol{U}_{t-1}) + \beta tr(\boldsymbol{U}_{t-1}^T \boldsymbol{D}_{t-1} \boldsymbol{U}_{t-1}), \end{aligned} \tag{13}$$

it goes,

$$\begin{aligned} &\sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_w \boldsymbol{U}_t)^i\|_2^2}{2\|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2} - \alpha \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_b \boldsymbol{U}_t)^i\|_2^2}{2\|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2} \\ &\quad + \beta \sum_{i=1}^{d} \frac{\|\boldsymbol{u}_t^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2} \leq \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2^2}{2\|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2} \\ &\quad - \alpha \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2^2}{2\|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2} + \beta \sum_{i=1}^{d} \frac{\|\boldsymbol{u}_{t-1}^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2}. \end{aligned} \tag{14}$$

Since the definition of $L_{2,1}-$ norm of matrix $\boldsymbol{A}$ is $\|\boldsymbol{A}\|_{2,1} = \sum \|\boldsymbol{a}^i\|_2$, the above inequality holds

$$\begin{aligned} &\sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_t)^i\|_2 - \left(\sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_t)^i\|_2 - \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_w \boldsymbol{U}_t)^i\|_2^2}{2\|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2} \right) \\ &\quad - \alpha \sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_b)^i\|_2 + \alpha \left(\sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_t)^i\|_2 \right. \\ &\quad \left. - \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_b \boldsymbol{U}_t)^i\|_2^2}{2\|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2} \right) + \beta \sum_{i=1}^{d} \frac{\|\boldsymbol{u}_t^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2} \\ &\leq \sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2 - \left(\sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2 \right. \\ &\quad \left. - \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2^2}{2\|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2} \right) \\ &\quad - \alpha \sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2 + \alpha \left(\sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2 \right. \\ &\quad \left. - \sum_{i=1}^{n} \frac{\|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2^2}{2\|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2} \right) + \beta \sum_{i=1}^{d} \frac{\|\boldsymbol{u}_{t-1}^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2}. \end{aligned} \tag{15}$$

In (15), we can always find proper values for α and β such that the following inequality holds from Lemma 2:

$$\begin{aligned} &\sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_t)^i\|_2 - \alpha \sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_t)^i\|_2 + \beta \sum_{i=1}^{d} \frac{\|\boldsymbol{u}_t^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2} \\ &\leq \sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2 - \alpha \sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2 + \beta \sum_{i=1}^{d} \frac{\|\boldsymbol{u}_{t-1}^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2}. \end{aligned} \tag{16}$$

That is,

$$\begin{aligned} &\sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_t)^i\|_2 - \alpha \sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_t)^i\|_2 + \beta \|\boldsymbol{U}_t\|_{2,1} \\ &\quad - \beta \sum_{i=1}^{d} \left(\|\boldsymbol{u}_t^i\|_2 - \frac{\|\boldsymbol{u}_t^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2} \right) \leq \sum_{i=1}^{n} \|(\boldsymbol{X}_w \boldsymbol{U}_{t-1})^i\|_2 \\ &\quad - \alpha \sum_{i=1}^{n} \|(\boldsymbol{X}_b \boldsymbol{U}_{t-1})^i\|_2 + \beta \|\boldsymbol{U}_{t-1}\|_{2,1} \\ &\quad - \beta \sum_{i=1}^{d} \left(\|\boldsymbol{u}_{t-1}^i\|_2 - \frac{\|\boldsymbol{u}_{t-1}^i\|_2^2}{2\|\boldsymbol{u}_{t-1}^i\|_2} \right). \end{aligned} \tag{17}$$

According to Lemma 2, we have

$$\left\|\boldsymbol{u}_t^i\right\|_2 - \frac{\left\|\boldsymbol{u}_t^i\right\|_2^2}{2\left\|\boldsymbol{u}_{t-1}^i\right\|_2} \leq \left\|\boldsymbol{u}_{t-1}^i\right\|_2 - \frac{\left\|\boldsymbol{u}_{t-1}^i\right\|_2^2}{2\left\|\boldsymbol{u}_{t-1}^i\right\|_2}. \tag{18}$$

Then, the following inequality holds

$$\begin{aligned}&\sum_{i=1}^{n}\|(\boldsymbol{X}_w\boldsymbol{U}_t)^i\|_2 - \alpha\sum_{i=1}^{n}\|(\boldsymbol{X}_b\boldsymbol{U}_t)^i\|_2 + \beta\|\boldsymbol{U}_t\|_{2,1}\\ &\leq \sum_{i=1}^{n}\|(\boldsymbol{X}_w\boldsymbol{U}_{t-1})^i\|_2 - \alpha\sum_{i=1}^{n}\|(\boldsymbol{X}_b\boldsymbol{U}_{t-1})^i\|_2 + \beta\|\boldsymbol{U}_{t-1}\|_{2,1}.\end{aligned} \tag{19}$$

That is,

$$\begin{aligned}&tr\left(\boldsymbol{U}_t^T\boldsymbol{X}_w^T(\boldsymbol{D}_w)_t\boldsymbol{X}_w\boldsymbol{U}_t\right) - \alpha tr\left(\boldsymbol{U}_t^T\boldsymbol{X}_b^T(\boldsymbol{D}_b)_t\boldsymbol{X}_b\boldsymbol{U}_t\right)\\ &+ \beta tr\left(\boldsymbol{U}_t^T\boldsymbol{D}_t\boldsymbol{U}_t^T\right) \leq tr\left(\boldsymbol{U}_{t-1}^T\boldsymbol{X}_w^T(\boldsymbol{D}_w)_{t-1}\boldsymbol{X}_w\boldsymbol{U}_{t-1}\right)\\ &-\alpha tr\left(\boldsymbol{U}_{t-1}^T\boldsymbol{X}_b^T(\boldsymbol{D}_b)_{t-1}\boldsymbol{X}_b\boldsymbol{U}_{t-1}\right) + \beta tr\left(\boldsymbol{U}_{t-1}^T\boldsymbol{D}_{t-1}\boldsymbol{U}_{t-1}^T\right).\end{aligned} \tag{20}$$

It goes,

$$\begin{aligned}&F((\boldsymbol{D}_w)_t, (\boldsymbol{D}_b)_t, \boldsymbol{U}_t, \boldsymbol{D}_t)\\ &\leq F((\boldsymbol{D}_w)_{t-1}, (\boldsymbol{D}_b)_{t-1}, \boldsymbol{U}_{t-1}, \boldsymbol{D}_{t-1}).\end{aligned} \tag{21}$$

The inequality (21) indicates that the objective function value of (3) will monotonically decrease by the iterative algorithm in Table I. For the proof of the convergence of (4), we can always use the similar proof by setting $\boldsymbol{S}_{2,1}^b = \boldsymbol{X}_b^T\boldsymbol{D}_b\boldsymbol{X}_b$ and $\boldsymbol{S}_{2,1}^w = \boldsymbol{X}_w^T\boldsymbol{D}_w\boldsymbol{X}_w$. Therefore, the proof of Theorem 1 is completed. ■

V. Experiment

In this section, experiments on five well-known databases were conducted to evaluate the performance of the proposed method in different cases, including various facial expressions and lighting conditions on face images and hyperspectral images. Some related methods are used for comparison. These methods include the classical component analysis methods (PCA, SPCA [24]), the marginal fisher analysis method (MFA [60]), the methods based on discriminant analysis (LDA [56], RILDA [41]), the methods based on $L_{2,1}-$ norm (UDFS [40], RFS [38], JELSR [59], CRFS [61]) and the state-of-the-art hyperspectral image dimensionality reduction methods (superPCA and superMPCA) [62]. In all experiments, the raw data are used as input for classification and this plays a role as baseline.

Experimental Settings: On each dataset, the images are divided into two parts, i.e., the gallery set and the probe set. The gallery set contains l (l is less than the number of images of each class) images of each class while the probe set is composed of the remaining images.

Since the dimensionality of the images is very high, it would cause unstable performance for the compared methods. For example, when the dimensionality of data is high but the number of training samples is small, the scatter matrices in LDA would be singular. Therefore, we need to preprocess the data. Base on this regard, the experiment is designed into three steps: First, PCA is used as preprocessing to reduce the dimensionality of the original data on AR, PIE, PaviaU, PolyU-HSFD and LFW database to dimension of 150, 200, 60, 150 and 200, respectively. Second, all methods are used to conduct feature selection or extraction. Finally, the nearest neighbor classifier (NN) or other classifiers are used for classification and the recognition rates are presented.

Fig. 1. Sample images on AR database.

Exploration of the parameters: We analyze the values of parameters α and β in the area of $[10^{-9}, 10^{-6}, \ldots, 10^{9}]$on each dataset. The higher average recognition rate corresponding to the values are the optimal parameter values for the proposed method. For the compared methods, we set their parameter values according to the original introduction. For example, both UDFS and RFS are reported to obtain the optimal performance with their parameters lying in the area of $[10^{-3}, 10^{-2}, \ldots, 10^{3}]$. Accordingly, those values are used in the experiments.

A. The Experiment on AR Database

A subsection of AR database [63] containing 2,400 images with size of 50×40 from 120 people are used in this experiment. The experiment is performed to test the performance of LJSME on the occasion when images are varying with facial expression and lighting conditions. The sample images of one individual on AR face database are shown in Fig. 1. The size and amount of images on different databases are listed in Table III.

In this experiment, we first explore the optimal parameter values of the proposed method. The recognition rate versus the variations of α and β is shown in Fig. 3(a). From Fig. 3(a), we can know that the optimal values of α and β are $[10^{-9}, 10^{0}]$, $[10^{-9}, 10^{3}]$, respectively.

In order to verify the effectiveness of the improvements in the objective function in (4), we explore three variants of the proposed method.

Variant 1: variant 1 is the case when $\alpha = 0$ and $\beta = 0$ in the objective function in (4).

Variant 2: variant 2 is the case when $\alpha \neq 0$ but $\beta = 0$. **Variant 3**: variant 3 is the case when $\alpha = 0$ but $\beta \neq 0$.

The performance of LJSME and variant 1, 2, 3 are shown in Fig. 3(b). From the comparison, we can know that variant 2 performs better than both variant 1 and variant 3. It indicates that the combination of first part and second part in the objective function of LJSME is usually more effective than one of them with the third part. However, only when $\alpha \neq 0$ and $\beta \neq 0$ and they are set as proper values, can the three parts work as a whole to obtain better performance. In another world, the improvements in the proposed method is effective and necessary.

In addition, several classifiers (i.e., nearest neighbor classifier (NN), K nearest neighbor classifier (KNN), support vector machine classifier (SVM) and random forest classifier (RF)) are

TABLE III
THE DESCRIPTION OF INPUT IMAGES ON ALL DATABASE

Database	AR	PIE	PaviaU	PolyU-HSFD	LFW
Image amount	2,400	1,632	103	1,410	4,324
Image size	50×40	32×32	610×340	44×36	112×96

TABLE IV
THE RECOGNITION RATE (%) OF DIFFERENT METHODS UNDER DIFFERENT CLASSIFIERS ON AR DATABASE

AR(l =5)	NN	KNN (K=3)	SVM	RF
Baseline	84.14	82.75	93.95	**90.00**
PCA	83.33	82.05	93.90	82.02
LDA	90.72	88.78	91.50	77.94
SPCA	83.34	82.07	93.91	82.29
RILDA	94.58	91.42	94.81	84.46
MFA	93.62	89.90	93.51	79.72
UDFS	83.57	82.05	93.90	84.44
RFS	94.22	92.92	89.32	87.20
JELSR	83.38	82.05	93.90	82.12
CRFS	93.33	**93.67**	91.28	89.00
LJSME	**94.91**	91.18	**95.33**	82.67

Fig. 2. Sample images on CMU PIE database.

used for classification in this experiment to evaluate the performance of the proposed LJSME. Table IV lists the average recognition rate of different methods based on 10 times. From Table IV, we can know that LJSME is effective in most cases. When using RF as classifier, the performance of most methods is inferior to the baseline. The potential reason is that RF classifier can handle high dimensional data without any feature selection/extraction and it is robust to outlier which benefits from its intrinsic property. The other reason might be that these feature extraction methods do not match the RF's classification principle so that they get low recognition rates.

The average recognition rate versus the dimension of all methods with $l = 3$ is shown in Fig. 3(c). Table V lists the maximal average recognition rates and the corresponding standard deviations and dimensions of each method.

From Fig. 3(c) and Table V, we can know that the proposed LJSME and RILDA as well as LDA can obtain good performance. This indicates the effectiveness of discriminant analysis in feature extraction and classification.

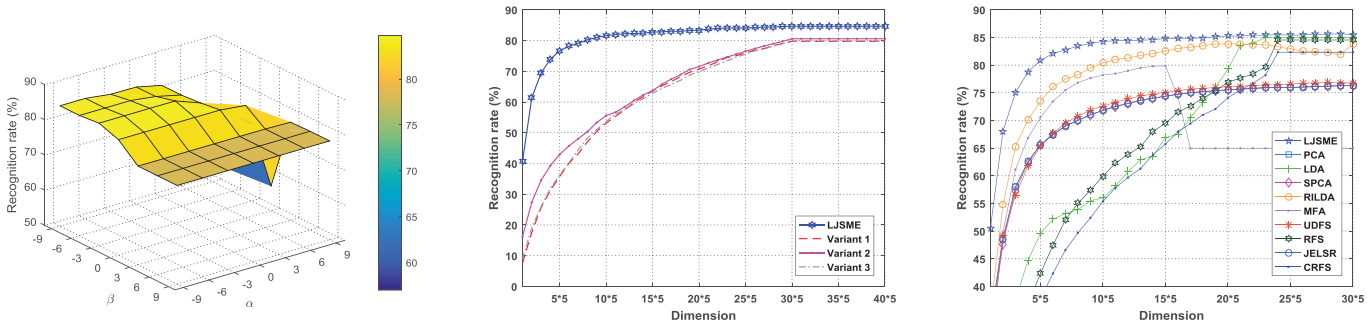
B. The Experiment on PIE Database

In order to evaluate the effectiveness of the proposed LJSME when there are pose variations and facial expressions on facial images, a subset (C29) from (PIE) [64] dataset is used in this experiment. C29 has 1,632 images with size of 32×32 from 68 people. The sample images of this dataset are shown in Fig. 2.

From Fig. 4(a), it is obvious that the optimal value of α and β are $[10^{-9}, 10^{-3}]$ and $[10^{-9}, 10^{3}]$, respectively.

To explore the performance of the proposed method while using different weight measurement in constructing the locality graph, related experiments were conducted. On this database, three weight measurements, i.e., K-nearest neighbor hard weight, Euclidean distance and heat kernel, are used. Fig. 4(b) presents the recognition rate of the proposed method under the three cases. The results indicate that using Euclidean distance as basic measurement in constructing locality graph is not as effective as using K-nearest neighbor hard weight and heat kernel. The potential reason for this phenomenon is that Euclidean distance measurement may provide a larger weight to the point pairs when there are noise/variation in face images since the images on PIE dataset is usually various in poses and facial expressions. Even so, the performance of LJSME is satisfactory overall no matter which graph is used. As such, we always use the KNN hard weight in this paper.

The proposed LJSME is a discriminant analysis method based on maximum margin criterion (MMC) and incorporates locality graph to redesign scatter matrices by using $L_{2,1}-$ norm as basic measurement (i.e., MMC + Graph + $L_{2,1}$). Since Fisher Discriminative Criterion (FDC) is usually used in many discriminant analysis methods, we conduct related experiments to evaluate the performance of the proposed method based on FDC (i.e., FDC + Graph + $L_{2,1}$). Fig. 4(c) presents the performance

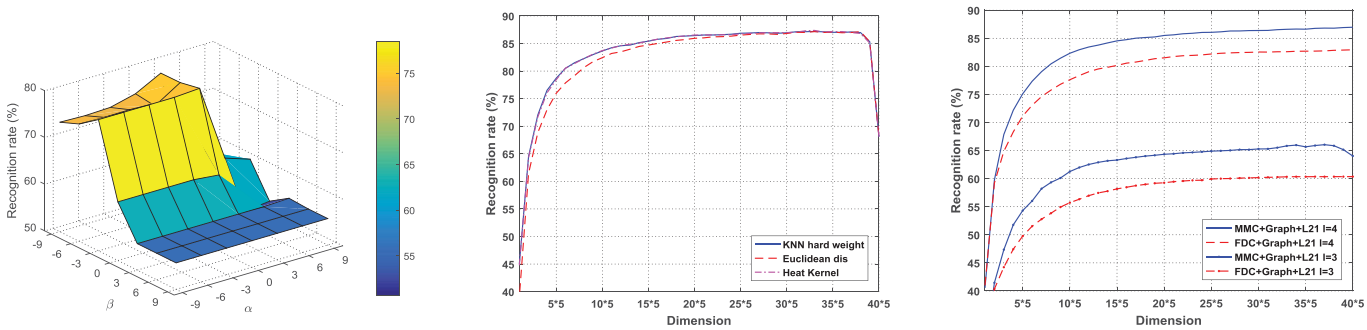


(a) (b) (c)

Fig. 3. (a) Recognition rate versus α and β on AR face database. (b) The recognition rate of different variants of LJSME. (c) The recognition rate versus dimension of all methods on the AR database with $l = 3$.

TABLE V
THE MAXIMAL RECOGNITION RATE (%), STANDARD DEVIATION, DIMENSION OF DIFFERENT METHODS ON AR DATABASE

l	Baseline	PCA	LDA	SPCA	RILDA	MFA	UDFS	RFS	JELSR	CRFS	**LJSME**
2	71.32 ±5.05	71.03 ±5.00 200	81.11 ±9.48 115	71.02 ±5.04 200	55.50 ±9.05 115	79.06 ±10.02 65	71.20 ±5.01 180	80.74 ±12.00 120	71.03 ±4.96 200	78.25 ±10.98 120	**81.30 ±8.34 135**
3	77.23 ±5.44	76.70 ±5.26 200	84.95 ±9.50 115	76.70 ±5.27 200	83.83 ±9.97 100	79.86 ±12.64 75	76.92 ±5.12 180	84.54 ±12.37 120	76.70 ±5.25 200	82.33 ±12.49 120	**85.64 ±10.01 145**
4	83.45 ±5.08	82.72 ±5.27 190	90.18 ±6.37 115	82.74 ±5.25 195	92.52 ±6.82 85	90.93 ±7.18 80	82.91 ±5.15 160	92.94 ±7.45 120	82.75 ±5.25 195	91.58 ±7.51 120	**93.14 ±6.17 150**
5	84.14 ±5.14	83.33 ±5.11 195	90.72 ±7.27 115	83.34 ±5.11 200	94.58 ±7.58 80	93.62 ±7.54 85	83.57 ±4.98 170	94.22 ±7.95 120	83.38 ±5.12 195	93.33 ±8.20 120	**94.91 ±6.89 75**



(a) (b) (c)

Fig. 4. (a) Recognition rate versus α and β on PIE face database, (b) The performance of LJSME with different weight measurements on locality graph on the PIE database with $l = 4$, (c) Recognition rate of LJSME with MMC/ FDC on PIE face database with $l = 3, 4$.

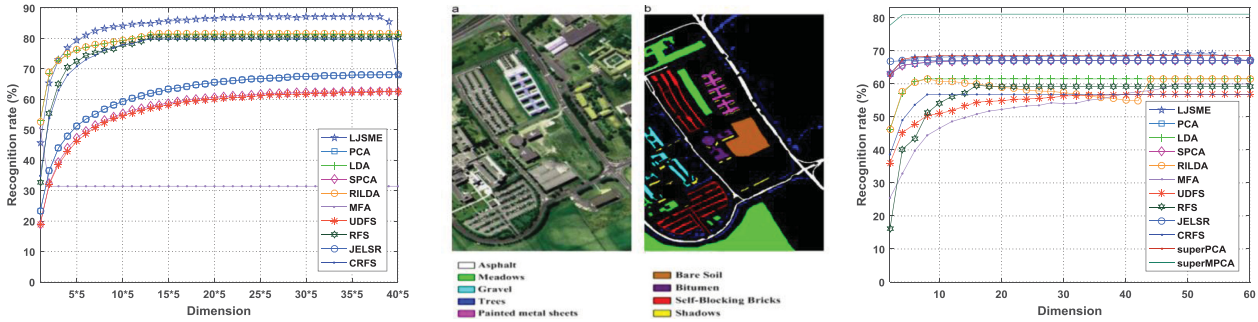
of the two methods. As we can see that MMC + Graph + $L_{2,1}$ is superior to FDC + Graph + $L_{2,1}$ for the proposed objective function. The potential reason is that FDC-based discriminant analysis method needs inverse operations, which would lead to low performance on the occasion when the sample number is much smaller than the dimensionality. That is, the SSS problem introduces singularity to scatter matrix so that the performance of FDC-based method is unsatisfactory. However, LJSME based on MMC avoids inverse operation and thus it is more robust in most cases.

Fig. 5(a) and Table VI show the experimental results on PIE database. From the results, we can clearly know that the proposed LJSME is more effective and robust than other methods as the recognition rate is much competitive.

C. The Experiment on Pavia University Database

In this experiment, we evaluate the performance of LJSME on hyperspectral image database. The Pavia University image has 103 features and the spatial dimensions is 610 × 340 [65]. The sample images are shown in Fig. 5(b).

Different from other experiments, in this experiment, superpixelwise PCA (superPCA) [62] and its multiscale extension (superMPCA) [62] are used as compared methods. SuperPCA and superMPCA are state-of-the-art methods that focus on learning the intrinsic low-dimensional features of hyperspectral images. The difference between superPCA and classical PCA methods is that superPCA extract potential low-dimensional features by supposing that different regions of the hyperspectral image should have different projections, while the classical PCA



(a) (b) (c)

Fig. 5. (a) The recognition rate versus dimension of all methods on the PIE database with $l = 4$, (b) Sample images on Pavia University dataset, (c) The performance of different methods on the PaviaU database with $l = 13$.

TABLE VI
THE MAXIMAL RECOGNITION RATE (%), STANDARD DEVIATION AND DIMENSION OF DIFFERENT METHODS ON THE PIE DATABASE

l	Baseline	PCA	LDA	SPCA	RILDA	MFA	UDFS	RFS	JELSR	CRFS	**LJSME**
4	68.35 ±9.32	68.09 ±9.33 200	81.40 ±10.75 65	62.54 ±10.14 200	81.64 ±9.38 150	50.53 ±11.68 5	62.54 ±10.15 200	80.18 ±8.81 65	68.09 ±9.34 200	79.77 ±10.35 65	**87.23** **±3.55 165**
5	72.14 ±7.37	71.82 ±7.31 200	88.15 ±7.24 65	72.46 ±13.56 200	88.30 ±5.67 115	56.42 ±13.53 5	72.46 ±13.57 200	87.40 ±7.42 65	71.82 ±7.32 200	86.53 ±9.20 65	**88.70** **±4.50 140**
6	74.3 ±7.70	73.72 ±7.61 195	89.84±5.20 65	76.72 ±12.43 200	89.45 ±4.60 65	58.04 ±11.35 5	76.72 ±12.45 200	88.26 ±7.08 65	73.72 ±7.61 195	87.82 ±8.00 65	**90.01** **±5.63 140**

TABLE VII
THE MAXIMAL RECOGNITION RATE (%), STANDARD DEVIATION AND DIMENSION OF DIFFERENT METHODS ON THE PAVIAU DATABASE

l	Baseline	PCA	LDA	SPCA	RILDA	MFA	UDFS	RFS	JELSR	CRFS	super PCA	super MPCA	**LJSME**
9	63.32 ±2.95	63.34 ±2.96 23	53.79 ±4.32 8	63.34 ±2.95 23	53.86 ±3.44 11	53.67 ±3.68 41	55.76 ±3.32 43	52.65 ±2.34 8	63.33 ±2.95 5	50.68 ±2.85 9	60.84 ±4.13 44	**77.60** **±1.77 30**	66.91 ±2.88 55
11	63.26 ±3.40	63.26 ±3.36 51	60.10 ±3.74 6	63.27 ±3.35 40	60.32 ±2.95 8	57.64 ±3.75 42	56.71 ±2.59 44	56.90 ±2.44 8	63.26 ±3.40 11	55.56 ±2.85 9	62.57 ±5.29 46	**79.08** **±3.92 25**	67.12 ±3.44 52
13	67.01 ±2.16	67.02 ±2.16 49	61.7 ±3.08 7	67.02 ±2.16 51	61.59 ±1.55 7	59.98 ±1.60 48	56.81 ±2.50 42	59.57 ±2.55 8	67.01 ±2.16 10	56.81 ±3.15 9	68.64 ±4.35 59	**81.11** **±1.88 26**	69.41 ±1.59 53
15	66.67 ±3.17	66.68 ±3.18 38	64.9 ±1.98 6	66.68 ±3.18 38	65.21 ±2.43 7	60.69 ±1.31 53	58.65 ±3.10 43	61.34 ±1.56 9	66.68 ±3.18 9	60.10 ±1.93 9	70.34 ±4.18 50	**81.72** **±2.22 30**	69.96 ±3.16 54

methods perform dimensionality reduction with a unified projection for an entire hyperspectral image.

The values of parameter α and β of the proposed method used on this dataset are $[10^{-3}, 10^{-2}, \ldots, 10^{9}]$ and $[10^{-9}, 10^{-8}, \ldots, 10^{9}]$, respectively. The experimental results on this dataset are shown in Table VII and Fig. 5(c). From Table VII and Fig. 5(c), we can know that even though the proposed LJSME is inferior to superMPCA, it outperforms superPCA and other compared methods in most cases.

D. The Experiment on PolyU Hyperspectral Face Database

The Hong Kong Polytechnic University Hyperspectral Face Database (PolyU-HSFD) [66] contains 1,410 images with size of 44×36 from 47 individuals. The sample images of this database are shown in Fig. 6(a).

We first explore the optimal values of parameter α and $\beta$and report them as $[10^{-6}, 10^{-3}]$, $[10^{-9}, 10^{0}]$, respectively. Table VIII and Fig. 6(b) show the recognition rates of different methods. All the results on this database show the good performance of the proposed LJSME.

E. The Discussion and Experiment on LFW Database

Current deep learning methods and the proposed LJSME are tools to extract features from images. They can be used for classification. However, they are different. First, LJSME is a traditional subspace learning method that is linear and easy to control because it only has two parameters compared to thousands of hundreds of parameters in many deep learning methods. Second, the training time of LJSME is much less than most deep learning methods. Third, when training deep learning methods, large amount of images are necessary for input. Otherwise, the performance cannot be guaranteed. That is, deep learning methods might be not suitable for the case when there are only few amount of inputs for training. However, the proposed LJSME is able to extract discriminant features even in such case. For example, on AR database, only 2 images of each individual (i.e., $2 \times 120 = 240$ images) were used for training and the remaining 18 images of each individual (i.e., $18 \times 120 = 2{,}160$ images) were used for testing, the proposed LJSME can obtain recognition rate of more than 80%.

Therefore, it is necessary to conduct fair experiment to explore the performance of LJSME in deep learning situation.

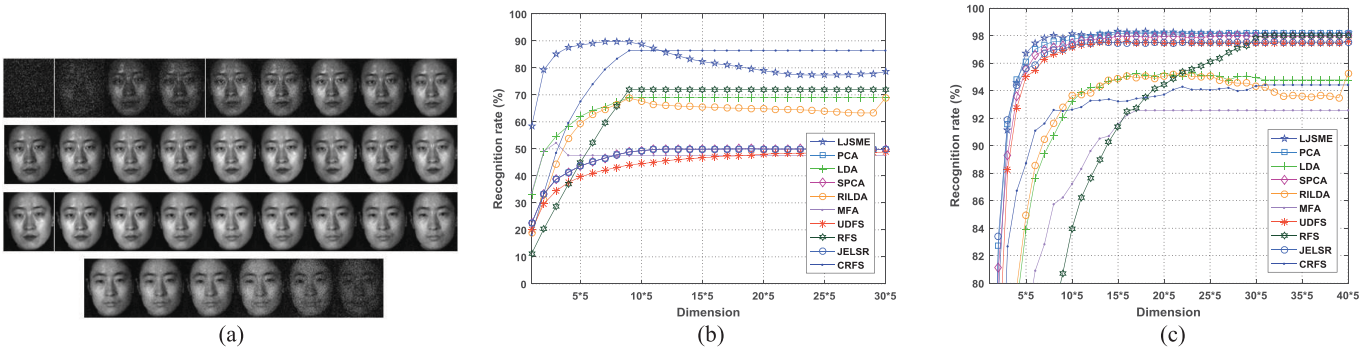


Fig. 6. (a) Sample images on PolyU-HSFD database, (b) The performance of different methods on the PolyU-HSFD database with $l = 5$, (c) The performance of different methods on the LFW database with $l = 3$.

TABLE VIII
THE MAXIMAL RECOGNITION RATE (%), STANDARD DEVIATION AND DIMENSION OF DIFFERENT METHODS ON THE POLYU-HSFD DATABASE

l	Baseline	PCA	LDA	SPCA	RILDA	MFA	UDFS	RFS	JELSR	CRFS	**LJSME**
5	49.57 ±8.94	50.04 ±8.52 95	69.00 ±11.10 45	50.03 ±8.50 95	68.84 ±11.43 45	52.22 ±9.94 15	49.08 ±9.12 200	72.03 ±12.42 45	49.99 ±8.49 95	86.43 ±9.25 45	**89.80** **±6.56 40**
6	52.33 ±10.57	52.89 ±9.98 130	75.98 ±8.42 45	52.93 ±10.00 80	74.69 ±7.96 45	57.73 ±11.68 20	52.52 ±10.13 200	72.85 ±10.05 45	52.92 ±10.00 65	87.29 ±9.34 45	**90.27** **±6.70 40**
7	62.29 ±8.30	63.06 ±8.80 140	82.93 ±8.35 45	63.05 ±8.80 140	82.14 ±8.03 45	60.44 ±12.66 20	62.63 ±8.74 200	77.44 ±9.31 45	63.07 ±8.82 135	88.35 ±9.40 45	**89.04** **±7.11 45**

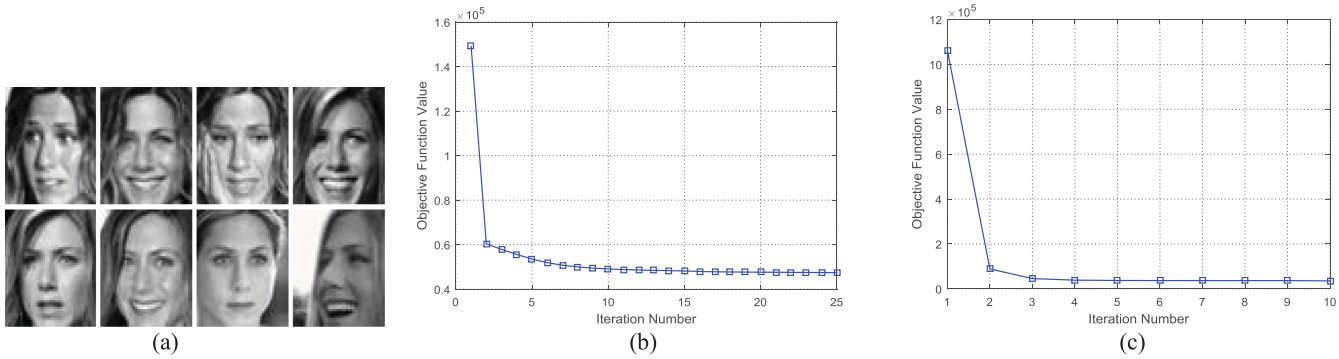


Fig. 7. (a) Sample image on LFW database. Convergence curve on (b) AR, (c) PolyU-HSFD database.

In this section, we conduct related experiments on LFW database. The LFW database is a well-known and challenging database. In the experiment, an aligned version of LFW database called LFW-a was used. There are total 4,324 images from 158 individuals on LFW-a dataset. The images are cropped and resized to 112 × 96 pixels. The sample images on LFW database are shown in Fig. 7(a).

The technique in this experiment is similar to [67]. Deep convolutional neural network (CNN) plays a role as feature extractor to get deep features of the data. After that, subspace learning methods are used to further perform feature extraction. Finally, the nearest neighbor classifier is used.

In this experiment, l ($l = 2, 3, 4$) images of each individual are selected to form the training set, and the rest are used for testing. The value of parameter α and β is the same with that on PolyU-HSFD database. The performance versus the dimension of all methods is shown in Fig. 6(c) while Table IX lists the maximal average recognition rate and the corresponding dimension of all methods.

From Fig. 6(c) and Table IX, we can find that under deep learning situation, all methods can obtain good performance though it may due to the help of CNN. However, the proposed LJSME is always better than other compared methods, which indicates that LJSME is able to extract discriminant features and improve the performance after deep feature extraction.

F. Speed Evaluation

Table X lists the average speed of different methods based on 10 times on all databases while Fig. 7(b) and (c) present the convergence curve of the proposed method. Compared with classical PCA and LDA, which are well-known efficient dimensionality reduction methods, the proposed LJSME is a litter slower, but it is still acceptable. Also, LJSME is faster than recent methods such as SPCA, RFS and UDFS in most of the

TABLE IX
THE MAXIMAL RECOGNITION RATE (%) AND DIMENSION OF DIFFERENT METHODS ON THE LFW DATABASE

Training samples	Baseline	PCA	LDA	SPCA	RILDA	MFA	UDFS	RFS	JELSR	CRFS	LJSME
2	97.95	98.03 100	82.49 100	97.78 170	58.81 120	76.02 80	96.76 100	97.36 155	97.38 110	92.81 155	**98.05 70**
3	98.18	98.18 135	95.30 115	98.03 140	95.27 110	92.57 85	97.56 80	98.03 155	97.56 65	94.42 155	**98.31 75**
4	98.35	98.35 135	97.56 105	98.35 190	97.45 85	95.88 100	97.37 85	98.24 155	98.32 80	95.42 155	**98.59 170**

TABLE X
THE AVERAGE SPEED (SECOND) OF DIFFERENT METHODS ON ALL DATABASES

Database	PCA	LDA	SPCA	RILDA	MFA	UDFS	RFS	JELSR	CRFS	LJSME
AR	0.2798	0.2519	30.5209	0.7299	0.0400	0.3334	2.1253	0.2193	0.0382	1.0355
CMU PIE	0.0303	0.0292	30.7146	0.4236	0.0258	0.2127	0.3079	0.2795	0.0137	1.4493
PaviaU	0.0010	0.0041	0.0278	0.0338	0.0044	0.0252	0.0095	0.0235	0.0030	0.0327
HSFD	0.2798	0.0443	18.4250	0.3791	0.0213	0.2092	0.2449	0.2224	0.0156	1.1549
LFW	0.1428	0.0160	3.3731	0.7637	0.0734	11.4974	3.2314	1.0612	0.0378	2.6145

cases. The convergence curves in Fig. 7(b) and (c) indicate that the proposed iterative algorithm is able to obtain the local optimal solution for LJSME in several steps. In all, we can say that LJSME is still an efficient method for feature selection and extraction.

G. Experimental Results and Discussions

The performance in terms of recognition rate of the proposed method and the compared methods (i.e., PCA, LDA, SPCA, MFA, RILDA, UDFS, RFS, JELSR, CRFS, superPCA and superMPCA) are presented above. According to the results and the intrinsic properties of those methods, we can have the following interesting points:

1) In most cases, RILDA and the proposed LJSME obtain better performance than other methods. The major reason of this phenomenon is that both the two methods use $L_{2,1}-$ norm instead of L_2 -norm to measure the reconstruction errors so that it can reduce the sensitivity to the outliers or the variations of facial expressions and various lighting conditions on face images. The difference between the two methods is that RILDA only focuses on reconstructing the scatter matrices with $L_{2,1}-$ norm. However, LJSME not only incorporates the graph embedding concept to the $L_{2,1}-$ norm based reconstructions to preserve the neighborhood relationship of the data, but also adds $L_{2,1}-$ norm penalty to the regularization term to guarantee the joint sparsity for discriminant analysis.
2) The performance of most compared methods is greatly influenced by the number of the training samples, especially LDA. As shown in Table V, VI and VIII, when the number of training images of each class is relatively small, LDA and RILDA obtain low performance. But their performance are improved greatly when the number of training samples increases. However, LJSME is totally different. It can obtain better performance than other methods in most cases. The potential reason is that it can avoid the SSS problem by utilizing the maximum margin criterion on the objective function so than it does not need inverse operation. As a result, LJSME has little or even no computational error during the computing process of feature selection or extraction.
3) Even though both LJSME and MFA use graph embedding to find the low-dimensional representations that best characterize the similarity between every pair of data, they have essential difference. First, MFA measures the distance of every data pair by using L_2- norm as the basic measurement, which would lead to enlarged errors because of the square operation. However, LJSME uses $L_{2,1}-$ norm instead of L_2- norm to avoid this potential risk so that it is more robust in dealing with the variations of images. Second, since MFA does not consider the sparsity of the projections, it lacks feature selection function. In contrast, LJSME using $L_{2,1}-$ norm penalty on the regularization term can obtain joint sparsity to perform feature extraction.
4) Even though UDFS, RFS, JELSR, CRFS and the proposed LJSME are the methods that use $L_{2,1}-$ norm on the regularization term, LJSME performs better in terms of recognition rate and robustness. The potential reason is that LJSME uses $L_{2,1}-$ norm as the basic measurement to redesign the scatter matrices, by which the robustness can be enhanced. On the other hand, using locality graph to preserve the neighborhood relationship among data pair can enhance the compactness of the same classes and meanwhile enlarge the distance between different classes. Therefore, LJSME is able to obtain discriminant information for classification and meanwhile improve the robustness.
5) From Table VII, we can know that the proposed LJSME is superior to superPCA in most cases, especially in the case when the scale of training data is small. The reason is

that in such case, the superpixels of the image obtained in superPCA are too small, which makes the oversegmented regions less distinctive. Thus, the principle components obtained from those segmented homogeneous regions have not much difference from that obtained from the whole hyperspectral image. In contrast, LJSME can obtain the discriminative features to improve the performance of classification by taking the label information into consideration and at the same time obtaining jointly sparse projections. In addition, we can see that superMPCA outperforms both superPCA and LJSME on PaviaU dataset. The phenomenon indicates that the multiscale segmentation strategy used in superMPCA can effectively improve the performance of hyperspectral image classification while the single-scale used in SuperPCA, and the general feature extraction strategy used in LJSME and other methods is not as such effective as that.

VI. Conclusion

In this paper, we proposed a locally joint sparse marginal embedding method for feature selection and extraction. The proposed method aims to improve the compactness of the same classes and meanwhile enlarge the distance of different classes based on maximum margin criterion. It can avoid the SSS problem and also enhance the robustness to outliers by reconstructing the scatter matrices with $L_{2,1}-$ norm instead of L_2-norm. Besides, it can guarantee the joint sparsity of the projections matrix so as to perform effective feature selection/extraction. An iterative algorithm is proposed to solve the optimization problem. The comparison and discussion between the proposed method and some related methods are shown and the theoretical analysis including computational complexity and convergence of the proposed algorithm are also presented. Experiments on several well-known databases including face images and hyperspectral images indicate that the proposed method obtains better performance than the traditional PCA, the classical LDA, the marginal fisher analysis (MFA), the sparse component analysis (SPCA), the $L_{2,1}-$ norm based methods (i.e., RFS, UDFS, JELSR, CRFS) and the rotational invariant LDAn (RILDA).

Appendix

Proof of Proposition 1

(a): Based on the definition of $L_{2,1}$–norm, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\|_2\boldsymbol{W}_{w,ij}= \left\|\begin{matrix}\boldsymbol{W}_{w,11}(\boldsymbol{x}_1-\boldsymbol{x}_1)^T\boldsymbol{U}\\ \boldsymbol{W}_{w,12}(\boldsymbol{x}_1-\boldsymbol{x}_2)^T\boldsymbol{U}\\ \vdots\\ \boldsymbol{W}_{w,1n}(\boldsymbol{x}_1-\boldsymbol{x}_n)^T\boldsymbol{U}\end{matrix}\right\|_{2,1} +\cdots+\left\|\begin{matrix}\boldsymbol{W}_{w,n1}(\boldsymbol{x}_n-\boldsymbol{x}_1)^T\boldsymbol{U}\\ \boldsymbol{W}_{w,n2}(\boldsymbol{x}_n-\boldsymbol{x}_2)^T\boldsymbol{U}\\ \vdots\\ \boldsymbol{W}_{w,nn}(\boldsymbol{x}_n-\boldsymbol{x}_n)^T\boldsymbol{U}\end{matrix}\right\|_{2,1}. \tag{22}$$

Let

$$\boldsymbol{X}_{w,1}=\begin{bmatrix}\boldsymbol{W}_{w,11}(\boldsymbol{x}_1-\boldsymbol{x}_1)^T\\ \boldsymbol{W}_{w,12}(\boldsymbol{x}_1-\boldsymbol{x}_2)^T\\ \vdots\\ \boldsymbol{W}_{w,1n}(\boldsymbol{x}_1-\boldsymbol{x}_n)^T\end{bmatrix},\ldots,\boldsymbol{X}_{w,n} =\begin{bmatrix}\boldsymbol{W}_{w,11}(\boldsymbol{x}_n-\boldsymbol{x}_1)^T\\ \boldsymbol{W}_{w,12}(\boldsymbol{x}_n-\boldsymbol{x}_2)^T\\ \vdots\\ \boldsymbol{W}_{w,1n}(\boldsymbol{x}_n-\boldsymbol{x}_n)^T\end{bmatrix},$$

we can simplify the equality (22) as below

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\right\|_2\boldsymbol{W}_{w,ij}=\|\boldsymbol{X}_{w,1}\boldsymbol{U}\|_{2,1} +\|\boldsymbol{X}_{w,2}\boldsymbol{U}\|_{2,1}+\cdots+\|\boldsymbol{X}_{w,n}\boldsymbol{U}\|_{2,1}.$$

From the property of $L_{2,1}-$ norm on matrix $\boldsymbol{A}$ that $\|\boldsymbol{A}\|_{2,1}=tr(\boldsymbol{A}^T\ \boldsymbol{D}\boldsymbol{A})$, where $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{ii}=\frac{1}{2\|\boldsymbol{A}^i\|_2}$ ($\boldsymbol{A}^i$ denotes the i-th row of $\boldsymbol{A}$), we have

$$\|\boldsymbol{X}_{w,1}\boldsymbol{U}\|_{2,1}+\|\boldsymbol{X}_{w,2}\boldsymbol{U}\|_{2,1}+\cdots+\|\boldsymbol{X}_{w,n}\boldsymbol{U}\|_{2,1}= tr(\boldsymbol{U}^T\boldsymbol{X}_{w,1}^T\boldsymbol{D}_{w,1}\boldsymbol{X}_{w,1}\boldsymbol{U})+\cdots+tr(\boldsymbol{U}^T\boldsymbol{X}_{w,n}^T\boldsymbol{D}_{w,n}\boldsymbol{X}_{w,n}\boldsymbol{U}) =tr\left(\boldsymbol{U}^T\left(\sum_{i=1}^{n}\boldsymbol{X}_{w,i}^T\boldsymbol{D}_{w,i}\boldsymbol{X}_{w,i}\right)\boldsymbol{U}\right), \tag{23}$$

where $(\boldsymbol{D}_{w,1})_{ii}=\frac{1}{2\|(\boldsymbol{X}_{w,1}\boldsymbol{U})^i\|_2},\ldots,(\boldsymbol{D}_{w,n})_{ii}=\frac{1}{2\|(\boldsymbol{X}_{w,n}\boldsymbol{U})^i\|_2}$.

Since

$$\sum_{i=1}^{n}\boldsymbol{X}_{w,i}^T\boldsymbol{D}_{w,i}\boldsymbol{X}_{w,i}= \begin{bmatrix}\boldsymbol{X}_{w,1}\\ \boldsymbol{X}_{w,2}\\ \vdots\\ \boldsymbol{X}_{w,n}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{D}_{w,1}0\cdots\cdots\cdots\cdots\cdots\cdots0\\ 0\quad\boldsymbol{D}_{w,2}\cdots\cdots\cdots\cdots\cdots0\\ \vdots\\ 0\cdots\cdots\cdots\cdots\cdots\cdots\cdots\boldsymbol{D}_{w,n}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_{w,1}\\ \boldsymbol{X}_{w,2}\\ \vdots\\ \boldsymbol{X}_{w,n}\end{bmatrix} =\boldsymbol{X}_w^T\boldsymbol{D}_w\boldsymbol{X}_w,$$

where

$$\boldsymbol{X}_w=\begin{bmatrix}\boldsymbol{X}_{w,1}\\ \boldsymbol{X}_{w,2}\\ \vdots\\ \boldsymbol{X}_{w,1n}\end{bmatrix},\boldsymbol{D}_w=\begin{bmatrix}\boldsymbol{D}_{w,1}0\cdots\cdots\cdots\cdots\cdots\cdots0\\ 0\quad\boldsymbol{D}_{w,2}\cdots\cdots\cdots\cdots\cdots0\\ \vdots\\ 0\cdots\cdots\cdots\cdots\cdots\cdots\cdots\boldsymbol{D}_{w,n}\end{bmatrix},$$

we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\right\|_2\boldsymbol{W}_{w,ij}=tr(\boldsymbol{U}^T\boldsymbol{X}_w^T\boldsymbol{D}_w\boldsymbol{X}_w\boldsymbol{U}). \tag{24}$$

(b): Similarly, the between-class scatter matrix is defined as

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\right\|_2\boldsymbol{W}_{b,ij}=$$

$$\left\|\begin{matrix}\boldsymbol{W}_{b,11}(\boldsymbol{x}_1-\boldsymbol{x}_1)^T\boldsymbol{U}\\ \boldsymbol{W}_{b,12}(\boldsymbol{x}_1-\boldsymbol{x}_2)^T\boldsymbol{U}\\ \vdots\\ \boldsymbol{W}_{b,1n}(\boldsymbol{x}_1-\boldsymbol{x}_n)^T\boldsymbol{U}\end{matrix}\right\|_{2,1}+\cdots+\left\|\begin{matrix}\boldsymbol{W}_{b,n1}(\boldsymbol{x}_n-\boldsymbol{x}_1)^T\boldsymbol{U}\\ \boldsymbol{W}_{b,n2}(\boldsymbol{x}_n-\boldsymbol{x}_2)^T\boldsymbol{U}\\ \vdots\\ \boldsymbol{W}_{b,nn}(\boldsymbol{x}_n-\boldsymbol{x}_n)^T\boldsymbol{U}\end{matrix}\right\|_{2,1}$$

$$=\|\boldsymbol{X}_{b,1}\boldsymbol{U}\|_{2,1}+\|\boldsymbol{X}_{b,2}\boldsymbol{U}\|_{2,1}+\cdots+\|\boldsymbol{X}_{b,n}\boldsymbol{U}\|_{2,1}$$

$$=tr(\boldsymbol{U}^T\boldsymbol{X}_{b,1}^T\boldsymbol{D}_{b,1}\boldsymbol{X}_{b,1}\boldsymbol{U})+\cdots+tr(\boldsymbol{U}^T\boldsymbol{X}_{b,n}^T\boldsymbol{D}_{b,n}\boldsymbol{X}_{b,n}\boldsymbol{U})$$

$$=tr\left(\boldsymbol{U}^T\left(\sum_{i=1}^{n}\boldsymbol{X}_{b,i}^T\boldsymbol{D}_{b,i}\boldsymbol{X}_{b,i}\right)\boldsymbol{U}\right)$$

$$=tr(\boldsymbol{U}^T\boldsymbol{X}_b^T\boldsymbol{D}_b\boldsymbol{X}_b\boldsymbol{U}), \tag{25}$$

where

$$\boldsymbol{X}_{b,1}=\begin{bmatrix}\boldsymbol{W}_{b,11}(\boldsymbol{x}_1-\boldsymbol{x}_1)^T\\ \boldsymbol{W}_{b,12}(\boldsymbol{x}_1-\boldsymbol{x}_2)^T\\ \vdots\\ \boldsymbol{W}_{b,1n}(\boldsymbol{x}_1-\boldsymbol{x}_n)^T\end{bmatrix},\ldots,\boldsymbol{X}_{b,n}=\begin{bmatrix}\boldsymbol{W}_{b,11}(\boldsymbol{x}_n-\boldsymbol{x}_1)^T\\ \boldsymbol{W}_{b,12}(\boldsymbol{x}_n-\boldsymbol{x}_2)^T\\ \vdots\\ \boldsymbol{W}_{b,1n}(\boldsymbol{x}_n-\boldsymbol{x}_n)^T\end{bmatrix},$$

$$(\boldsymbol{D}_{b,1})_{ii}=\frac{1}{2\left\|(\boldsymbol{X}_{b,1}\boldsymbol{U})^i\right\|_2},\ldots,(\boldsymbol{D}_{b,n})_{ii}=\frac{1}{2\left\|(\boldsymbol{X}_{b,n}\boldsymbol{U})^i\right\|_2}$$

$$\boldsymbol{X}_b=\begin{bmatrix}\boldsymbol{X}_{b,1}\\ \boldsymbol{X}_{b,2}\\ \vdots\\ \boldsymbol{X}_{b,n}\end{bmatrix},\boldsymbol{D}_b=\begin{bmatrix}\boldsymbol{D}_{b,1}0\cdots\cdots\cdots\cdots\cdots\cdots0\\ 0\quad\boldsymbol{D}_{b,2}\cdots\cdots\cdots\cdots\cdots\cdots0\\ \vdots\\ 0\cdots\cdots\cdots\cdots\cdots\cdots\cdots\boldsymbol{D}_{b,n}\end{bmatrix}.$$

That is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{U}\right\|_2\boldsymbol{W}_{b,ij}=tr(\boldsymbol{U}^T\boldsymbol{X}_b^T\boldsymbol{D}_b\boldsymbol{X}_b\boldsymbol{U}). \tag{26}$$

Therefore, the proof of Proposition 1 is completed. ■

REFERENCES

[1] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 716–727, Aug. 2006.

[2] F. Wu *et al.*, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[3] W. Xie, L. Shen, and J. Jiang, "A novel transient wrinkle detection algorithm and its application for expression synthesis," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 279–292, Feb. 2017.

[4] Y. Lu *et al.*, "Nuclear norm-based 2DLPP for image classification," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2391–2403, Nov. 2017.

[5] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[6] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[7] H. Yan and J. Yang, "Sparse discriminative feature selection," *Pattern Recogn.*, vol. 48, no. 5, pp. 1827–1835, 2015.

[8] F. Zhong, J. Zhang, and D. Li, "Discriminant locality preserving projections based on L_1-norm maximization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2065–2074, Nov. 2014.

[9] M. Turk, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[10] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Mach. Learn. Res.*, vol. 6, pp. 483–502, 2005.

[11] C. Luo, B. Ni, S. Yan, and M. Wang, "Image classification by selective regularized subspace learning," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 40–50, Jan. 2016.

[12] L. Chen, H. M. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recogn.*, vol. 33, no. 10, pp. 1713–1726, 2000.

[13] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recogn.*, vol. 34, no. 10, pp. 2067–2070, 2001.

[14] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear discriminant analysis using rotational invariant L_1-norm," *Neurocomputing*, vol. 73, nos. 13–15, pp. 2571–2579, 2010.

[15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[16] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[17] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," *Comput. Vis. Pattern Recogn.*, vol. 2, pp. 564–569, 2004.

[18] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recogn. Lett.*, vol. 26, pp. 527–532, 2005.

[19] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Jul. 2004, pp. 1569–1576.

[20] S. Yan *et al.*, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, Jan. 2007.

[21] R. Hu, W. Jia, D. Huang, and Y. Lei, "Maximum margin criterion with tensor representation," *Neurocomputing*, vol. 73, nos. 10–12, pp. 1541–1549, 2010.

[22] Y. Koren and L. Carmel, "Robust linear dimensionality reduction," *IEEE Trans. Vis. Comput. Graph.*, vol. 10, no. 4, pp. 459–470, Aug. 2004.

[23] Y. Lu and Q. Tian, "Discriminant subspace analysis: An adaptive approach for image classification," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1289–1300, Nov. 2009.

[24] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 1–30, 2004.

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

[26] A. Aspremont, F. Bach, I. Willow, and L. El Ghaoui, "Optimal solutions for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 9, no. 100, pp. 1269–1294, 2008.

[27] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, "Multilinear sparse principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1942–1950, Oct. 2014.

[28] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data," *IAENG Int. J. Appl. Math.*, vol. 39, no. 1, pp. 48–60, 2009.

[29] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L_1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[30] Z. Fujin and Z. Jiashu, "Linear discriminant analysis based on L_1-norm maximization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.

[31] Q. Feng and Y. Zhou, "Kernel combined sparse representation for disease recognition," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 1956–1968, Oct. 2016.

[32] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2010–2021, Sep. 2017.

[33] S. Wang and W. Guo, "Sparse multigraph embedding for multimodal feature representation," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1454–1466, Jul. 2017.

[34] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2429–2443, Jun. 2016.

[35] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. Int. Joint Conf. Artificial Intell.*, vol. 55, 2011, pp. 1294–1299.
[36] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. Int. Joint Conf. Artificial Intell.*, 2011, pp. 1324–1329.
[37] J. Huang, G. Li, Q. Huang, and X. Wu, "Joint feature selection and classification for multilabel learning," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 876–889, Mar. 2018.
[38] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint L2,1 norms minimization," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[39] X. Shi, Y. Yang, Z. Guo, and Z. Lai, "Face recognition by sparse discriminant analysis via joint L2,1-norm minimization," *Pattern Recogn.*, vol. 47, no. 7, pp. 2447–2453, 2014.
[40] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$L_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artificial Intell.*, 2011, pp. 1589–1594.
[41] Z. Lai, Y. Xu, J. Yang, L. Shen, and D. Zhang, "Rotational invariant dimensionality reduction algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3733–3746, Nov. 2017.
[42] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.
[43] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[44] K. Fukunaga, "Introduction to statistical pattern recognition," *New York Acad.*, vol. 60, no. 12-1, pp. 2133–2143, 1990.
[45] J. Zhang, J. Yu, J. Wan, and Z. Zeng, "$L_{2,1}$ norm regularized fisher criterion for optimal feature selection," *Neurocomputing*, vol. 166, pp. 455–463, 2015.
[46] X. He and P. Niyogi, "Locality preserving projections," *Neural Inf. Process. Syst.*, vol. 16, no. 1, pp. 186–197, 2004.
[47] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[48] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, vol. 2, 2005, pp. 1208–1213.
[49] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[50] Y. Pang, L. Zhang, Z. Liu, N. Yu, and H. Li, "Neighborhood preserving projections (NPP): A novel linear dimension reduction method," in *Proc. Int. Conf. Adv. Intell. Comput.*, 2005, pp. 117–125.
[51] Y. Cui and L. Fan, "A novel supervised dimensionality reduction algorithm: Graph-based Fisher analysis," *Pattern Recogn.*, vol. 45, no. 4, pp. 1471–1481, 2012.
[52] S. Yan, D. Xu, B. Zhang, and H. Zhang, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[53] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.
[54] W. Yang, J. Wang, M. Ren, and J. Yang, "Feature extraction based on Laplacian bidirectional maximum margin criterion," *Pattern Recogn.*, vol. 42, no. 11, pp. 2327–2344, 2009.
[55] J. Wen, Z. Lai, Y. Zhan, and J. Cui, "The L2, 1-norm-based unsupervised optimal feature selection with applications to action recognition," *Pattern Recogn.*, vol. 60, pp. 515–530, 2016.
[56] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[57] X. Shi, Y. Yang, Z. Guo, and Z. Lai, "Face recognition by sparse discriminant analysis via joint $L_{2,1}$-norm minimization," *Pattern Recogn.*, vol. 47, no. 7, pp. 2447–2453, 2014.
[58] Z. Lai, D. Mo, J. Wen, L. Shen, and W. Wong, "Generalized robust regression for jointly sparse subspace learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 756–772, Mar. 2019.
[59] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.
[60] S. Yan, D. Xu, B. Zhang, and H. Zhang, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[61] R. He, T. Tan, L. Wang, and W. Zheng, "$L_{2,1}$ regularized correntropy for robust feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2012, pp. 2504–2511.
[62] J. Jiang *et al.*, "SuperPCA : A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.
[63] A. A. Martinez and R. Benavente, "The AR face database," CVC Tech. Rep. 24, 1998.
[64] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illuminlation, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[65] *Hyperspectral Remote Sensing Scenes*. 2014. [Online]. Available: http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes
[66] W. Di, L. Zhang, D. Zhang, and Q. Pan, "Studies on hyperspectral face recognition in visible spectrum with feature band selection," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 6, pp. 1354–1361, Nov. 2010.
[67] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

The authors' photographs and biographies not available at the time of publication.