Rotational Invariant Dimensionality Reduction Algorithms

Zhihui Lai, Yong Xu, *Member, IEEE*, Jian Yang, Linlin Shen, and David Zhang, *Fellow, IEEE*

***Abstract*—A common intrinsic limitation of the traditional subspace learning methods is the sensitivity to the outliers and the image variations of the object since they use the L_2 norm as the metric. In this paper, a series of methods based on the $L_{2,1}$-norm are proposed for linear dimensionality reduction. Since the $L_{2,1}$-norm based objective function is robust to the image variations, the proposed algorithms can perform robust image feature extraction for classification. We use different ideas to design different algorithms and obtain a unified rotational invariant (RI) dimensionality reduction framework, which extends the well-known graph embedding algorithm framework to a more generalized form. We provide the comprehensive analyses to show the essential properties of the proposed algorithm framework. This paper indicates that the optimization problems have global optimal solutions when all the orthogonal projections of the data space are computed and used. Experimental results on popular image datasets indicate that the proposed RI dimensionality reduction algorithms can obtain competitive performance compared with the previous L_2 norm based subspace learning algorithms.**

***Index Terms*—Dimensionality reduction, image classification, image feature extraction, rotational invariant (RI) subspace learning.**

I. Introduction

FEATURE extraction and dimensionality reduction methods have been paid much attention in past several decades. The classical linear dimensionality reduction methods such as principle component analysis (PCA) [1]–[3] and linear discriminant analysis (LDA) [4] and its variations [5], [6] are widely used in the fields of pattern recognition, computer vision, and data mining. It is known that these classical methods (i.e., PCA and LDA) only focus on the global structure of a dataset in dimensionality reduction. With the fast development of the manifold learning based techniques [7]–[10], the local geometry structure has been taken into account in designing different linear dimensionality reduction methods. For example, locality preserving projection (LPP, also called Laplacianfaces) [11] and orthogonal LPP [12] were proposed for face recognition. Yan *et al.* [13] proposed a unified graph embedding framework for linear and nonlinear dimensionality reduction, and marginal fisher analysis (MFA) and its extension [14] were proposed the for face and gait feature extraction.

All the above methods, however, use the L_2 or Frobenius norm based metric to characterize the scatter of the dataset, thus these methods are sensitive to the outliers. Recently, other measurement such as L_1 norm was widely explored due to its robustness in different applications. For example, the L_1 norm was used in sparse regression [15]–[17], sparse representation classifier designation [18], [19], subspace learning [20]–[25], sparse subspace learning [26], [27], and sparse coding for image representation [28]. In addition, the sparse L_1 graph was also used in subspace learning, spectral clustering [29], and label propagation [30]. But one drawback of these L_1 norm based methods is that the L_1 norm terms are just used as the regularization and the L_2 or Frobenius norm terms are still dominant in the optimization problems. Thus, these methods are still sensitive to the outliers in a certain sense in dimensionality reduction.

Although various L_1 norm based subspace learning methods, such as those in [25] and [31]–[34], have shown promising performance, these methods still have some unsolved problems. For example, some of them have very high computational costs in computing the (local) optimal solutions, and the theoretical relation between the optimal solutions of L_1 norm based methods and the traditional/classical ones was still unclear. Recently, a new measurement called rotational invariance (RI) L_1 norm or $L_{2,1}$ norm has attracted much attention in the fields of patter recognition and computer vision [35], multitask learning and tensor factorization [36]. Previous studies show that the pure $L_{2,1}$ norm based regression is more robust than the L_1 norm regression in pattern recognition [37]–[39], and thus was widely used in

Manuscript received May 23, 2015, revised November 11, 2015; accepted May 24, 2016. This work was supported in part by the Natural Science Foundation of China under Grant 61573248, Grant 61203376, Grant 61375012, Grant 61272050, Grant 61362031, Grant 61332011, and Grant 61370163, in part by the General Research Fund of Research Grants Council of Hong Kong under Project 531708, in part by the Science Foundation of Guangdong Province under Grant 2014A030313556, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20150324141711637. This paper was recommended by Associate Editor P. Tino.

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

Y. Xu is with the Bio-Computing Research Center and Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: yongxu@ymail.com).

J. Yang is with the School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

L. Shen is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: llshen@szu.edu.cn).

D. Zhang is with the Biometrics Research Centre, Department of Computing, Hong Kong Polytechnic University, Hong Kong (e-mail: csdzhang@comp.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/TCYB.2016.2578642

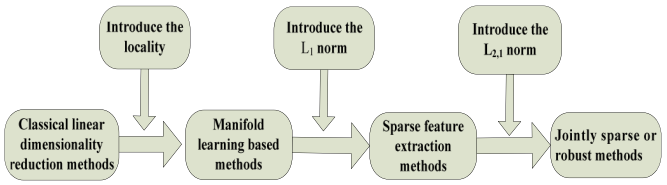


Fig. 1. Development route of the dimensionality reduction methods mentioned in this paper.

joint feature selection and subspace learning [40]–[42], image recognition [43], Web image annotation [44], and multimedia data understanding [45]. The brief development route of the dimensionality reduction methods mentioned in this paper is shown in Fig. 1.

Robustness is an important issue in feature extraction. One tractable method is to introduce the robust measurement, i.e., replace the Frobenius norm with other norms which are robust to the outliers. A relative ideal norm for feature extraction and recognition should contain the following aspects: 1) the measurement should be robust to the outliers; 2) the derived model using this measurement is easy to solve; and 3) it is better to bridge the strong theoretical connections between the previous methods and the new ones using the introduced norm. Based on these three aspects, the RI L_1 norm or $L_{2,1}$ norm is a very suitable candidate owing to its robustness to outliers and different variations in the dataset [36]–[39], the simplicity for solving the derived models as indicated in [38], and the close theoretical connections to previous methods (which will be shown on Section III of this paper).

This paper focuses on designing the robust linear dimensionality reduction methods using the RI L_1 norm or $L_{2,1}$ norm. The difference between this paper and the previous works is that this paper not only focuses on a set of concrete robust subspace learning methods, but also builds a unified framework to conclude the proposed methods using the RI L_1 norm or $L_{2,1}$ norm. From the algorithm aspect, the significant difference between the proposed algorithms and the previous ones is the algorithms presented in this paper can achieve its robustness automatically or intrinsically (without introducing any other parameters).

The main contributions of this paper are as follows.

1) We propose four representative RI subspace learning methods, i.e., RI PCA (RIPCA), RI LDA (RILDA), RI LPP (RILPP), and RI MFA (RIMFA) for image feature extraction. Besides these newly proposed algorithms, we also propose a unified robust and RI subspace learning framework. It is shown that the framework proposed in this paper indeed extends the well-known graph embedding framework proposed in [13] to a more general form for linear dimensionality reduction.
2) The comprehensive analyses, including the convergence, computational complexity, and the theoretical connections between this framework and the previous graph embedding algorithm framework, are presented to show the essential properties of the proposed algorithm framework. And more importantly, the optimization problems derived by the new metric are easy to solve and the codes are also very easy to implement (the codes can be downloaded from http://www.scholat.com/laizhihui).
3) Extensive experiments show that the proposed RI subspace learning algorithms perform better than the previous ones for image feature extraction in most cases.

The rest of this paper is organized as follows. In Section II, four RI subspace learning algorithms are proposed. The theoretical analyses of the proposed framework are presented in Section III. Experiments are carried out in Section IV to test these RI subspace learning algorithms where the objects in the databases have different variations, and the conclusions are given in Section V.

II. Proposed Algorithms

In this section, some notations are given at first and then the algorithms are presented. Let matrix $X = [x_1, x_2, \ldots, x_N]$ be the data matrix including all the training samples $\{x_i\}_{i=1}^N \in R^m$ in its columns. In practice, the feature dimension m is often very high. The goal of feature extraction is to transform the data from the originally high-dimensional space to a low-dimensional one. In other words, sample $x \in R^m$ should be transformed into $y \in R^d$ ($d \ll m$) by using

$$y = U^T x \in R^d \tag{1}$$

where $U = (u_1, u_2, \ldots, u_d)$ and $u_i (i = 1, \ldots, d)$ is an m-dimension column vector.

A. Definitions of Different Norms

For a given matrix $A = [a_{ij}] \in R^{n \times m}$, we denote the ith row of A by A^i. The Frobenius norm of matrix A is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2} = \sqrt{\sum_{i=1}^{n} \|A^i\|_2^2}. \tag{2}$$

It can be seen that the sensitivity of the Frobenius norm comes from the squared operation, which makes the larger values of $\|A^i\|_2^2$ significantly dominate the final result. Differing from the Frobenius norm, L_1-norm of a matrix A is defined as

$$\|A\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|. \tag{3}$$

The $L_{2,1}$-norm of a matrix is defined as

$$\|A\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} a_{ij}^2} = \sum_{i=1}^{n} \|A^i\|_2. \tag{4}$$

Since for any rotational matrix R, $\|AR\|_{2,1} = \|A\|_{2,1}$, $L_{2,1}$-norm is RI (this is the reason why the proposed algorithms are called as RI in this paper). As indicated in [38], the robustness of the $L_{2,1}$-norm or RI L_1-norm is originated from its special definition, where there is no squared operation. Note that, if A degrades to be a high-dimensional row vector, its $L_{2,1}$-norm or RI L_1-norm will degrade to the Frobenius norm.

TABLE I
SUMMARY OF THE ALGORITHMS

Data matrix/graph	Weight
$X_w = X_{Rt} = \left((x_1 - \bar{x}),...,(x_N - \bar{x})\right) / G_w = G_I = I_{m\times m}$ $X_b = X_I = I_{m\times m} / G_b = G_I = I_{m\times m}$	$D_w = D_{Rt} = diag(\frac{1}{2\left\|(x_1^T - \bar{x}^T)U\right\|_2},...,\frac{1}{2\left\|(x_N^T - \bar{x}^T)U\right\|_2})$ $D_b = D_I = I_{m\times m}$
$X_w = X_{Rw} = \left((x_1^1 - \bar{x}_1),...,(x_1^{N_1} - \bar{x}_{N_1}),...,(x_c^1 - \bar{x}_c),...,(x_c^{N_c} - \bar{x}_c)\right)$ $/ G_w = G_{Iw} = I_{m\times m}$ $X_b = X_{Rb} = \left(N_1(\bar{x}_1 - \bar{x}),...,N_c(\bar{x}_c - \bar{x})\right) / G_b = G_{Ib} = I_{c\times c}$	$D_w = D_{Rw} = diag(\frac{1}{2\left\|(x_1^1 - \bar{x}_1)^T U\right\|_2},...,\frac{1}{2\left\|(x_1^{N_1} - \bar{x}_{N_1})^T U\right\|_2}$ $,...,\frac{1}{2\left\|(\bar{x}_c^{N_1} - \bar{x}_c)^T U\right\|_2},...,\frac{1}{2\left\|(\bar{x}_c^{N_c} - \bar{x}_c)^T U\right\|_2})$ $D_b = D_{Rb} = diag(\frac{1}{2\left\|N_1(\bar{x}_1^T - \bar{x}^T)U\right\|_2},...,\frac{1}{2\left\|N_c(\bar{x}_c^T - \bar{x}^T)U\right\|_2})$
$X_w = X_R = \left((x_1 - x_1),...,(x_1 - x_N),...,(x_N - x_1),...,(x_N - x_N)\right)^T /$ $G_w = G_{Rl} = diag(W_{11}, W_{12},...,W_{1N},...,W_{N1}, W_{N2},...,W_{NN})$ $X_b = X = [x_1, x_2,..., x_N] / G_b = G_{Rd} = diag(D_{11}, D_{22},..., D_{NN})$	$D_w = D_{Rl} = diag(\frac{1}{2W_{11}\left\|(x_1 - x_1)^T U\right\|_2},...,\frac{1}{2W_{1N}\left\|(x_1 - x_N)^T U\right\|_2},$ $...,\frac{1}{2W_{N1}\left\|(x_N - x_1)^T U\right\|_2},...,\frac{1}{2W_{NN}\left\|(x_N - x_N)^T U\right\|_2})$ $D_b = D_{Rd} = diag(\frac{1}{2D_{11}\left\|x_1^T U\right\|_2},\frac{1}{2D_{22}\left\|x_2^T U\right\|_2},...,\frac{1}{2D_{NN}\left\|x_N^T U\right\|_2})$
$X_w = X_R = \left((x_1 - x_1),...,(x_1 - x_N),...,(x_N - x_1),...,(x_N - x_N)\right)^T /$ $G_w = G_{Rc} = diag(W_{11}^c, W_{12}^c,..., W_{1N}^c,..., W_{N1}^c, W_{N2}^c,..., W_{NN}^c)$ $X_b = X_R = \left((x_1 - x_1),...,(x_1 - x_N),...,(x_N - x_1),...,(x_N - x_N)\right)^T /$ $G_b = G_{Rp} = diag(W_{11}^p, W_{12}^p,..., W_{1N}^p,..., W_{N1}^p, W_{N2}^p,..., W_{NN}^p)$	$D_w = D_{Rc} = diag(\frac{1}{2W_{11}^c\left\|(x_1 - x_1)^T U\right\|_2},...,\frac{1}{2W_{1N}^c\left\|(x_1 - x_N)^T U\right\|_2},$ $...,\frac{1}{2W_{N1}^c\left\|(x_N - x_1)^T U\right\|_2},...,\frac{1}{2W_{NN}^c\left\|(x_N - x_N)^T U\right\|_2})$ $D_b = D_{Rp} = diag(\frac{1}{2W_{11}^p\left\|(x_1 - x_1)^T U\right\|_2},...,\frac{1}{2W_{1N}^p\left\|(x_1 - x_N)^T U\right\|_2},$ $...,\frac{1}{2W_{N1}^p\left\|(x_N - x_1)^T U\right\|_2},...,\frac{1}{2W_{NN}^p\left\|(x_N - x_N)^T U\right\|_2})$

B. Discussions and the Motivations of This Paper

Yan *et al.* [13] proposed a general framework to unify the subspace learning methods, including PCA, LDA, LPP, and MFA, by using the model

$$\min_U \quad \mathrm{tr}\left[U^T X(\bar{D}^w - \bar{W}^w)X^T U\right] \tag{5}$$

$$\text{s.t.} \quad \mathrm{tr}\left(U^T X\left(\bar{D}^b - \bar{W}^b\right)X^T U\right) = \mathrm{cons} \tag{6}$$

where cons denotes the constant, $\bar{W}^w$ and $\bar{W}^b$ are the graphs defined on the dataset, and $\bar{D}_{ii}^b = \sum_i \bar{W}_{ij}^b$, $\bar{D}_{ii}^w = \sum_i \bar{W}_{ij}^w$. The optimal solutions of the above problem can be given by the following generalized eigenequation:

$$X(\bar{D}^w - \bar{W}^w)X^T U = X\left(\bar{D}^b - \bar{W}^b\right)X^T U\Lambda. \tag{7}$$

Many linear dimensionality reduction methods can be included in this graph embedding algorithm framework [13]. However, as it is mentioned in Section I, the crucial drawback of this framework is its sensitiveness to the outliers or the variations of the images since it uses the L_2 or Frobenius norm as the metric [35], [36], [38], [39]. The drawback of the previous framework (or previous subspace learning algorithms) inspires us to develop new framework (or algorithms) which is robust in linear dimensionality reduction. Motivated by the previous robust subspace learning algorithms [35], [36], [38], [39], [41], [43], we introduce the RI L_1 norm or $L_{2,1}$ norm to develop a set of algorithms for robust linear dimensionality reduction. Four simple but effective and efficient algorithms (i.e., RIPCA, RILDA, RILPP, and RIMFA) are first proposed and then a unified framework is obtained for robust linear dimensionality reduction. For ease of reading and comparison, all the detailed information of the four algorithms presented in this paper will be summarized in Table I and the algorithm steps are shown in Table II.

C. RIPCA

PCA aims to find a set of projections that can characterize the most of the variances of the data points by using the square norm. But for RIPCA, the RI L_1-norm (i.e., $L_{2,1}$-norm) is used as the measurement among the data points. According to the definition of $L_{2,1}$-norm and the formulations presented in [38], the RI L_1-norm total scatter value is defined as follows:

$$\sum_{i=1}^N \left\|\left(x_i^T - \bar{x}^T\right)U\right\|_2 = \left\|\begin{matrix}\left(x_1^T - \bar{x}^T\right)U \\ \cdots \\ \left(x_N^T - \bar{x}^T\right)U\end{matrix}\right\|_{2,1}$$
$$= \mathrm{tr}\left(U^T X_{Rt} D_{Rt} X_{Rt}^T U\right) = \mathrm{tr}\left(U^T S_{Rt} U\right) \tag{8}$$

TABLE II
ALGORITHM STEPS OF RI SUBSPACE LEARNING FRAMESORK

Input:: Samples $X=[x_1,x_2,...,x_N]$, the numbers of iterations T (and neighborhood size if needed)
Output: Low-dimensional features y_i（$i=1,2,...,N$）
Step 1: Construct matrix X_w and X_b,
Step 2: Initialize U as arbitrary columnly-orthogonal $m\times d$ matrix.
Step 3: For $j=1:T$ do
-Compute the matrix D_w and D_b, compute $X_bG_bD_bG_bX_b^T$ and $X_wG_wD_wG_wX_w^T$.
-Solve the eigenequation of (29) to obtain the eigenvectors $[u_1,u_2,...,u_d]$,
-Update $U\leftarrow[u_1,u_2,...,u_d]$.
Step 4: Output the final matrix U for feature analysis.
Step 5: Project the samples onto the low-dimensional subspace to obtain $y_i=U^Tx_i$（$i=1,2,...,N$） for classification.

TABLE III
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, AND DIMENSION] OF DIFFERENT ALGORITHMS ON FERET DATA SET

L	PCA	LPP	LDA	MFA	RIPCA	RILPP	RILDA	RIMFA
3	50.27±8.66 (185)	46.25±7.60 (200)	68.55±6.41 (70)	69.28±4.33 (45)	71.93±3.92 (195)	45.59±7.83 (200)	83.76±3.18 (25)	81.86±3.75 (25)
4	59.90±5.32 (195)	61.42±5.68 (185)	73.15±6.37 (45)	73.24±6.12 (60)	73.62±4.99 (190)	60.21±5.69 (200)	86.87±4.66 (25)	87.20±2.70 (25)
5	62.10±6.05 (185)	73.66±5.15 (190)	79.28±6.36 (50)	79.53±6.45 (40)	78.43±3.68 (195)	75.21±5.41 (200)	89.20±5.40 (30)	89.20±3.43 (30)

where $\bar{x}$ denotes the mean of the data and $S_{Rt}=X_{Rt}D_{Rt}X_{Rt}^T$ is referred to RI total scatter matrix and diagonal matrix

$$D_{Rt}=\mathrm{diag}\left(\frac{1}{2\left\|\left(x_1^T-\bar{x}^T\right)U\right\|_2},\ldots,\frac{1}{2\left\|\left(x_N^T-\bar{x}^T\right)U\right\|_2}\right)$$

and

$$X_{Rt}=((x_1-\bar{x}),\ldots,(x_N-\bar{x})).$$

PCA seeks projection directions with maximal variances. In other words, it finds and removes the projection direction with minimal variance [13]. Similarly, RIPCA also finds the minimal variance directions and discards them in dimensionality reduction. We alternatively use the minimum optimization problem (9) for the further analyses in this paper. Thus, the objective function of RIPCA can be presented as

$$\min_U \mathrm{tr}\left(U^TX_{Rt}D_{Rt}X_{Rt}^TU\right)\quad \text{s.t.}\quad U^TU=I. \tag{9}$$

As a result, the optimal projections needed to be removed in RIPCA are the eigenvectors corresponding to the smallest eigenvalus of the following standard eigenequation:

$$X_{Rt}D_{Rt}X_{Rt}^TU=U\Lambda \tag{10}$$

where $U=[u_1,u_2,\ldots,u_{m-d},u_{m-d+1},\ldots,u_m]$ is an $m\times m$ matrix containing all the eigenvectors corresponding to the eigenvalue matrix $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_{m-d},\lambda_{m-d+1},...,\lambda_m)$, where $\lambda_1\le\lambda_2\le...\le\lambda_m$. Then the optimal projection matrix of RIPCA is $U^*=[u_{m-d+1},\ldots,u_m]$. In other words, the projections $u_1,u_2,\ldots,u_{m-d}$ that characterize the minimal variance should be discarded. Thus, one can get the low-dimensional (i.e., d dimensional) representation for the original sample x by using $y=U^{*T}x$. It is easy to find that the D_{Rt} in (10) is related to the variable U, thus an iterative strategy is used to compute the optimal U^*. The computational procedures are shown in Table III. Since all the proposed methods are in the same way, we will not mention this point if it is not necessary.

D. RILDA

Similar to classical LDA algorithms, RILDA needs to define the RI between-class and within-class scatter values. The RI between-class scatter value is defined as

$$\begin{aligned}\sum_{i=1}^{c}N_i\left\|\left(\bar{x}_i^T-\bar{x}^T\right)U\right\|_2&=\left\|\begin{matrix}N_1\left(\bar{x}_1^T-\bar{x}^T\right)U\\\cdots\\N_c\left(\bar{x}_c^T-\bar{x}^T\right)U\end{matrix}\right\|_{2,1}\\&=\mathrm{tr}\left(U^TX_{Rb}D_{Rb}X_{Rb}^TU\right)=\mathrm{tr}\left(U^TS_{Rb}U\right)\end{aligned} \tag{11}$$

where $\bar{x}_i$ is the mean of the ith class, $S_{Rb}=X_{Rb}D_{Rb}X_{Rb}^T$ is referred to RI between-class scatter matrix, and the data matrix

$$X_{Rb}=(N_1(\bar{x}_1-\bar{x}),\ldots,N_c(\bar{x}_c-\bar{x}))$$

the diagonal matrix

$$D_{Rb}=\mathrm{diag}\left(\frac{1}{2\left\|N_1\left(\bar{x}_1^T-\bar{x}^T\right)U\right\|_2},\ldots,\frac{1}{2\left\|N_c\left(\bar{x}_c^T-\bar{x}^T\right)U\right\|_2}\right).$$

Similarly, the within-class scatter value is defined as

$$\begin{aligned}\sum_{i=1}^{c}\sum_{j=1}^{N_i}\left\|\left(x_i^j-\bar{x}_i\right)^TU\right\|&=\left\|X_{Rw}^TU\right\|_{2,1}\\&=\mathrm{tr}\left(U^TX_{Rw}D_{Rw}X_{Rw}^TU\right)\\&=\mathrm{tr}\left(U^TS_{Rw}U\right)\end{aligned} \tag{12}$$

where x_i^j denotes the jth sample in ith class, $S_{Rw} = X_{Rb} D_{Rw} X_{Rb}^T$ is called RI within-class scatter matrix, the diagonal matrix

$$D_{Rw} = \operatorname{diag}\left(\frac{1}{2\left\| \left(x_1^1 - \bar{x}_1\right)^T U \right\|_2}, \ldots, \frac{1}{2\left\| \left(x_1^{N_1} - \bar{x}_{N_1}\right)^T U \right\|_2}, \right.$$
$$\left. \ldots, \frac{1}{2\left\| \left(\bar{x}_c^{N_1} - \bar{x}_c\right)^T U \right\|_2}, \ldots, \frac{1}{2\left\| \left(\bar{x}_c^{N_c} - \bar{x}_c\right)^T U \right\|_2} \right)$$

and the data matrix

$$X_{Rw} = \left(\left(x_1^1 - \bar{x}_1\right), \ldots, \left(x_1^{N_1} - \bar{x}_{N_1}\right), \ldots, \left(x_c^1 - \bar{x}_c\right), \ldots, \left(x_c^{N_c} - \bar{x}_c\right) \right).$$

The common optimization criterions in classical discriminant analysis include maximizing the ratio of between-class scatter vs. within-class scatter or minimizing the ratio of within-class scatter vs. between-class scatter [46]. We take the later strategy in this paper and the objective function of RILDA can be written as

$$\begin{aligned} \min_{U^T U = I} \quad & \operatorname{tr}\left(U^T X_{Rw} D_{Rw} X_{Rw}^T U\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(U^T X_{Rb} D_{Rb} X_{Rb}^T U\right) = \text{cons.} \end{aligned} \tag{13}$$

Equation (13) can be rewritten as the following optimization problem:

$$\min_{U^T U = I} \frac{\operatorname{tr}\left(U^T X_{Rw} D_{Rw} X_{Rw}^T U\right)}{\operatorname{tr}\left(U^T X_{Rb} D_{Rb} X_{Rb}^T U\right)}. \tag{14}$$

As a result, the optimal projections of RILDA are the eigenvectors corresponding to the smallest eigenvalue of

$$\left(X_{Rb} D_{Rb} X_{Rb}^T\right)^{-1} X_{Rw} D_{Rw} X_{Rw}^T U = U\Lambda \tag{15}$$

which can be solved by the standard eigen-decomposition of the matrix $(X_{Rb} D_{Rb} X_{Rb}^T)^{-1} X_{Rw} D_{Rw} X_{Rw}^T$.

E. RILPP

LPP aims to preserve the local geometric structure in the low-dimensional subspace. In LPP, the graph is defined as

$$W_{ij} = \begin{cases} \exp\left(-\left\| x_i - x_j \right\|^2 / t\right) \text{ or } 1, & \text{if } x_i \in N_k(x_j) \\ & \text{or } x_j \in N_k(x_i) \\ 0, & \text{otherwise} \end{cases}$$

where t is the heat kernel parameter to measure the local similarity. In RILPP, the local geometric structure is preserved by minimizing the weighted RI L_1-norm, which is different from LPP. RILPP aims to minimize the following numerator:

$$\begin{aligned} &\sum_i \sum_j \left\| \left(x_i - x_j\right)^T U \right\|_2 W_{ij} \\ &\quad = \sum_i \sum_j \left\| W_{ij} \left(x_i - x_j\right)^T U \right\|_2 = \left\| G_{Rl} X_R^T U \right\|_{2,1} \\ &\quad = \operatorname{tr}\left(U^T X_R G_{Rl} D_{Rl} G_{Rl} X_R^T U\right) = \operatorname{tr}\left(U^T S_{Rl} U\right) \end{aligned} \tag{16}$$

where $S_{Rl} = X_R^T G_{Rl} D_{Rl} G_{Rl} X_R$ is called RI local scatter matrix, and the data matrix

$$X_R = ((x_1 - x_1), \ldots, (x_1 - x_N), \ldots, (x_N - x_1), \ldots, (x_N - x_N))^T$$

and

$$G_{Rl} = \operatorname{diag}(W_{11}, W_{12}, \ldots, W_{1N}, \ldots, W_{N1}, W_{N2}, \ldots, W_{NN}),$$
$$D_{Rl} = \operatorname{diag}\left(\frac{1}{2W_{11}\left\| (x_1 - x_1)^T U \right\|_2}, \ldots, \frac{1}{2W_{1N}\left\| (x_1 - x_N)^T U \right\|_2}, \right.$$
$$\ldots, \frac{1}{2W_{N1}\left\| (x_N - x_1)^T U \right\|_2},$$
$$\left. \ldots, \frac{1}{2W_{NN}\left\| (x_N - x_N)^T U \right\|_2} \right).$$

Note that, in this paper we define $(1/(2W_{ii}\|(x_i - x_i)^T U\|_2)) \times (x_i - x_i) = \infty \times 0 \triangleq 0$, thus this does not affect the computation in (16) (similar cases exist in RIMFA).

From the local neighborhood graph W, we can obtain the diagonal matrix D with the diagonal elements $D_{ii} = \sum_i W_{ij}$. The imposed constraint part of RILPP is measured by

$$\begin{aligned} \sum_i \left\| x_i^T U \right\|_2 D_{ii} &= \sum_i \left\| D_{ii} x_i^T U \right\|_2 = \left\| G_{Rd} X^T U \right\|_{2,1} \\ &= \operatorname{tr}\left(U^T X G_{Rd} D_{Rd} G_{Rd} X^T U\right) = \operatorname{tr}\left(U^T S_{Rd} U\right) \end{aligned} \tag{17}$$

where

$$G_{Rd} = \operatorname{diag}(D_{11}, D_{22}, \ldots, D_{NN})$$
$$S_{Rd} = X G_{Rd} D_{Rd} G_{Rd} X^T$$

and the diagonal matrix

$$D_{Rd} = \operatorname{diag}\left(\frac{1}{2D_{11}\left\| x_1^T U \right\|_2}, \frac{1}{2D_{22}\left\| x_2^T U \right\|_2}, \right.$$
$$\left. \ldots, \frac{1}{2D_{NN}\left\| x_N^T U \right\|_2} \right).$$

With the above preparations, the objective function of RILPP can be stated as

$$\min_{U^T U = I} \frac{\operatorname{tr}\left(U^T X_R G_{Rl} D_{Rl} G_{Rl} X_R^T U\right)}{\operatorname{tr}\left(U^T X G_{Rd} D_{Rd} G_{Rd} X^T U\right)}. \tag{18}$$

The optimal solutions can be obtained by the eigen-decomposition of

$$\left(X G_{Rd} D_{Rd} G_{Rd} X^T\right)^{-1} X_R G_{Rl} D_{Rl} G_{Rl} X_R^T U = U\Lambda. \tag{19}$$

The optimal projections of RILPP are the eigenvectors corresponding to the smaller eigenvalures of (19).

F. RIMFA

MFA was introduced in [13] as a representative algorithm for the graph embedding framework, where the squared distance between the point pairs in local neighborhood was measured. When using the RI L_1-norm, we can obtain significantly different graph embedding method. The interclass separability and the compactness of RIMFA by using the RI L_1-norm are denoted as S_{Rp} and S_{Rc}, respectively. With the same definition of the penalty graph W^p and the compactness graph W^c as in MFA, the interclass separability S_{Rp} is defined as follows:

$$\sum_i \sum_{(i,j)\in P_{k_1}(c_i) \text{ or } (i,j)\in P_{k_1}(c_j)} \left\|U^T x_i - U^T x_j\right\|_2 W_{ij}^p = \operatorname{tr}\left(U^T X_R G_{Rp} D_{Rp} G_{Rp} X_R^T U\right) = \operatorname{tr}\left(U^T S_{Rp} U\right) \quad (20)$$

where

$$\begin{aligned} S_{Rp} &= X_R G_{Rp} D_{Rp} G_{Rp} X_R^T \\ G_{Rp} &= \operatorname{diag}\left(W_{11}^p, W_{12}^p, \ldots, W_{1N}^p, \ldots, W_{N1}^p, W_{N2}^p, \ldots, W_{NN}^p\right) \\ D_{Rp} &= \operatorname{diag}\left(\frac{1}{2W_{11}^p\left\|(x_1-x_1)^T U\right\|_2}, \ldots, \frac{1}{2W_{1N}^p\left\|(x_1-x_N)^T U\right\|_2}, \right. \\ &\qquad \ldots, \frac{1}{2W_{N1}^p\left\|(x_N-x_1)^T U\right\|_2}, \\ &\qquad \left. \ldots, \frac{1}{2W_{NN}^p\left\|(x_N-x_N)^T U\right\|_2}\right). \end{aligned}$$

Similarly, the intraclass compactness S_{Rc} is defined as follows:

$$S_{Rc} = \sum_i \sum_{i\in N_{k_2}^+(j) \text{ or } j\in N_{k_2}^+(i)} \left\|x_i^T U^T - x_j^T U^T\right\|_2 W_{ij}^c = \operatorname{tr}\left(U^T X_R G_{Rc} D_{Rc} G_{Rc} X_R^T U\right) = \operatorname{tr}\left(U^T S_{Rc} U\right) \quad (21)$$

where

$$\begin{aligned} S_{Rc} &= X_R G_{Rc} D_{Rc} G_{Rc} X_R^T \\ G_{Rc} &= \operatorname{diag}\left(W_{11}^c, W_{12}^c, \ldots, W_{1N}^c, \ldots, W_{N1}^c, W_{N2}^c, \ldots, W_{NN}^c\right) \\ D_{Rc} &= \operatorname{diag}\left(\frac{1}{2W_{11}^c\left\|(x_1-x_1)^T U\right\|_2}, \ldots, \frac{1}{2W_{1N}^c\left\|(x_1-x_N)^T U\right\|_2}, \right. \\ &\qquad \ldots, \frac{1}{2W_{N1}^c\left\|(x_N-x_1)^T U\right\|_2}, \\ &\qquad \left. \ldots, \frac{1}{2W_{NN}^c\left\|(x_N-x_N)^T U\right\|_2}\right). \end{aligned}$$

With the above preparation, we obtain the objective function of RIMFA

$$\min_{U^T U=I} \frac{\operatorname{tr}\left(U^T X_R G_{Rc} D_{Rc} G_{Rc} X_R^T U\right)}{\operatorname{tr}\left(U^T X_R G_{Rp} D_{Rp} G_{Rp} X_R^T U\right)}. \quad (22)$$

The optimal projections are the eigenvectors corresponding to the smaller eigenvalues of the following eigenequation:

$$\left(X_R G_{Rp} D_{Rp} G_{Rp} X_R^T\right)^{-1} X_R G_{Rc} D_{Rc} G_{Rc} X_R^T U = U\Lambda. \quad (23)$$

G. Unified Framework

In this subsection, we present the unified framework based on the RI L_1-norm. It is easy to find that the objective function of RIPCA can be rewritten as

$$\begin{aligned} \min_{U^T U=I} \quad & \operatorname{tr}\left(U^T X_{Rt} G_I D_{Rt} G_I X_{Rt}^T U\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(U^T X_I G_I D_I G_I X_I^T U\right) = \text{cons} \end{aligned} \quad (24)$$

where X_I, G_I and D_I are $m \times m$ identity matrix

Similarly, the objective function of RILDA can be represented as

$$\begin{aligned} \min_{U^T U=I} \quad & \operatorname{tr}\left(U^T X_{Rw} G_{Iw} D_{Rw} G_{Iw} X_{Rw}^T U\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(U^T X_{Rb} G_{Ib} D_{Rb} G_{Ib} X_{Rb}^T U\right) = \text{cons} \end{aligned} \quad (25)$$

where G_{Iw} and G_{Ib} are the identity matrix with a suitable size.

Comparing the above rewritten RIPCA and RILDA models and the ones of RILPP and RIMFA, we obtain a unified framework as follows to conclude all the methods presented in the above four sections:

$$\begin{aligned} \min_{U^T U=I} \quad & \operatorname{tr}\left(U^T X_w G_w D_w G_w X_w^T U\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(U^T X_b G_b D_b G_b X_b^T U\right) = \text{cons} \end{aligned} \quad (26)$$

where $X_w = (\tilde{x}_1, \ldots, \tilde{x}_{N_w})$ and $X_b = (\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_{N_b})$ uniformly denote the new data matrices in the optimization problems in (18) and (23)–(25) , respectively. And the diagonal matrices and the weight matrices defined on the new datasets of X_w and X_b corresponding to the same way as in RIPCA, RILDA, RILPP, and RIMFA are uniformly denoted as D_w, G_w and D_b, G_b, respectively. The summation of the algorithms on the proposed framework is shown in Table I.

Thus, comparing (5) and (6) with (26), we can find that the graph embedding framework is extended to more generalized form by using the RI L_1-norm, which can be called generalized graph embedding algorithm framework. The proposed framework can be rewritten as

$$\min_{U^T U=I} \frac{\operatorname{tr}\left(U^T X_w G_w D_w G_w X_w^T U\right)}{\operatorname{tr}\left(U^T X_b G_b D_b G_b X_b^T U\right)} \quad (27)$$

or

$$\min_{\substack{\operatorname{tr}\left(U^T X_b G_b D_b G_b X_b^T U\right)=\text{cons} \\ U^T U=I}} \operatorname{tr}\left(U^T X_w G_w D_w G_w X_w^T U\right). \quad (28)$$

Similarly, the optimal solution of the above problem can be obtained by the eigen-decomposition of eigenequation

$$\left(X_b G_b D_b G_b X_b^T\right)^{-1} X_w G_w D_w G_w X_w^T U = U\Lambda. \quad (29)$$

Differing from the graph embedding framework which only needs to solve a simple (standard or generalized) eigenequation, the proposed methods need to iteratively solve a series of (standard) eigenequations since the D_w and D_b are correlated to the U in (29) in each iteration. The procedures of the proposed RI subspace learning algorithm framework are shown in Table II.

Comparing to the previous graph embedding linear dimensionality reduction framework, the advantage of the generalized framework proposed in this paper is that it can

automatically generate the weight matrices derived by the $L_{2,1}$ norm in the iterative procedure. These weight matrices are closely related to the intrinsic relationship among the data points and thus can discover a more effective subspace for feature extraction and classification.

The proposed methods are significantly different from the one in reference [42], which mainly focus on the usage of the $L_{2,1}$ norm as a regularized term for jointly sparse regression instead of as a basic metric. This paper also differs from [25] using the L_1 norm as the measurement, and only the LDA case is discussed in [25]. Moreover, the optimization methods of this paper and [25] are also diverse far from each other. Although both this paper and [47] introduce the locality of the data, the goal of [47] is to learn a graph to preserve the locality instead of to learn a projection matrix for linear dimensionality reduction. In short, the goal, the objective function and optimization method of the proposed methods are all different from those in [25], [42], and [47].

III. Theoretical Analysis

In this section, theoretical analyses will be presented to show the properties of the proposed algorithm framework. The convergent analysis and the computational complexity are firstly presented. And then the essence of the framework is explored. We discuss a special case to explore the equivalence between the RI subspace learning framework and the previous graph embedding algorithm framework. At last, it is shown that the proposed algorithm framework has the global optimal solution when two scatter matrices are full rank matrices.

A. Convergence of the Algorithm

Since the algorithms are iterative methods, in this section, we analyze its convergence. Firstly, we need the following lemma and corollary presented in [38].

Lemma 1: For any nonzero vector a and b, the following inequality holds:

$$\|a\| - \frac{\|a\|^2}{2\|b\|} \le \|b\| - \frac{\|b\|^2}{2\|b\|}. \tag{30}$$

Corollary 1: For any nonzero vectors $a^i, b^i (i = 1, 2, \ldots, N)$, the following inequality holds:

$$\sum_i \|a^i\| - \sum_i \frac{\|a^i\|^2}{2\|b^i\|} \le \sum_i \|b^i\| - \sum_i \frac{\|b^i\|^2}{2\|b^i\|}. \tag{31}$$

Theorem 1: The iterative algorithm for the RI L_1-norm will monotonically decrease the objective function value in each iteration.

Proof: For simplicity, we denote the ith row of matrix A by A^i, i.e., $A^i = A(i, :)$. Supposed in the t iteration we have

$$U_t = \arg\min_{\substack{\mathrm{tr}(U^T X_b G_b D_b^t G_b X_b^T U)=\mathrm{cons} \\ U^T U = I}} \mathrm{tr}\left(U^T X_w G_w D_w^t G_w X_w^T U\right).$$

This indicates that

$$\begin{aligned}
&\mathrm{tr}\left(U_t^T X_w G_w D_w^t G_w X_w^T U_t\right) \\
&\quad \le \mathrm{tr}\left(U_{t-1}^T X_w G_w D_w^t G_w X_w^T U_{t-1}\right) \\
&\quad \Rightarrow \sum_i \frac{\left\|\left(G_w X_w^T U_t\right)^i\right\|^2}{2\left\|\left(G_w X_w^T U_{t-1}\right)^i\right\|} \le \sum_i \frac{\left\|\left(G_w X_w^T U_{t-1}\right)^i\right\|^2}{2\left\|\left(G_w X_w^T U_{t-1}\right)^i\right\|} \\
&\quad \Rightarrow \sum_i \left\|\left(G_w X_w^T U_t\right)^i\right\| \\
&\qquad - \left(\sum_i \left\|\left(G_w X_w^T U_t\right)^i\right\| - \sum_i \frac{\left\|\left(G_w X_w^T U_t\right)^i\right\|^2}{2\left\|\left(G_w X_w^T U_{t-1}\right)^i\right\|}\right) \\
&\quad \le \sum_i \left\|\left(G_w X_w^T U_{t-1}\right)^i\right\| \\
&\qquad - \left(\sum_i \left\|\left(G_w X_w^T U_{t-1}\right)^i\right\| - \sum_i \frac{\left\|\left(G_w X_w^T U_{t-1}\right)^i\right\|^2}{2\left\|\left(G_w X_w^T U_{t-1}\right)^i\right\|}\right).
\end{aligned}$$

According to the Corollary 1, we get the following inequality:

$$\begin{aligned}
&\sum_i \left\|\left(G_w X_w^T U_t\right)^i\right\| \le \sum_i \left\|\left(G_w X_w^T U_{t-1}\right)^i\right\| \\
&\Rightarrow \left\|G_w X_w^T U_t\right\|_{2,1} \le \left\|G_w X_w^T U_{t-1}\right\|_{2,1}.
\end{aligned} \tag{32}$$

This gives

$$\mathrm{tr}\left(U_t^T X_w G_w D_w^{t+1} G_w X_w^T U_t\right) \le \mathrm{tr}\left(U_{t-1}^T X_w G_w D_w^t G_w X_w^T U_{t-1}\right).$$

That is to say the objective function value $\mathrm{tr}(U^T X_w G_w D_w^t G_w X_w^T U)$ with the constraint of $\mathrm{tr}(U^T X_b G_b D_b G_b X_b^T U) = \mathrm{cons}$ will monotonically decrease in the iteration. ■

According to Theorem 1, we can conclude that the iterative procedures will converge to the local optimal U of problems (26) or (28).

B. Computational Complexity Analysis

The main computational complexity of the RI L_1-norm algorithm is to solve the eigenequation of (29). Thus, the computational complexity is $O(m^3)$ for each iteration. If the algorithm needs T iteration steps, then the total computational complexity is $O(Tm^3)$. As demonstrated in the experimental section, the algorithms usually converge very fast. Thus T is usually a smaller number. Therefore, although the computational complexity of the RI L_1-norm is larger than those of previous algorithms, the proposed algorithms are still efficient in most cases.

C. Essence of the Framework

In Section II-G, a unified framework is proposed for RI subspace learning. The following theorem reveals the essence of the RI subspace learning framework.

Theorem 2: The essences of the RI subspace learning methods (RIPCA, RILDA, RILPP, and RIMFA) are the reweighted versions of the corresponding methods (PCA, LDA, LPP, and MFA, respectively) by rescaling the weights, which are related

to the original high-dimensional data points and the learned low-dimensional subspace.

Proof: From (26) we can define

$$\tilde{W}^w \triangleq G_w D_w G_w = \text{diag}\left(\frac{W^w_{11}}{2\left\|\tilde{x}_1^T U\right\|}, \ldots, \frac{W^w_{N_w N_w}}{2\left\|\tilde{x}_{N_w}^T U\right\|}\right) \tag{33}$$

$$\underset{\sim}{W}^b \triangleq G_b D_b G_b = \text{diag}\left(\frac{W^b_{11}}{2\left\|\underset{\sim}{x}_1^T U\right\|}, \ldots, \frac{W^b_{N_b N_b}}{2\left\|\underset{\sim}{x}_{N_b}^T U\right\|}\right) \tag{34}$$

where diagonal matrices W^w and W^b are weights on new data sets X_w and X_b, respectively. Then the two terms in (26) can be rewritten as

$$\text{tr}\left(U^T X_w G_w D_w G_w X_w^T U\right) = \sum_i^{N_w} \left\|\tilde{x}_i^T U\right\|^2 \tilde{W}^w_{ii} \tag{35}$$

$$\text{tr}\left(U^T X_b G_b D_b G_b X_b^T U\right) = \sum_i^{N_b} \left\|\underset{\sim}{x}_i^T U\right\|^2 \underset{\sim}{W}^b_{ii}. \tag{36}$$

These two equations indicate that the scatter values are the sum of the rescaled scatter of X_w and X_b by the weighted matrices $\tilde{W}^w$ and $\underset{\sim}{W}^b$, which are decided by the original high-dimensional data points and the learned low-dimensional subspaces. ■

From Theorem 2, we can find that the RI subspace learning framework is the reweighted version of the previous graph embedding framework.

D. Special Case 1: The Equivalence to the Graph Embedding Algorithm Framework

In this section, we discuss the special case to explore the equivalence between the RI subspace learning framework and the previous graph embedding algorithm framework by ignoring the null space of the data (the original data can be preprocessed by PCA to remove the null space of the data, which will be discussed in the next section). We have the following theorem.

Theorem 3: Suppose $N = d \ll m$ and U is an $m \times d$ orthogonal matrix spanned by the original data space. If we define two special graphs using the following weights:

$$\tilde{W}^w_{ii} = \left\|\tilde{x}_i^T\right\| * \text{cons1}, \underset{\sim}{W}^b_{ii} = \left\|\underset{\sim}{x}_i\right\| * \text{cons2} \tag{37}$$

where cons1 and cons2 are two constants, then the RI subspace learning algorithm framework is equivalent to the graph embedding algorithm framework.

Proof: From Theorem 2, it is easy to find that we only need to show whether if there exist some kind of definitions such that

$$\begin{aligned}\tilde{W}^w &= G_w D_w G_w = \text{diag}\left(\frac{W^w_{11}}{2\left\|\tilde{x}_1^T U\right\|}, \ldots, \frac{W^w_{N_w N_w}}{2\left\|\tilde{x}_{N_w}^T U\right\|}\right) \\ &= I * \text{cons1}\end{aligned} \tag{38}$$

and

$$\begin{aligned}\underset{\sim}{W}^b &= G_b D_b G_b = \text{diag}\left(\frac{W^b_{11}}{2\left\|\underset{\sim}{x}_1^T U\right\|}, \ldots, \frac{W^b_{N_b N_b}}{2\left\|\underset{\sim}{x}_{N_b}^T U\right\|}\right) \\ &= I * \text{cons2}.\end{aligned} \tag{39}$$

Since U is a $d \times m$ orthogonal matrix spanned by the whole data space, in this special case, it is easy to see that

$$\begin{aligned}\left\|\tilde{x}_1^T U\right\| &= \left\|\tilde{x}_1^T\right\|, \ldots, \left\|\tilde{x}_{N_b}^T U\right\| = \left\|\tilde{x}_{N_b}^T\right\| \\ \left\|\underset{\sim}{x}_1^T U\right\| &= \left\|\underset{\sim}{x}_1^T\right\|, \ldots, \left\|\underset{\sim}{x}_{N_w}^T U\right\| = \left\|\underset{\sim}{x}_{N_w}^T\right\|.\end{aligned}$$

If and only if we define $W^w_{ii} = \|\tilde{x}_i^T\| * \text{cons1}(i = 1, 2, \ldots, N_w)$, $W^b_{ii} = \|\underset{\sim}{x}_i^T\| * \text{cons2}(i = 1, 2, \ldots, N_b)$, then we have the following relationship from (18), (19), (35), (36), (38), and (39):

$$X_w G_w D_w G_w X_w^T \propto X\left(\bar{D}^w - \bar{W}^w\right)X^T \tag{40}$$

$$X_b G_b D_b G_b X_b^T \propto X\left(\bar{D}^b - \bar{W}^b\right)X^T. \tag{41}$$

Therefore, these two frameworks are equivalent. ■

It is obvious that if U does not full fill the data space, then $\|\tilde{x}_i^T U\| \neq \|\tilde{x}_i^T\|$ and $\|\underset{\sim}{x}_i^T U\| \neq \|\underset{\sim}{x}_i^T\|$. Thus, (40) and (41) will not be satisfied. In this case, the derived subspaces (projective vectors) of these two frameworks will be different from each other.

E. Special Case 2: Global Optimal Solution

Since the original data can be preprocessed by PCA (or RIPCA) such that $X_w G_w D_w G_w X_w^T$ and $X_b G_b D_b G_b X_b^T$ are full rank matrices (or they can add matrix εI so that they are full rank matrices, where ε is a sufficiently small real number). Without loss of generality, we suppose these two matrices are full rank matrices in the RI subspace learning framework. Then, the proposed algorithm framework has an elegant property, which exists in the classic subspace learning methods such as LDA and LPP. That is, the proposed algorithm framework has a global optimal solution when all the projections corresponding to the data space are computed.

Theorem 4: Suppose $X_w G_w D_w G_w X_w^T$ and $X_b G_b D_b G_b X_b^T$ are full rank matrices. Since U is an $m \times m$ orthogonal matrix (in the iteration), then the rotation invariant subspace learning algorithm framework (29) has a global optimal solution (up to a rotation matrix), and thus the iterative procedures can be avoided.

Proof: We can initialize the U as the SVD of $X_b G_b D_b G_b X_b^T$, i.e., $X_b G_b D_b G_b X_b^T = \hat{U}\hat{D}\hat{V}^T$ and let $U = \hat{U}$. Since all the eigenvectors of the eigenequation (29) span the same subspace as U and the eigenvectors are orthogonal, it is obvious that $(W^w_{ii}/2\|\tilde{x}_i^T U\|)$ and $(W^b_{ii}/2\|\underset{\sim}{x}_i^T U\|)$ are constants. As a result, $X_w G_w D_w G_w X_w^T$ and $X_b G_b D_b G_b X_b^T$ are also invariant in the iterative procedures, thus the eigenvectors are also invariant (up to a rotation matrix). Since the optimization problem is convex, it has a global optimal solution which can be obtained by solving the eigenequation (29) only once. ■

Theorem 4 shows that in this special case, the number of iteration is 1, which indicates that no any iteration step is needed. Therefore, the algorithms will become simpler and more efficient. In this special case, the essential difference between the graph embedding algorithm framework and the RI subspace learning framework is that the later needs an orthogonal matrix for initialization so as to compute the eigenvectors but the former can directly compute the eigenvectors. Another difference is the RI subspace learning framework learns the orthogonal projections by standard eigen-decomposition

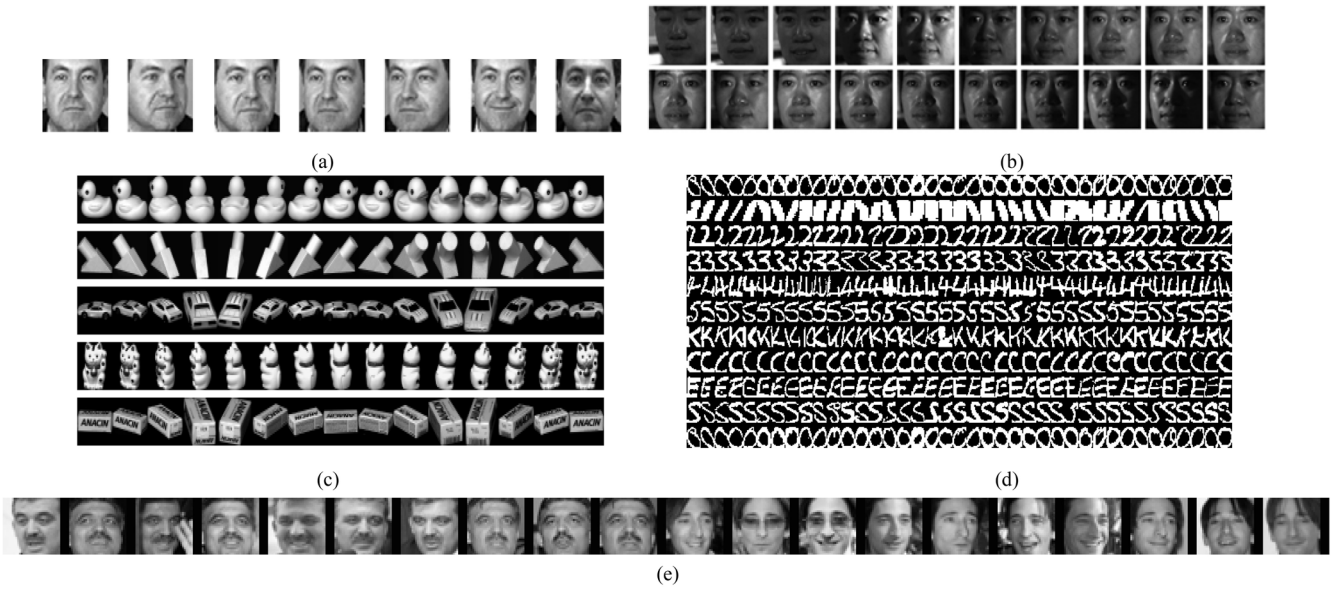


Fig. 2. Image samples used in the experiments. (a) FERET face database. (b) CMU PIE face database. (c) COIL100 objective image database. (d) Binary alpha image database. (e) LFW face database.

while the graph embedding framework obtains the nonorthogonal projections since it solves the generalized eigen-function.

IV. Experiments

In this section, a set of experiments are presented to show the effectiveness of the proposed RI subspace learning algorithms for image feature extraction and recognition against the classical subspace learning methods (i.e., PCA and LDA), the most related manifold learning methods (i.e., LPP and MFA). The FERET face database is used to explore the robustness of the proposed RI subspace learning algorithms on the variations in expressions and illumination. The CMU PIE (Pose29, light, and illumination change) face database is used to evaluate the performance of these methods when face poses and lighting conditions vary dramatically. The COIL100 objective image database is employed to test the performance of these algorithms when there are rotational variations. The binary alpha digits image database and label faces in the wild (LFW) database [48] are used to test the robustness of the proposed algorithms when there are very similar images of different objectives. The nearest neighbor classifier with the Euclidean distance is used in all the experiments. The MATLAB codes of the proposed methods can be available from http://www.scholat.com/laizhihui.

A. Description of the Databases

The FERET face database is a result of the FERET program, which was sponsored by the U.S. Department of Defense through the DARPA program [49]. It has become a standard database for testing and evaluating state-of-the-art face recognition algorithms. The proposed method was tested on a subset of the FERET database. This subset includes 1400 images of 200 individuals (each individual has seven images) and involves variations in facial expression, illumination, and pose. In the experiment, the facial portion of each original image was automatically cropped based on the location of the eyes, and the cropped images was resized to 40×40 pixels. The sample images of one person are shown in Fig. 2(a).

The CMU PIE face database [50] contains 68 individual with 41 368 face images as a whole. The face images were captured under varying pose, illumination and expression. In our experiments, we select a subset (C29) which contains1632 images of 68 individuals (each individual has 24 images). The C29 subset involves variations in illumination, facial expression and pose. All of these face images are aligned based on eye coordinates and cropped to 32×32. Fig. 2(b) shows the sample images from this database.

The COIL100 objective image database (http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php) consists of $100 \times 72 = 7200$ images of 100 objects where the images of the object were taken at pose intervals of 5°, i.e., 72 poses per object. The original mages were normalized to be 128×128 pixels. Some sample images of five objects are shown in Fig. 2(c). All images are converted to a gray-scale image of 32×32 pixel for computational efficiency in the experiments.

Binary alpha digits image database (http://www.cs.nyu.edu/~roweis/data.html) is composed of 1404 binary images of handwritten digits from "0" to "9" and also characters from "A" to "Z," totally 36 classes. Each object has 39 images. The resolution of each image is 20×16 pixels. Some sample images are shown in Fig. 2(d).

The LFW database contains images of 5749 different individuals in unconstrained environment [48]. A subdatabase with 158 subjects from LFW-a is used in this paper. The final size of images is normalized to be 32×32 and no other preprocessing is performed on the images since we want to test the robustness of the proposed algorithms. Some sample images of LFW-a database are shown in Fig. 2(e).

TABLE IV
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, AND DIMENSION] OF DIFFERENT ALGORITHMS ON CMU PIE DATA SET

L	PCA	LPP	LDA	MFA	RIPCA	RILPP	RILDA	RIMFA
4	59.12±11.44 (145)	75.37±8.21 (145)	75.78±8.23 (67)	78.93±7.05 (150)	60.88±5.52 (145)	77.46±8.27 (150)	78.62±7.55 (145)	79.15±7.68 (150)
5	68.61±13.77 (140)	81.95±6.43 (140)	84.13±8.12 (67)	85.83±6.40 (70)	70.37±6.63 (150)	83.10±6.85 (145)	86.37±7.25 (85)	86.06±7.38 (140)
6	76.02±10.19 (140)	86.94±6.39 (145)	90.06±3.65 (67)	91.08±3.73 (110)	78.10±4.86 (145)	88.95±5.84 (140)	93.31±3.37 (60)	92.92±3.45 (60)

TABLE V
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, AND DIMENSION] OF DIFFERENT ALGORITHMS ON COIL100 DATA SET

L	PCA	LPP	LDA	MFA	RIPCA	RILPP	RILDA-1084	RIMFA
6	77.47±2.15 (30)	77.22±2.24 (24)	80.21±2.51 (19)	80.26±2.38 (17)	77.53±6.71 (30)	78.20±2.02 (25)	81.80±2.52 (18)	81.29±2.38 (17)
8	82.79±2.28 (34)	82.35±2.07 (28)	85.16±2.06 (20)	85.36±2.12 (21)	82.72±2.36 (31)	83.10±1.86 (35)	86.49±2.31 (19)	86.24±1.93 (18)
10	86.87±2.28 (38)	86.14±2.21 (29)	88.55±2.11 (21)	88.59±2.08 (20)	86.76±2.36 (41)	86.92±2.15 (45)	89.62±2.18 (18)	89.84±2.39 (15)
12	88.58±2.27 (45)	88.19±2.35 (30)	90.47±2.00 (21)	90.38±2.17 (22)	88.36±2.34 (31)	88.47±2.46 (35)	91.29±2.28 (18)	91.61±2.56 (18)

TABLE VI
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, AND DIMENSION] OF DIFFERENT ALGORITHMS ON BINARY ALPHA DIGITS DATA SET

L	PCA	LPP	LDA	MFA	RIPCA	RILPP	RILDA	RIMFA
10	62.44±1.65 (30)	62.66±3.80 (29)	61.62±1.51 (27)	63.69±1.57 (26)	64.82±1.38 (29)	63.02±1.52 (40)	64.18±1.26 (32)	63.84±2.54 (35)
15	67.19±1.37 (29)	67.05±0.86 (34)	66.48±0.62 (28)	68.68±0.91 (27)	72.24±1.20 (29)	71.65±1.23 (34)	72.39±1.52 (26)	71.39±1.25 (40)
20	69.99±1.75 (30)	71.115±1.69 (29)	70.20±1.12 (31)	72.59±2.33 (23)	78.37±2.20 (26)	78.61±2.57 (28)	78.80±1.49 (25)	78.19±1.88 (38)
25	72.48±1.72 (28)	73.75±1.29 (29)	72.70±1.55 (25)	75.22±1.58 (23)	84.16±1.80 (26)	84.12±2.00 (24)	84.03±1.75 (28)	83.80±1.55 (40)

TABLE VII
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, AND DIMENSION] OF DIFFERENT ALGORITHMS ON LFW FACE DATA SET

L	PCA	LPP	LDA	MFA	RIPCA	RILPP	RILDA	RIMFA
6	13.40±0.78 (200)	16.34±1.37 (200)	25.61±5.30 (157)	21.79±1.07 (200)	43.40±0.81 (175)	42.77±2.00 (160)	49.24±0.31 (130)	54.34±0.07 (35)

B. Experiment Setup

In the experiments, L images of each individual were randomly selected and used as the training set, and one half of the remaining images were used as the validation set and test set, respectively. The best parameters determined by the validation set were used to learning the projections for feature extraction and classification. The L was set as different numbers according the size of each individual/object on the data sets. That is, $L = 3, 4, 5$, $L = 4, 5, 6$, $L = 6, 8, 10, 12$, and $L = 10, 15, 20, 25$ for FERET, CMU PIE, COIL100, and binary alpha digits image databases, respectively.

For improving the computational efficiency and avoiding the singular problem, the data were preprocessed by using PCA to preserve most of the image energy to pursuit the best performance of each method. The neighborhood parameters were selected from the set $\{1, 2, \ldots, 6, 2^3, 2^4, \ldots, N-1\}$. The numbers of final subspace dimensions for FERET, CMU PIE and LFW-a face databases were varied from 5 to 200 with step 5. For COIL100 and binary alpha digits image databases, the numbers of the final subspace dimensions were varied from 10 to 40 with step 1 since the dimensions corresponding to the best recognition rates are in this range. In each run, the best parameters determined by the validation set were used to learn the optimal projections for feature extraction. The algorithms were independently run 10 times. And the average recognition rate, standard deviation and the corresponding dimension on the test set are reported in Tables III–VII. The recognition rates versus the variations of the dimensions are also shown in Figs. 3 and 4.

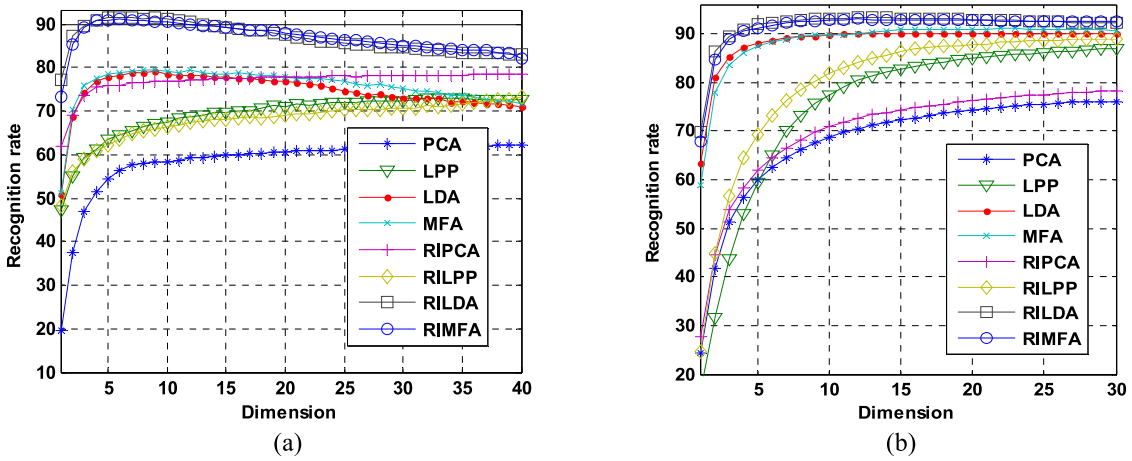


Fig. 3. Average recognition rates (%) versus the variations of the dimensions. (a) Five training samples on the FERET face database. (b) Six training samples on CMU PIE face database.

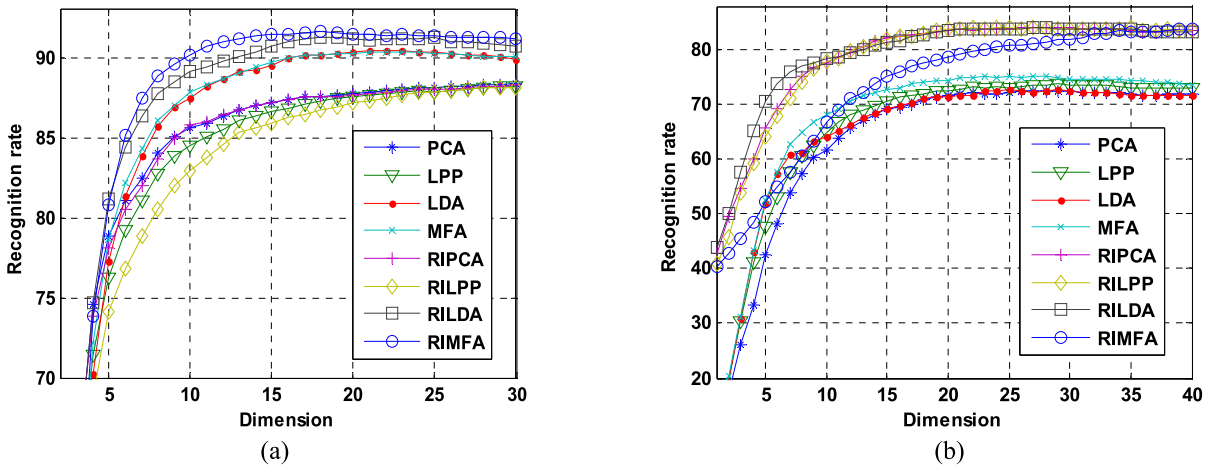
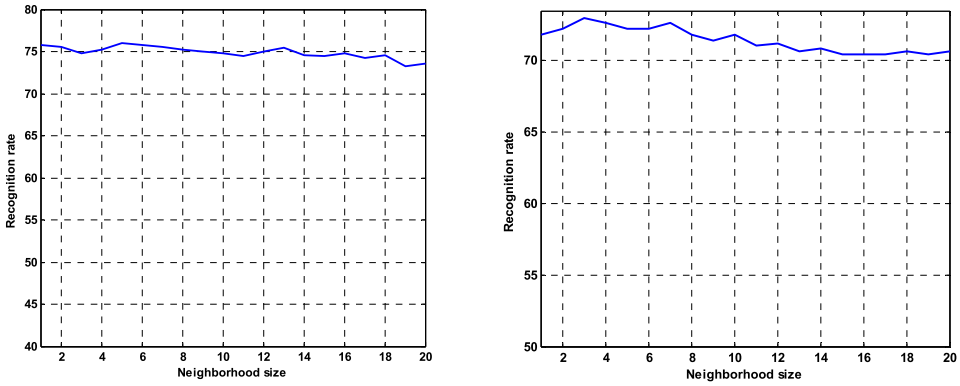


Fig. 4. Average recognition rates (%) versus the dimensions. (a) 12 training samples on the COIL100 image databases. (b) 25 training samples on binary alpha image databases.

C. Experimental Result and Analysis

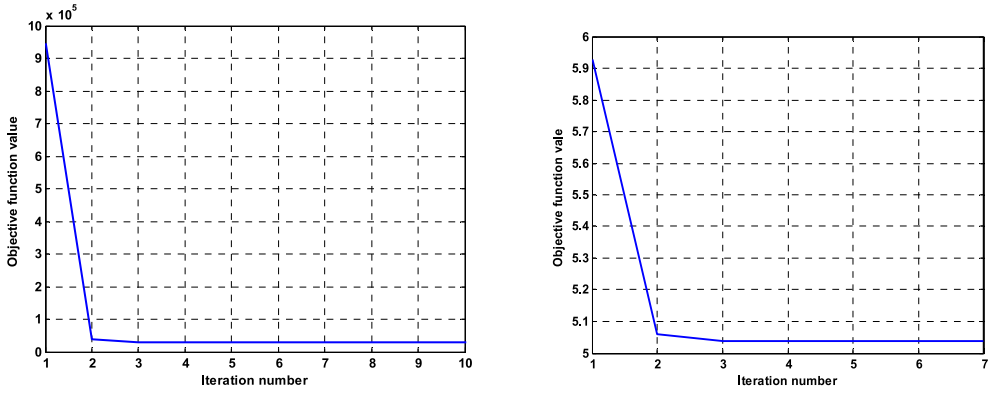
From the above tables and the figures, we can obtain some interesting observations and conclusions.

1) The RI dimensionality reduction framework using the L_1 norm or $L_{2,1}$ norm performs better than the graph embedding algorithm framework which uses the Frobenius norm as the measurement. As can be found from the top recognition rates listed in the tables, the highest recognition rate in each case is always achieved by the newly proposed RI feature extraction algorithms. In most cases, RIPCA performs better than (denote as ">") PCA, RILPP>LPP, RILDA>LDA, and RIMFA>MFA. This indicates that using RI L_1 norm or $L_{2,1}$ norm as a measurement does obtain better performance in feature extraction, no matter there are variations in expression, pose and illumination or rotations of the images. Thus the proposed RI subspace learning algorithms are more robust than the compared algorithms in these cases.
2) As can be seen from Figs. 3 and 4, RI algorithms usually achieve higher recognition rate in a lower dimensional subspace compared with the previous algorithms. This phenomenon is consistent on the four databases and more salient in FERET and COIL100 databases. This indicates that RI subspace learning methods have more powerful ability in dimensionality reduction.
3) Though the proposed algorithms perform better than the corresponding ones based on the L_2 norm, there are the cases where the formers obtain only similar performance to the latter (or even the formers obtain slightly lower recognition rates). The key reason may be the extremely lack of training samples. For example, on the FERET face database, RILPP performs better than LPP when there are five training samples per individual, but it obtains lower recognition rate when there are only three and four training samples per individual. Another important reason is that there lacks discriminant information for RILPP to guard the algorithm to assign a suitable weighted matrix D_w to enhance the discriminate ability. That is, only to enhance the robustness for preserving the locality in RILPP cannot guarantee to obtain the stronger discriminant ability than LPP in some special cases.
4) Using the discriminant information, RILDA and RIMFA usually perform better than LDA and MFA in



(a) (b)

Fig. 5. Recognition rates versus the neighborhood size on binary alpha image database. (a) RILPP. (b) RIMFA.



(a) (b)

Fig. 6. Convergence on (a) FERET face data set and (b) binary alpha data set.

all cases. This indicates that introducing the $L_{2,1}$ norm to the Fisher criterion for discriminant feature extraction can enhance the accuracy in image recognition. But using the different criterions in the RI subspace learning, the accuracies of different methods are usually different from each other. Strictly speaking, among the RI dimensionality reduction methods, there is no complete winner that can always perform the best in all kinds of scenarios (or in all the databases). The essential reason for the excellent performance of RILDA and RIMFA is that they use the label information in their criterions and a better weighted matrix D_w could be signed to the data points to reduce the influence of the outliers, or the variations of illumination, pose and expression in face images or the variation of the object's rotations.

5) For the extremely case as in LFW-a dataset, where facial images exhibit extreme pose, illumination and background variations, occlusions, and inaccurate alignment, the recognition rates reported in Table VII show that the proposed methods also perform much better than the classical methods. About 20 percentage points on the recognition rates are increased by the proposed methods. This indicates that the $L_{2,1}$ norm as a metric is more robust than L_2 norm in subspace learning for potential real world applications.

D. Parameters Sensitivity Study

In the proposed RI dimensionality reduction framework, we present four methods. Among these methods, the parameter d, i.e., the number of dimension, is an important parameter in the models. From Figs. 2–4, we can find that this parameter significantly varies on different databases. This phenomenon demonstrates that the parameter sensitivity is presumably related to the properties of the different data sets, and there is no consistent rule suitable for different database. However, another important parameter, i.e., the neighborhood size, in RILPP and RIMFA has a certain rule. Usually, as shown in Fig. 5, when the local neighborhood size is set to about 5, RILPP and RIMFA can achieve better performances. Similar properties can also be found in LPP and MFA. This indicates that RILPP and RIMFA inherit similar local geometric preserving properties from LPP and MFA.

E. Convergence Study

As indicated in the previous section, the objective function of the proposed framework will converge to the local optimum. For practical applications it is very interesting for us to show how fast our algorithm converges.

Fig. 6 shows the convergence curves of our optimization algorithm with respect to the objective function value. Fig. 6(a) and (b) shows the convergent properties of the representative method, i.e., RIMFA, on FERET and binary

alpha digits image databases. It can be found that the proposed method converges very fast. Similar properties can be found in other algorithms. Generally, the proposed framework can converge within as few as 3–5 iterations.

V. Conclusion

In this paper, we develop four linear dimensionality reduction methods, i.e., RIPCA, RILDA, RILPP, and RIMFA. Then a novel dimensionality reduction framework using the RI L_1 norm or $L_{2,1}$ norm is proposed. Comprehensive analyses, including theoretical analyses on the convergence, computational complexity and the essence of the framework, are presented. Particularly, when all the projections are computed and used, the proposed framework has global optimal solution. The proposed framework inherits the simplicity of the previous Frobenius norm based graph embedding learning methods. We show that the RI dimensionality reduction framework can be degraded to the graph embedding algorithm framework by defining a special weight for each pair of data points. The proposed framework is easy to solve since it just needs to iteratively solve the standard eigenequation. Experiments on five image databases show that the RI dimensionality reduction framework (or the four representative methods proposed in this paper, i.e., RIPCA, RILDA, RILPP, and RIMFA) can perform better than the previous graph embedding algorithm framework (or the corresponding four methods, i.e., PCA, LDA, LPP, and MFA) in most cases.

References

[1] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. America A Opt. Image Sci. Vis.*, vol. 4, no. 3, pp. 519–524, 1987.

[2] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

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

[4] 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.

[5] W.-H. Yang and D.-Q. Dai, "Two-dimensional maximum margin feature extraction for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 4, pp. 1002–1012, Aug. 2009.

[6] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA: A novel fast feature extraction technique for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 946–952, Aug. 2006.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[10] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, 2004.

[11] 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.

[12] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal Laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.

[13] S. Yan *et al.*, "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.

[14] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2811–2821, Nov. 2007.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[19] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of L_1-optimizer in pattern classification," *Pattern Recognit.*, vol. 45, no. 3, pp. 1104–1118, Mar. 2012.

[20] Z. Lai, W. K. Wong, Y. Xu, C. Zhao, and M. Sun, "Sparse alignment for robust tensor learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1779–1792, Oct. 2014.

[21] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

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

[23] S.-J. Wang *et al.*, "Sparse tensor discriminant color space for face verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 876–888, Jun. 2012.

[24] Z. Lai, W. K. Wong, Z. Jin, J. Yang, and Y. Xu, "Sparse approximation to the eigensubspace for discrimination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1948–1960, Dec. 2012.

[25] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[26] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 723–735, Apr. 2016.

[27] Z. Lai, Y. Xu, J. Yang, J. Tang, and D. Zhang, "Sparse tensor discriminant analysis," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3904–3915, Oct. 2013.

[28] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[29] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with L1–graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[30] F. Zang and J.-S. Zhang, "Label propagation through sparse neighborhood and its applications," *Neurocomputing*, vol. 97, pp. 267–277, Nov. 2012.

[31] Q. Ke and T. Kanade, "Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 739–746.

[32] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[33] Y. Pang and Y. Yuan, "Outlier-resisting graph embedding," *Neurocomputing*, vol. 73, nos. 4–6, pp. 968–974, Jan. 2010.

[34] W. Zheng, Z. Lin, and H. Wang, "L1-norm kernel discriminant analysis via bayes error bound optimization for robust feature extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 793–805, Apr. 2014.

[35] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2004, pp. 281–288.

[36] H. Huang and C. Ding, "Robust tensor factorization using R1 norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.

[37] W. K. Wong, Z. Lai, Y. Xu, J. Wen, and C. P. Ho, "Joint tensor feature analysis for visual object recognition," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2425–2436, Nov. 2015.

[38] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 1813–1821.

[39] C.-X. Ren, D.-Q. Dai, and H. Yan, "Robust classification using $\ell_{2,1}$-norm based regression model," *Pattern Recognit.*, vol. 45, no. 7, pp. 2708–2718, Jul. 2012.
[40] H. Kong, Z. Lai, X. Wang, and F. Liu, "Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning," *Neurocomputing*, vol. 177, pp. 198–205, Feb. 2016.
[41] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 55. Barcelona, Spain, 2011, pp. 1294–1299.
[42] 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.
[43] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, 2011, pp. 1324–1329.
[44] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.
[45] Z. Ma *et al.*, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.
[46] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic, 1990.
[47] Y.-M. Zhang, K. Huang, X. Hou, and C.-L. Liu, "Learning locality preserving graph from data," *IEEE Trans. Cybern.*, vol. 44, no. 11, pp. 2088–2098, Nov. 2014.
[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
[49] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[50] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, Guangzhou, China, in 2002, the M.S. degree from Jinan University, Guangzhou, in 2007, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2011.

He has been a Research Associate, a Post-Doctoral Fellow, and a Research Fellow with Hong Kong Polytechnic University, Hong Kong. He has published over 50 scientific articles, including 20 papers published on top-tier IEEE TRANSACTIONS. His current research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization, and applications in the fields of intelligent robot research.

Dr. Lai is currently an Associate Editor of the *International Journal of Machine Learning and Cybernetics*.

Yong Xu (M'06) received the B.S. and M.S. degrees from the Air Force Institute of Meteorology, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2005.

From 2005 to 2007, he was a Post-Doctoral Research Fellow with the Shenzhen Graduate School, Harbin Institute of Technology (HIT), Harbin, China. He is currently a Professor with the Shenzhen Graduate School, HIT. He also acts as a Research Assistant Researcher with Hong Kong Polytechnic University, Hong Kong, from 2007 to 2008. He has published over 40 scientific papers. His current research interests include pattern recognition, biometrics, and machine learning.

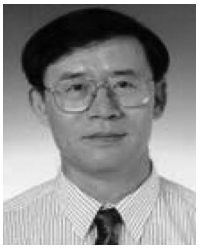
Jian Yang received the B.S. degree in mathematics from Xuzhou Normal University, Xuzhou, China, in 1995, the M.S. degree in applied mathematics from Changsha Railway University, Changsha, China, in 1998, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002.

In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the School of Computer Science and Technology, NUST. He has authored over 80 scientific papers in pattern recognition and computer vision. He has over 2000 ISI Web of Science and 4000 Google Scholar citations. His current research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang was a recipient of the RyC Program Research Fellowship sponsored by the Spanish Ministry of Science and Technology, in 2003.

Linlin Shen received the B.Sc. degree from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 2005.

He was a Research Fellow with the Medical School, University of Nottingham, researching brain image processing of magnetic resonance imaging. He is currently a Professor and the Director of the Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His current research interests include Gabor wavelets, face/palmprint recognition, medical image processing, and hyperspectral image classification.



David Zhang (F'08) received the B.S. degree in computer science from Peking University, Beijing, China, the M.Sc. degree in computer science and the first Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, and then an Associate Professor with the Academia Sinica, Beijing. He is currently a Head of the Department of Computing and the Chair Professor with Hong Kong Polytechnic University, Hong Kong. He also serves as the Visiting Chair Professor with Tsinghua University, and an Adjunct Professor in Peking University, Shanghai Jiao Tong University, Shanghai, China, HIT, and the University of Waterloo. He has authored over ten books and 200 journal papers.

Prof. Zhang is the Founder and an Editor-in-Chief of the *International Journal of Image and Graphics*, a Book Editor of *Springer International Series on Biometrics*, an Organizer of the International Conference on Biometrics Authentication, and an Associate Editor of over ten international journals including the IEEE TRANSACTIONS and *Pattern Recognition.* He is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a fellow of International Association for Pattern Recognition.