Neurocomputing 216 (2016) 672–683

Contents lists available at ScienceDirect

Neurocomputing

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

Recursively global and local discriminant analysis for semi-supervised and unsupervised dimension reduction with image analysis

Shangbing Gao [a,b,*], Jun Zhou [a], Yunyang Yan [a], Qiao Lin Ye [c]

[a] Faculty of Computer and Software Engineering, Jiangsu Provincial Key Laboratory for Advanced Manufacturing Technology, Huaiyin Institute of Technology, Huai'an, PR China
[b] Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology), Nanjing, PR China
[c] School of Information Technology, Nanjing Forestry University, Nanjing, PR China

ARTICLE INFO

Article history:
Received 6 April 2016
Received in revised form
22 July 2016
Accepted 8 August 2016
Communicated by: Dacheng Tao
Available online 20 August 2016

Keywords:
Semisupervised dimension reduction
"Concave-convex" programming problem
Feature extraction
Fisher linear discriminant analysis
Manifold regularization

ABSTRACT

Semi-supervised discriminant analysis (SDA) is a recently-developed semi-supervised dimension reduction method for improving the performance of Fisher linear discriminant analysis (FDA), which attempts to mine the local structures of both labeled and unlabeled data. In this paper, we develop new semi-supervised and unsupervised discriminant analysis techniques. Our semi-supervised method, referred as to as recursively global and local discriminant analysis (RGLDA), is modeled based on the characterizations of "locality" and "non-locality", such that the manifold regularization in the formulation has a more direct connection to classification. The objective of RGLDA is a "concave-convex" programming problem based on the hinge loss. Its solution follows from solving multiple related SVM-type problems. In addition, we also propose a simple version (called URGLDA) for unsupervised dimension reduction. The experiments tried out on several image databases show the effectiveness of RGLDA and URGLDA.

1. Introduction

In many real applications, such as face recognition, we are usually faced with high-dimensional data. The data may be represented by original images or many kinds of visual features [45]. In such cases, exacting good features is crucial for mitigating the so-called "curse of dimensionality" and improving the performance of any pattern classifier. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) [1,18] and Fisher Linear Discriminant Analysis (FDA) [2,14,35], are developed for this purpose. PCA and FDA have been widely applied in the fields of pattern recognition and computer vision. Many experimental studies have shown that FDA outperforms PCA significantly [3,4,36]. PCA can also be used for the data reconstruction. Same as PCA, the k-dimensional coding schemes [41,42] attempt to represent data using a set of representative k-dimensional vectors. Bian et al. [44] presented analysis on the bound of FDA.

Recently, there is much interest in developing manifold learning algorithms, e.g., Isometric Feature Mapping (ISOMAP) [5], Local Linear Embedding (LLE) [6], Laplacian Eigenmap (LE) [7], and Locality Preserving Projections (LPP) [8]. A main problem of ISOMAP, LLE and LE is that they y cannot map the unknown points [8]. LPP was proposed to deal with this problem [37,38]. Such an algorithm, however, have no direct connection to classification [9], which only characterizes local scatter and ignores the characterization of "non-locality". Recently, Yang et al. [9] proposed Unsupervised Discriminant Projection (UDP) to resolve this issue. UDP is modeled based on "locality" and "non-locality" [9]. In these unsupervised methods, the prior class information is not used.

Supervised learning algorithms may not generalize well enough when there is no sufficiently supervised information, although they generally outperform unsupervised learning algorithms [39,40]. Furthermore, collecting labeled data is generally more involved than collecting unlabeled clearly [10], since it requires expensive human labor, and a large amount of noise and outlier data are easy to be introduced. In [43], Liu et al. have attempted to use importance reweighting to solve the classification problems where the labeled samples are corrupted. In order to sufficiently exploit unlabeled data for better classification, many semi-supervised learning algorithms, such as Transductive SVM (TSVM) [11] and graph-based semi-supervised learning algorithms [10,12], have been recently developed. Among them, Manifold Regularization (MR) [10] is the most attractive [10]. However, these algorithms are developed for classification problems. Based on distance learning [46], Yu et al. [47] proposed semisupervised multiview distance metric learning. In dimension reduction, Cai et al., [13] proposed a

* Corresponding author.
E-mail address: luxiaofen_2002@126.com (S. Gao).

http://dx.doi.org/10.1016/j.neucom.2016.08.018
0925-2312/

novel method, called semi-supervised discriminant analysis (SDA). It not only preserves the discriminating information of labeled data but also the local structure of both labeled and unlabeled data. Specifically, both labeled and unlabeled data are used to build a graph Laplacian [13] that is incorporated into the FDA to smooth the mapping. SDA inherits the respective superiorities of FDA and manifold learning. However, SDA focuses only on the local geometry, such that it has no direct connection to classification [39]. A basic assumption behind SDA is that nearby points will have similar embeddings [13], which is consistent with the manifold assumption [10]. In practice, if the manifold assumption holds, the points from different classes could lie in different manifolds, which means that if we would like to obtain a better projection than that yielded by SDA, then, like UDP [9], the "non-locality" should be also taken into account in the framework of SDA, such that as much discriminant information of labeled and unlabeled points as possible can be mined. Based on SDA, Huang et al. [32] proposed TR-FLDA. By employing the similar idea in TR-FLDA, Dornaika et al. [33] proposed a new graph-based semisupervised DR (GSDR) method. However, these two methods suffer from the same problems of SDA. For existing eigenvalue-based dimension reduction techniques, it is difficult to introduce the sparsity in the projection matrix [15]. The last few years has seen few elegant sparse dimension reduction techniques in multi-class setting, such as Sparse PCA (SPCA) [15], which uses L_1-penalized regression on regular projection axes. With the similar technique, we can also develop a multi-class sparse SDA method. However, they are two-stage approaches. Intuitively, it is too naive to believe that there is no any information loss in the two-stage processing. We believe that sparsity should be introduced into the model by entirely following the genuine geometrical interpretation of SDA for the guarantee of obtaining a good performance.

In this paper, we propose a new semi-supervised dimension reduction method, called Recursively Global and Local Discriminant Analysis (RGLDA) which includes two steps. First, a novel dimension reduction framework for semi-supervised learning, called Global and Local Discriminant Analysis (GLDA), is proposed. In addition to incorporating the basic idea behind SDA of finding an optimal projection and estimating the local geometry of a relatively limited amount of labeled data as well as considerable unlabeled data, GLDA simultaneously mines the underlying non-local structure of labeled and unlabeled data. Second, the projections are generated by using the similar recursive procedure as suggested in RFDA [19]. The new algorithm is essentially developed from the SDA but has a significant performance advantage. Different from SDA which maximizes the squared sum of all the pairwise distances between class means, our method maximizes every pairwise distance between class means and allows some pairwise distances commit the maximization limit. Moreover, the prior work [16,17,32] on various recognition tasks has demonstrated that casting an eigenvalue-based problem as a related SVM-type problem can lead to better recognition rates. In addition, we extend RGLDA to a simple version (called RGLDA) for unsupervised dimension reduction. The main contributions of this study are summarized as follows:

1) We propose a framework for semi-supervised and unsupervised dimension reduction, whose unique idea can be used in most of the up-to-date semisupervised dimension reduction methods, such as TR-FLDA [33], and GSDR [34].
2) Our work characterizes not only the local but also the non-local quantities, such that the mapping has a direct connection to classification.
3) Our proposed framework is flexible. To be specific, the special formulation allows us to easily develop the sparse semi-supervised model by incorporating various regularization techniques into the formulation in future work. Unlike the recently proposed two-stage sparse methods, the sparsity is directly introduced to the model.

We organize this paper as follows. Section II gives a brief review of FDA and SDA. In section III, we give a basic framework for dimension reduction that casts an eigenvalue-based problem as an SVM-type problem, proposes GLDA, and shows its solution. Section IV develops recursive GLDA (RGLDA), which aims at generating multiple projection axes. In section V, a simplified dimension reduction method for unsupervised learning is proposed. In section VI, we evaluate our algorithms on several image databases. Section VII gives a method of constructing sparse dimension reduction methods based on the formulation of RGLDA. In the last section, we draw some conclusions.

2. Review of FDA and SDA

Let $\mathbf{x} = \{\mathbf{x}_i|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \mathbf{x}_{l+2} \ldots \mathbf{x}_m\} \in \mathbb{R}^{n\times m}$ be the data matrix, where m and n are the number and dimensionality of data, and $\mathbf{x}_i|_{i=1}^{l}$ and $\mathbf{x}_i|_{i=l+1}^{m}$ are labeled and unlabeled data, respectively. The labeled data is from c different classes. There are N_u samples in each class, which are represented by $\boldsymbol{D}_u$, $u = 1, 2, \ldots, c$. Define by $\mathbf{z} = \in \mathbb{R}^d (1 \le d \le n)$ a low-dimensional representation of a high-dimensional sample $\mathbf{x}$ in the original space, where d is the dimensionality of the reduced space. The purpose of dimension reduction is to seek for a transformation matrix $\mathbf{W}$, such that a lower representation $\mathbf{z}$ of the sample $\mathbf{x}$ can be calculated as $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, where "T" denotes the transpose. Other notations are listed in Table 1, each being explained when it is first used.

2.1. Fisher linear discriminant analysis (FDA)

FDA, as a popular supervised dimension reduction method, aims at finding the projection that simultaneously maximizes the between-class scatter and minimize the within-class scatter. That is, we can obtain the projection by maximizing the following objection function

$$J_F(\mathbf{W}) = \frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}, \tag{1}$$

where the within-class scatter matrix $\mathbf{S}_W$ is defined by

$$\mathbf{S}_\mathbf{W} = \sum_{u=1}^{c}\sum_{x\in\boldsymbol{D}_u}(\mathbf{x}-\boldsymbol{\mu}^{(u)})(\mathbf{x}-\boldsymbol{\mu}^{(u)})^T = \mathbf{X}^T\mathbf{L}_F\mathbf{X} \tag{2}$$

Table 1
Notations.

$\mathbf{S}_w \in \mathbf{R}^{n\times n}$	The within-class scatter;
$\mathbf{S}_B \in \mathbf{R}^{n\times n}$	The between-class scatter;
$\mathbf{S}_t \in \mathbf{R}^{n\times n}$	The global scatter;
$\boldsymbol{\mu}^{(u)} \in \mathbf{R}^n$	The mean vector of the samples in class u;
$\boldsymbol{\mu} \in \mathbf{R}^n$	The Total mean of the samples;
$\mathbf{H} \in \mathbf{R}^{m\times m}$	The adjacency matrix;
ρ	A regularization parameter;
ν	A regularization parameter;
$\mathbf{D} \in \mathbf{R}^{m\times m}$	A diagonal matrix whose entries on diagonal are column or row sum of $\mathbf{H}$;
$\mathbf{L} = \mathbf{D} - \mathbf{H} \in \mathbf{R}^{m\times m}$	The Laplacian matrix;
$\mathbf{S}_L = \mathbf{X}\mathbf{L}\mathbf{X}^T \in \mathbf{R}^{n\times n}$	A graph scatter matrix;
f_i and g_i	Two real-values convex functions on a vector space $\mathbf{Z}$;
$T_1\{f, \mathbf{z}\}(\mathbf{z}')$	The first order Taylor expansion of f at location $\mathbf{z}$;
$\partial_\mathbf{z} f(\mathbf{z})$	The gradient of the function f at location $\mathbf{z}$;
$\xi \in \mathbf{R}^m$	The Hinge loss function;
$\mathbf{X}^{(u)}$	The sample matrix of class u;
$\mathbf{e}^{(u)}$	A column vector of ones of size($\mathbf{X}^{(u)}$) dimensions.

in which $\mathbf{L}_F = \mathbf{I} - \mathbf{F}$ is a graph Laplacian matrix, $\mathbf{F}$ is a adjacency matrix whose elements are $1/N_u$ if any pair of points belong to the same class and are 0's otherwise, $\boldsymbol{\mu}^{(u)}$ is the mean of the labeled samples of class u, and the between-class scatter $\mathbf{S}_B$ is defined by

$$\mathbf{S}_B = \sum_{u=1}^{c} N_u(\boldsymbol{\mu}^{(u)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(u)} - \boldsymbol{\mu})^{\mathrm{T}}, \tag{3}$$

where $\boldsymbol{\mu}$ is the average vector of the labeled point. The solutions of FDA are the generalized eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d$ ($\mathbf{w}_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, d$) of $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$ corresponding to the first d largest eigenvalues $\lambda_1 \geq \lambda_2, \ldots, \lambda_d$.

2.2. Semi-supervised Discriminant Analysis (SDA) [13]

SDA defines an adjacency matrix $\mathbf{H}$, whose elements are gives as follows:

$$H_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \text{ is among } k \text{ nearest neighbors of } \mathbf{x}_i \\ & \text{or } \mathbf{x}_i \text{ is among } k \text{ nearest neighbors of } \mathbf{x}_j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Obviously, $\mathbf{H}$ is a symmetric matrix. SDA incorporates a manifold regularization term into the FDA framework to try to find the transformation that draws the close samples closer together. Then, the desired transformation should be one that minimizes both the within-class scatter of labeled samples and the local scatter of data and simultaneously maximizes the between-class scatter of labeled samples. As it happens, the transformation can be obtained by maximizing the following criterion:

$$J_S(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w} + \delta\mathbf{w}^{\mathrm{T}}\mathbf{S}_L\mathbf{w}}, \tag{5}$$

where δ is a scaling factor, and the local scatter matrix is defined as

$$\begin{aligned} \mathbf{w}^{\mathrm{T}}\mathbf{S}_L\mathbf{w} &= \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_j\right)^2 \\ &= \left(\mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}}\mathbf{w}\right) \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{w} \end{aligned} \tag{6}$$

in which $\mathbf{L} = \mathbf{D} - \mathbf{H}$ is a Laplacian matrix, and $\mathbf{D}$ is a diagonal matrix whose entries on diagonal are column or row sum of $\mathbf{H}$. It is clear that one basic assumption behind SDA is manifold assumption, i.e., any pair of neighboring points has the same labels. A set of projection axes of SDA can be obtained by selecting the generalized eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d$ of $\mathbf{S}_N\mathbf{w} = \lambda\mathbf{S}_L\mathbf{w}$ corresponding to the first d largest eigenvalues $\lambda_1 \geq \lambda_2, \ldots, \lambda_d$.

2.3. Constrained "Concave-Convex" Procedure (CCP)

The constrained "concave-convex" procedure (CCP) [30] is designed for solving the optimization problems with a concave-convex function and concave-convex constraints. The CCP aims at solving the following optimization problem [22,23,31]

$$\begin{aligned} &\min_{\mathbf{z}} \ f_0(\mathbf{z}) - g_0(\mathbf{z}) \\ &\text{s.t.} \ \ f_i(\mathbf{z}) - g_i(\mathbf{z}) \leq \pi_i, \ i = 1, 2, \ldots, l, \end{aligned}$$

where f_i and g_i are two real-values convex functions on a vector space $\mathbf{Z}$ for all $i = 1, 2, \ldots, n$ and $\pi_i \in \mathbb{R}$ for all $i = 1, 2, \ldots, n$. Denote by $T_1\{f, \mathbf{z}\}(\mathbf{z}')$ the first order Taylor expansion of f at location $\mathbf{z}$, that is $T_1\{f, \mathbf{z}\}(\mathbf{z}') = f(\mathbf{z}) + \partial_{\mathbf{z}} f(\mathbf{z})(\mathbf{z}' - \mathbf{z})$, where $\partial_{\mathbf{z}} f(\mathbf{z})$ is the gradient of the function f at $\mathbf{z}$. For non-smooth functions, $\partial_{\mathbf{z}} f(\mathbf{z})$ can be replaced by the subgradient [22,23]. A subgradient of f at $\mathbf{z}$ is any vector $\mathbf{g}$ that satisfies the inequality $f(\mathbf{y}) \leq f(\mathbf{x}) + \mathbf{g}'(\mathbf{y} - \mathbf{z})$ for all $\mathbf{y}$. The subdifferential of f at $\mathbf{z}$ is the set of all subgradients of f at $\mathbf{z}$ [22]. Initialize $\mathbf{z}_0$ with a random value or a best guess. The CCP calculates $\mathbf{z}_{t+1}$ from $\mathbf{z}_t$ by replacing $g_i(\mathbf{z})$ with $T_1\{g_i, \mathbf{z}_t\}(\mathbf{z})$, and then sets $\mathbf{z}_{t+1}$ to the solution of the following convex optimization problem

$$\begin{aligned} &\min_{\mathbf{z}} \ f_0(\mathbf{z}) - T_1\left\{g_0, \mathbf{z}_t\right\}(\mathbf{z}) \\ &\text{s.t.} \ \ f_i(\mathbf{z}) - T_1\left\{g_i, \mathbf{z}_t\right\}(\mathbf{z}) \leq c_i, \ i = 1, 2, \ldots, l. \end{aligned}$$

The above recursive procedure continues until $\mathbf{z}_t$ converges. Smolaet al. [22] has proved the fast convergence of CCP that is guaranteed to find a local minimum.

3. Global and Local Discriminant Analysis (GLDA) for semi-supervised learning

First, we give a general formulation of an eigenvalue based method for classification problems, and then disclose that many classical dimension reduction techniques can be also reformulated as related SVM-type problems.

3.1. Casting an eigenvalue-based technique as a related SVM-type problem

The idea of reformulating an eigenvalue-based technique as a related SVM-type problem derives from the up-to-date multiplane classification method, called Twin Support Vector Machine (TWSVM) [16], which casts Proximal Support Vector Machines via Generalized (GEPSVM) [20] as a related SVM-type problem. Recently, similar idea is also used in our works [17,21] for classification. In general, the objective function of an eigenvalue-based optimization problem can be written as

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}^{\mathrm{T}}\mathbf{G}\|^2}{\|\mathbf{w}^{\mathrm{T}}\mathbf{H}\|^2}, \tag{7}$$

where $\mathbf{G}$ and $\mathbf{H}$ represent some distances. In the multiplane classifier GEPSVM, $\mathbf{G}$ calculates the distances of the plane to each of points for its own class and $\mathbf{H}$ for the other classes. In our pervious work [21] where MVSVM is proposed, $\mathbf{G}$ computes the distances of each of projected points for one particular class to their mean ($\mathbf{m}$), while $\mathbf{H}$ calculates the distance between $\mathbf{m}$ and the mean of projected points for the other classes. According to [16,17], we can cast (7) as the following problem

$$\begin{aligned} &\min_{\mathbf{w},\xi} \ \mathbf{w}^{\mathrm{T}}\mathbf{G}\mathbf{G}^{\mathrm{T}}\mathbf{w} + \mathrm{v}\sum_{i=1}^{l}\xi_i \\ &\text{s.t.} \ \ |\mathbf{w}^{\mathrm{T}}\mathbf{H}_i| + \xi_i \geq 1, \ \xi_i \geq 0. \end{aligned} \tag{8}$$

where $\mathbf{H}_i$ denotes the ith column vector of $\mathbf{H}$, $|\cdot|$ the absolute, v a constant, and ξ_i a loss function which relaxes the hard constrains, i.e., minimizing the variance of the error committed when the hard constraints violate. Note that $|\cdot|$ is convex. Instead of maximizing the sum of the squares of the distances in $\mathbf{H}$ in (7), the constrains in (8) maximize each of distances in $\mathbf{H}$. We can arbitrarily select the non-negative constant in the right hand side of the constraint term in (8), and changing it to any other positive constant δ results in $\mathbf{w}$ being replaced by $\delta\mathbf{w}$. In [16,17], the analysis and results have shown that the reformulation of an eigenvalue-based technique provides performance improvement. In addition, it is very easy to introduce model sparsity by directly incorporating various regularization techniques.

For binary classification problems, we can easily remove the $|\cdot|$ in (8) and then obtain the simple models as in [16,17]. For dimension reduction in multi-class setting, the absolute function in (8) will be remained, such that the first constraint in (8) is non-

linear. Letting $\mathbf{G} = \left[\mathbf{X}^{(1)} - \boldsymbol{\mu}^{(1)}\mathbf{e}^{(1)},\ \mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)}\mathbf{e}^{(2)}, \ldots, \mathbf{X}^{(c)} - \boldsymbol{\mu}^{(c)}\mathbf{e}^{(c)}\right]$, and $\mathbf{H} = \left[\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}, \boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}, \ldots, \boldsymbol{\mu}^{(c)} - \boldsymbol{\mu}\right]$, where $\mathbf{X}^{(u)}$ is the sample matrix of class u, and $\mathbf{e}^{(u)}$ is a column vector of ones of size($\mathbf{X}^{(u)}$) dimensions, we can obtain a SVM-type FDA [31] model.

Clearly, the minimization (8) satisfies the condition of the CCP: simply define

$$f_0(\mathbf{w}, \xi) = \mathbf{w}^{\mathrm{T}}\mathbf{G}\mathbf{G}^{\mathrm{T}}\mathbf{w} + \mathrm{v}\sum_{i=1}^{l}\xi_i, f_i(\mathbf{w}, \xi) = 1 - \xi_i, \tag{9}$$

$$g_0(\mathbf{w}, \xi) = 0, \quad \text{and} \quad g_i(\mathbf{w}, \xi) = |\mathbf{w}^{\mathrm{T}}\mathbf{H}_i|. \tag{10}$$

We also set π_i equal to zero for all i. By using the CCP, it is easy to solve the optimization (8).

3.2. The GLDA method

The key to semi-supervised learning algorithm is the prior assumption of consistency which is generally called manifold assumption. For dimension reduction, it can be interpreted as that nearby points will have the similar embedding [13]. In the objective of SDA, the manifold regularizer considers the local geometry of samples, which has no direct connection to classification. Therefore, we can characterize the non-local scatter as in [9] to address this problem:

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\mathbf{S}_N\mathbf{w} &= \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(1 - H_{ij}\right)\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_j\right)^2 \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{S}_T\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_L\mathbf{w},\end{aligned} \tag{11}$$

where $\mathbf{S}_T$ is the global scatter of both the labeled and unlabeled points. We construct the following semi-supervised problem to be maximized

$$\mathbf{J}_S(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w} + \rho\mathbf{w}^{\mathrm{T}}\mathbf{S}_N\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w} + \rho\mathbf{w}^{\mathrm{T}}\mathbf{S}_L\mathbf{w}}, \tag{12}$$

where ρ is a regularization parameter.

Theorem 1. *The optimization problem (12) can be replaced by the following problem to be minimized.*

$$J_G(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w} + \rho\mathbf{w}^{\mathrm{T}}\mathbf{S}_L\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_T^{'}\mathbf{w} + \rho\mathbf{w}^{\mathrm{T}}\mathbf{S}_T\mathbf{w}} \tag{13}$$

in the course of seeking for the discriminant vectors of the optimal set, where $\mathbf{S}_{\mathbf{T}}^{'}$ is the global scatter of labeled samples.

Proof. Obviously, the projection axes of the optimization problem (12) to be maximized are eigenvectors corresponding to the largest eigenvalues of $(\mathbf{S}_B + \rho\mathbf{S}_N)\mathbf{w} = \lambda(\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w}$. As seen from (11), $\mathbf{S}_N = \mathbf{S}_T - \mathbf{S}_L$, thus we have.

$$\begin{aligned}&(\mathbf{S}_B + \rho\mathbf{S}_N)\mathbf{w} = \lambda(\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w} \\ &\Rightarrow \left[(\mathbf{S}_T^{'} - \mathbf{S}_W) + \rho(\mathbf{S}_T - \mathbf{S}_L)\right]\mathbf{w} = \lambda(\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w} \\ &\Rightarrow (\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w} = \frac{1}{1+\lambda}(\mathbf{S}_T^{'} + \rho\mathbf{S}_T) \\ &\Rightarrow (\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w} = \lambda'(\mathbf{S}_T^{'} + \rho\mathbf{S}_T)\end{aligned} \tag{14}$$

where $\mathbf{S}_T^{'}$ is the global scatter matrix of labeled samples, and $\lambda' = 1/(1+\lambda)$. A set of projection axes of minimization (13) can be obtained by selecting the generalized eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d$ ($\mathbf{w}_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, d$) of $(\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w} = \lambda(\mathbf{S}_T^{'} + \rho\mathbf{S}_T)\mathbf{w}$ corresponding to the first d smallest eigenvalues $\lambda_1 \geq \lambda_2, \ldots, \lambda_d$. Therefore, the eigenvector of $(\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w} = \lambda'(\mathbf{S}_T^{'} + \rho\mathbf{S}_T)\mathbf{w}$ associated with smallest eigenvalue λ' is equivalent to the eigenvector of $(\mathbf{S}_B + \rho\mathbf{S}_N)\mathbf{w} = \lambda(\mathbf{S}_W + \rho\mathbf{S}_L)\mathbf{w}$ associated with largest eigenvalue λ. □

From (13), we can see that GLDA considers both the global scatter and the local scatter. Obviously, we can yield FDA if setting ρ as zero. We now cast the problem (13) as a related SVM-type problem. The denominator $\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w} + \rho\mathbf{w}^{T}\mathbf{S}_L\mathbf{w}$ in (13) is equivalent to

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\mathbf{S}_T^{'}\mathbf{w} + \rho\mathbf{w}^{\mathrm{T}}\mathbf{S}_T\mathbf{w} &= \sum_{i=1}^{l}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}\right)(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}} \\ &\quad + \rho\sum_{i=1}^{M}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \mathbf{w}^{\mathrm{T}}\mathbf{m}\right)\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \mathbf{w}^{\mathrm{T}}\mathbf{m}\right)^{\mathrm{T}} \\ &= \mathbf{w}^{\mathrm{T}}[(\mathbf{x}_1 - \boldsymbol{\mu}), \ldots, (\mathbf{x}_l - \boldsymbol{\mu}), \rho(\mathbf{x}_1 - \mathbf{m}), \ldots, \rho(\mathbf{x}_M - \mathbf{m})] \\ &\quad \cdot[(\mathbf{x}_1 - \boldsymbol{\mu}), \ldots, (\mathbf{x}_l - \boldsymbol{\mu}), \rho(\mathbf{x}_1 - \mathbf{m}), \ldots, \rho(\mathbf{x}_M - \mathbf{m})]^{\mathrm{T}}\mathbf{w}\end{aligned} \tag{15}$$

where $\mathbf{m}$ is the total mean vector. Using (15), we rewrite the formulation (13) as

$$J_G(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w} + \rho\mathbf{w}^{T}\mathbf{S}_L\mathbf{w}}{\|\mathbf{w}^{\mathrm{T}}\mathbf{H}\|^2}, \tag{16}$$

which has a similar formulation as that in (7), where $\mathbf{H} = \left[(\mathbf{x}_1 - \boldsymbol{\mu}), \ldots, (\mathbf{x}_l - \boldsymbol{\mu}), \rho(\mathbf{x}_1 - \mathbf{m}), , \ldots, \rho(\mathbf{x}_M - \mathbf{m})\right]$. We can first cast (16) as the formulation as in (9) and then obtain the following minimization problem in multi-class setting

$$\begin{aligned}&\min_{\mathbf{w},\xi}\ \mathbf{w}^{\mathrm{T}}\mathbf{S}_D\mathbf{w} + \mathrm{v}\sum_{i}^{l+M}\xi_i \\ &\text{s.t.}\quad \left|\mathbf{w}^{\mathrm{T}}\mathbf{H}_i\right| + \xi_i \geq 1, \quad \xi_i > 0,\end{aligned} \tag{17}$$

where $\mathbf{S}_D = \mathbf{S}_W + \rho\mathbf{S}_L$, and $\mathbf{H}_i$ denotes the ith column vector of $\mathbf{H}$. The optimization problem (17) can be solved with the CCP introduced in the above section. Letting $\mathbf{F} = \left[\mathrm{sign}\left(\mathbf{w}_t^{\mathrm{T}}\mathbf{H}_1\right)\left(\mathbf{w}^{\mathrm{T}}\mathbf{H}_1\right), \ldots, \mathrm{sign}\left(\mathbf{w}_t^{\mathrm{T}}\mathbf{H}_{l+M}\right)\left(\mathbf{w}^{\mathrm{T}}\mathbf{H}_{l+M}\right)\right]^{\mathrm{T}}$, we need to solve the following optimization problem for each update

$$\begin{aligned}&\min_{\mathbf{w},\xi}\ \mathbf{w}^{\mathrm{T}}\mathbf{S}_D\mathbf{w} + \nu\mathbf{e}^{\mathrm{T}}\xi \\ &\text{s.t.}\quad \mathbf{F}\mathbf{w} + \xi \geq \mathbf{e}, \xi \geq 0.\end{aligned} \tag{18}$$

According to the CCP, the solution $\mathbf{w}_t$ obtained from the minimization (17) is then replaced with $\mathbf{w}_{t+1}$. It is easy to check that the optimization (18) can be emulated by a regularized SVM without threshold when a new training set, consisting of $l + M$ training samples $(\mathbf{k}_1, y_1), \ldots, (\mathbf{k}_{l+M}, y_{l+M})$, is constructed, where $\mathbf{k}_i = \mathbf{H}_i, y_i \in \{+1, -1\}$, $i = 1, 2, \ldots, l + M$ are the class labels estimated by computing sign $\left(\mathbf{w}_t^{\mathrm{T}}\mathbf{H}_i\right)$. This shows each update of GLDA has a direct connection to SVM. According to the CCP, we solve the GLDA optimization problem using the following algorithm. We set the stopping criterion in the CCP as the difference of two iterations less than 0.01. GLDA obtains the solution by solving the problems as defined in (17) iteratively. The aforementioned computation aids in the generation of only one projection vector. In the following, we show how to generate multi-projection vectors by a recursive procedure.

4. Recursive GLDA (RGLDA)

In this section, we first detail our RGLDA method, which calculates the features by a recursive procedure.

4.1. Algorithm

A to find multiple projection axes. We utilize of the similar idea as suggested in [19,31] to generate the new sample set, which consists of

the following steps: (1) determines the optimal projection direction $\mathbf{w}$ according to (11), and (2) generate new data samples by projecting the original samples into a subspace. Specifically, the RGLDA algorithm is described as follows.

Algorithm 1. Recursive GLDA (RGLDA).

(i) Initialization. Let the number of iterations $p = 1$, and the original training sample set $\mathbf{X}$ defined in (1).

(ii) Determine the first RGLDA feature by $\tilde{\mathbf{w}}_1 = (\mathbf{S}_D)^{-1}\mathbf{F}^{\mathrm{T}}\boldsymbol{\alpha}$, by solving the optimization problem (18) on set $\mathbf{X}$.

(iii) Normalize $\mathbf{w}_p$ and then generate the following two sample data sets, i.e., labeled sample set $\mathrm{lab}(\mathbf{X}_{p+1})$ and training sample set $\mathbf{X}_{p+1}$ by projecting the samples into the subspace

$$\mathrm{lab}(\mathbf{X}_{p+1}) = \{(\mathbf{x}_i)_{p+1} | (\mathbf{x}_i)_{p+1} = (\mathbf{x}_i)_p - \mathbf{w}_p^{\mathrm{T}}(\mathbf{x}_i)_p\mathbf{w}_p,\ i = 1, 2, \ldots, l\},\ \mathbf{X}_{p+1} = \{(\mathbf{x}_i)_{p+1} | (\mathbf{x}_i)_{p+1} = (\mathbf{x}_i)_p - \mathbf{w}_p^{\mathrm{T}}(\mathbf{x}_i)_p\mathbf{w}_p,\ i = 1, 2, \ldots, M\}.$$

(iv) Calculate $(\mathbf{S}_D)_{p+1}$ associated with the new sets $\mathrm{lab}(\mathbf{X}_{p+1})$ and $\mathbf{X}_{p+1}$

(v) Seek for the $(p+1)$th optimal projection direction $\mathbf{w}_{p+1}$ by solving the optimization problem as defined in (11) on sets $\mathrm{lab}(\mathbf{X}_{p+1})$ and $\mathbf{X}_{p+1}$.

(vi) Terminate when reduce the expected dimension of reduced space.

(vii) Increment p by 1, i.e. $p \leftarrow p + 1$, and go back (iii).

The features available from RGLDA can be generated by the Algorithm 1. The recursive process may continue as long as matrix $(\mathbf{S}_D)_p$ and $\mathbf{H}_p$ are non-zero matrix. When $(\mathbf{S}_D)_p$ and $\mathbf{H}_p$ are zero matrix, the process naturally stops because the separability cannot be maximized by projection.

4.2. RGLDA in small sample size cases

In real application domain, such as face recognition, the number of samples in the training set (m) tends to be much smaller than that number of features in each sample (n), such the matrix $\mathbf{S}_D$ is singular. Recently, many dimension reduction algorithms, such as [2,8,9], are employed in the PCA subspace, that is, the samples are first projected to a PCA subspace. There are three considerations for this. First, the singularity occurring can be avoided [2]. Second, some useless information, such as noise, can be eliminated [24] by keeping most of energy. Third, the computational efficiency can be improved [9]. Therefore, we can first project the training sample set to a PCA subspace to avoid the singularity of $\mathbf{S}_D$, and then employ the RGLDA in the PCA subspace. Hereafter, we suppose that the matrix $\mathbf{S}_D$ is singular. Similar to [9,17,31], let $\mathbb{B} = \mathrm{span}\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q\}$ be the subspace and denote by $\mathbb{B}^{\perp} = \mathrm{span}\{\boldsymbol{\beta}_{q+1}, \ldots, \boldsymbol{\beta}_n\}$ its orthogonal complement, where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q$ are the first q eigenvectors of $\mathbf{S}_T$ corresponding to positive eigenvalues. Obviously, $\mathbb{B}^{\perp}$ is the null space of $\mathbf{S}_L$. Then, we get the following theorem.

Theorem 2. *Let $\mathbf{w} = \mathbf{u} + \boldsymbol{\theta}$ be a decomposition of $\mathbf{w}(\mathbf{w} \in \mathbb{R}^n)$ into a part $\mathbf{u} \in \mathbb{B}$ and a part $\boldsymbol{\theta} \in \mathbb{B}^{\perp}$, then the constrained optimization problem (17) which generates the first RGLDA feature vector is equivalent to.*

$$\begin{aligned} &\min_{\mathbf{w},\xi}\ \mathbf{u}^{\mathrm{T}}\mathbf{S}_D\mathbf{u} + \mathrm{v}\sum_{i}^{l+M}\xi_i \\ &\text{s.t.}\ \ |\mathbf{u}^{\mathrm{T}}\mathbf{H}_i| + \xi_i \geq 1,\ \ \xi_i > 0, \end{aligned} \tag{19}$$

Proof. Every $\boldsymbol{\theta}$ can be decomposed as a linear combination of the orthogonal eigenvectors of $\mathbf{S}_T$ that correspond to zero eigenvalues. Since$\boldsymbol{\theta} \in \mathbb{B}^{\perp}$, we have $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\theta} = 0$ and $\mathbf{S}_T\boldsymbol{\theta} = 0$. Since $\mathbf{S}_T = \mathbf{S}_L + \mathbf{S}_N$, one can gets $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_L\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\theta} = 0$. We further gets $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_L\boldsymbol{\theta} = 0$ and $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\theta} = 0$, which implies $\mathbf{S}_L\boldsymbol{\theta} = 0$ and $\mathbf{S}_N\boldsymbol{\theta} = 0$. Since.

$$\begin{aligned} \mathbf{S}_{\mathrm{T}} &= \sum_{i=1}^{M}\sum_{j=1}^{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \\ &= \sum_{i=1}^{l}\sum_{j=1}^{l}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} + \sum_{i=l+1}^{M}\sum_{j=1}^{l}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \\ &\quad + \sum_{i=1}^{l}\sum_{j=l+1}^{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} + \sum_{i=l+1}^{M}\sum_{j=l+1}^{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \\ &= \mathbf{S}'_{\mathrm{T}} + \mathbf{S}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{S} &= \sum_{i=1}^{l}\sum_{j=1}^{l}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} + \sum_{i=l+1}^{M}\sum_{j=1}^{l}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \\ &\quad + \sum_{i=1}^{l}\sum_{j=l+1}^{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} + \sum_{i=l+1}^{M}\sum_{j=l+1}^{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}. \end{aligned}$$

It is easy to check that $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}'_{\mathrm{T}}\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_W\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_B\boldsymbol{\theta} = 0$. One further gets $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_W\boldsymbol{\theta} = 0$, meaning that $\mathbf{S}_W\boldsymbol{\theta} = 0$. Therefore,

$$\begin{aligned} \mathbf{w}^{\mathrm{T}}\mathbf{S}_D\mathbf{w} &= (\mathbf{u} + \boldsymbol{\theta})^{\mathrm{T}}\mathbf{S}_D(\mathbf{u} + \boldsymbol{\theta}) \\ &= \mathbf{u}^{\mathrm{T}}\mathbf{S}_W\mathbf{u} + \rho\mathbf{u}^{\mathrm{T}}\mathbf{S}_L\mathbf{u} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_W\mathbf{u} + \rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_L\mathbf{u} \\ &\quad + \mathbf{u}^{\mathrm{T}}\mathbf{S}_W\boldsymbol{\theta} + \rho\mathbf{u}^{\mathrm{T}}\mathbf{S}_L\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_W\boldsymbol{\theta} + \rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_L\boldsymbol{\theta} \\ &= \mathbf{u}^{\mathrm{T}}\mathbf{S}_W\mathbf{u} + \rho\mathbf{u}^{\mathrm{T}}\mathbf{S}_L\mathbf{u} \\ &= \mathbf{u}^{\mathrm{T}}\mathbf{S}_D\mathbf{u} \end{aligned}$$

Based on the above formulas, we can get $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}'_T\boldsymbol{\theta} + \rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\theta} = 0$. We further have $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}'_T\boldsymbol{\theta} + \rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{T}}(\mathbf{R}_1 + \mathbf{R}_2 + \ldots + \mathbf{R}_{l+M})\boldsymbol{\theta} = 0$. It is easy to check that $\mathbf{R}_i = \mathbf{H}_i^{\mathrm{T}}\mathbf{H}_i$, are semipositive definite matrices, thus we have $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{R}_i\boldsymbol{\theta} = 0$, $i = 1,\ 2 \ldots, l + M$ and we can further get $\mathbf{R}_i\boldsymbol{\theta} = 0$. Hence,

$$\begin{aligned} \mathbf{w}^{\mathrm{T}}\mathbf{R}_i\mathbf{w} &= (\mathbf{u} + \boldsymbol{\theta})^{\mathrm{T}}\mathbf{R}_i(\mathbf{u} + \boldsymbol{\theta}) \\ &= \mathbf{u}^{\mathrm{T}}\mathbf{R}_i\mathbf{u} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{R}_i\mathbf{u} + \mathbf{u}^{\mathrm{T}}\mathbf{R}_i\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{R}_i\boldsymbol{\theta} \\ &= \mathbf{u}^{\mathrm{T}}\mathbf{R}_i\mathbf{u}. \end{aligned}$$

Since $|\mathbf{w}^{\mathrm{T}}\mathbf{H}_i| = \mathrm{sqrt}(\mathbf{w}^{\mathrm{T}}\mathbf{R}_i\mathbf{w})$, one can get $|\mathbf{w}^{\mathrm{T}}\mathbf{H}_i| = |\mathbf{u}^{\mathrm{T}}\mathbf{H}_i|$. Therefore, the projection axis $\mathbf{w}$ needing to be estimated can be replaced with $\mathbf{u}$. Thus, optimization problem (17) is equivalent to that in (19). □

Theorem 3 discloses the fact that the solution of (17) can be produced in the subspace $\mathbb{B}$ without any loss of the information. Let $\mathbf{P}$ denote a transformation matrix of q dimensions, each column vector of which is corresponding to a non-zero eigenvalue of $\mathbf{S}_T$. By linear algebra theory, $\mathbb{B}$ is isomorphic to the q-dimensional Euclidean space $\mathbb{R}^m$ [9,17]. The isomorphic mapping is exactly the transformation matrix $\mathbf{P}$ [9], one has

$$\mathbf{u} = \mathbf{P}\boldsymbol{\eta},\ \mathbf{u} \in \mathbb{B},\ \boldsymbol{\eta} \in \mathbb{R}^m, \tag{20}$$

where $\mathbf{P} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q)$. By the above mapping, the formulation (19) can be re-formulated as the following problem

$$\begin{aligned} &\min_{\boldsymbol{\eta},\xi}\ \frac{1}{2}\boldsymbol{\eta}^{\mathrm{T}}\tilde{\mathbf{S}}_D\boldsymbol{\eta} + \nu\sum_{i=1}^{l+M}\xi_i \\ &\text{s.t.}\ \ |\boldsymbol{\eta}^{\mathrm{T}}\mathbf{D}| + \xi_i \geq 1,\ \xi_i > 0, \end{aligned} \tag{21}$$

where $\tilde{\mathbf{S}}_D = \mathbf{P}^T\mathbf{S}_D\mathbf{P}$, and $\mathbf{D} = \mathbf{P}^T\mathbf{H}_i$. Also we know $\mathbf{S}_D = \mathbf{S}_W + \rho\mathbf{S}_L = \mathbf{X}(\mathbf{L}_F + \rho\mathbf{L})\mathbf{X}^T$. Thus, $\tilde{\mathbf{S}}_D = \tilde{\mathbf{X}}(\mathbf{L}_F + \rho\mathbf{L})\tilde{\mathbf{X}}^T$. Since $\mathbf{P} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q\right)$ and its all column vectors are principal eigenvectors of $\mathbf{S}_T$, $\tilde{\mathbf{X}} = \mathbf{P}^T\mathbf{X}$ is the PCA transformation of data matrix $\mathbf{X}$. Therefore, $\boldsymbol{\eta}$, as the solution of (21), can be generated in the PCA subspace. If $\boldsymbol{\eta}^*$ is the solution to (21), then, $\mathbf{u}^* = \mathbf{P}\boldsymbol{\eta}^*$ is the first RGLDA optimal projection. With the recursive procedure in Algorithm 1, the solution $\boldsymbol{\eta}_p^*$ to (21) on the set$\mathbf{X}_p$ can be calculated. Then the d optimal projection axes of RGLDA are $\mathbf{u}_p^* = \mathbf{P}_p\boldsymbol{\eta}_p^*$, $p = 1, 2, \ldots, d$.

5. Unsupervised RGLDA (URGLDA)

In this section, we propose a simplified version of unsupervised learning by removing the supervised information in (22). We aim to solve the following optimization problem for the projection $\mathbf{w}$

$$\min_{\mathbf{w},\xi} \ \mathbf{w}^T\mathbf{S}_L\mathbf{w} + \mathrm{v}\sum_i^M \xi_i$$
$$\text{s.t.} \ \left|\mathbf{w}^T\mathbf{H}_i\right| + \xi_i \geq 1, \xi_i > 0, \tag{22}$$

where $\mathbf{H} = \left[\mathbf{x}_1 - \mathbf{m}, \mathbf{x}_2 - \mathbf{m}, \ldots, \mathbf{x}_M - \mathbf{m}\right]$ and $\mathbf{H}_i$ denotes the ith column vector of $\mathbf{H}$. Similar to Section 4, the solution of (22) is ultimately reduced to solving multiple SVM-type problems iteratively. The GLDA can be viewed as a unified semi-supervised learning framework for FDA and Unsupervised Discriminant Projection (UDP). Naturally, GLDA is equivalent to UDP [9] when the label information is not used. Therefore, the problem (22) is a reformulation of UDP. By the similar recursive procedure as that of RGLDA, we can obtain multiple projection axes.

6. Experiments

In this section, we first compare RGLDA with the supervised and semi-supervised subspace learning algorithms, and then compare URGLDA with the representative unsupervised subspace learning algorithms. We use the Nearest Neighbor (NN) classifier for its simplicity and efficiency. We implemented all algorithms in MATLAB 7.1 and carried out experiments on a PC with Intel (R) Core2Duo processor (2.79 GHz), 1 GB RAM.

6.1. Experiments on RGLDA

In this subsection, we evaluate the proposed RGLDA on several image databases. We implemented six algorithms, namely, Baseline, FDA, RFDA [19], supervised LPP (SLPP) [25], SDA [13], and RGLDA. For Baseline, we directly implement the NN without using any DR method. For SLPP, we perform it directly using the codes from http://www.zjucadcg.cn/dengcai. In all experiments, PCA is applied first to preserve 95% energy of the images, similarly to [8,9]. As in algorithms such as SDA, how to set parameters is still an open problem. Therefore, we empirically set the parameters. Specifically, we choose the neighborhood size k between 1 and $m-1$ at sampled intervals of 2 and the parameters α and v between 10^{-4} and 10^4 at sampled intervals of 10. We report the best test results and the optimal dimensionality from the best parameter configuration. Simultaneously, based on the chosen best parameter configuration we also report the best recognition rate on unlabeled data.

1) ORL: The ORL face database [26] contains 400 images of 40 individuals and has become a standard database for testing. The images were captured at different times with the different variations, including expression like open or closed eyes and smiling or non-smiling, and facial details like glasses or no glasses. The images were aligned, cropped and scaled to 32×32 pixels. Each image is represented by a 32×32 (i.e., 1024) dimensional vector in image space. 8 images per individual are randomly selected for training, and the remaining 3 images are used for testing.
Among the training set, we randomly labeled two samples per class and treat other training samples as unlabeled data. The test was repeated ten times and we then computed the mean recognition results of the six algorithms. Table 2 reports the best recognition rates of each method along with the corresponding dimensionality. As can be seen, semi-supervised algorithms SDA and RGLDA outperform the supervised algorithms Baseline, FDA, SLPP, and RFDA, meaning utilizing of unlabeled data can improve the performance of supervised methods. We can also see that our RGLDA method performs better than SDA. The test accuracy for SDA, and RGLDA are 83.3% and 85.4%, respectively. It is worthwhile to note that SLPP works better than FDA. This result is consistent to the observation in [25].
2) UMIST: The UMIST face database [27] contains 575 multi-view grayscale images of 20 individuals, covering a wide range of poses from profile to frontal views. The images were aligned, cropped and are scaled to 23×28 pixels, each of which is represented by a 23×28 (i.e., 644) dimensional vector in image space. Fig. 1 shows some cropped images of one individual. We randomly selected 15 images per class for training, and the rest for testing. Among the training image set, we randomly labeled $l(=2, 3, 4)$ samples per class and the other training images are selected as unlabeled data. We repeated these trials independently ten times and calculated the mean recognition rates of each method. Table 3 gives the results. The conclusions are similar as those drawn from the first experiment on ORL. SDA and RGLDA outperform other methods, while RGLDA performs better than SDA. Fig. 2 shows plots of recognition rates versus the dimensions. Note that for FDA and SDA, the reduced dimension is upper bounded by $c - 1$. As can be seen, RGLDA achieves the best results when the dimensions are within $[1, c - 1]$. From the Fig. 2 and Table 3, it is easy to observe that the performance improvement of Baseline, FDA, RFDA and SLPP is very obvious as the labeled samples increase.

3) COIL20 dataset: The COIL-20 database [28] contains 1440 images of 20 objects, and each object has 72 images captured from varying angles at intervals of five degrees. We resized each image into 32×32 pixels. Half of the images per object, i.e., 36 images, were randomly selected as the training set. Among the training set, we randomly labeled l $(=2, 3, 4)$ samples. For each l, we averaged the results over 10 splits. Table 4 reports the results of Baseline, FDA, RFDA, SLPP, SDA, and RGLDA. The results show that SDA and RGLDA outperform other methods. AS can be seen, the performance of RGLDA is better than that of SDA, which is consistent with the conclusion draw from the above experiments. Fig. 3 plots the recognition rates of all the six algorithms on the labeled dataset and the unlabeled

Table 2
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA Over 10 random runs on ORL database. For each row, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses.

Method	Baseline	FDA	RFDA	SLPP	SDA	RGLDA
Unlabeled Acc. (%)	69.2 (1024)	76.7(33)	75.6 (77)	80.9(39)	83.5(39)	**84.0** (51)
Test Acc. (%)	73.8 (1024)	76.5 (19)	77.9(29)	78.3(41)	83.3(39,)	**85.4** (43)

Fig. 1. Sample images of one individual from UMIST.

Table 3
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA Over 10 random runs on UMIST database. For each row, the results shown in bold are the best among all the results. the optimal dimensions are shown in parentheses.

Method	$l=2$		$l=3$		$l=4$	
	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)
Baseline	60.9 (644)	61.9(644)	71.8 (644)	71.6 (644)	80.3 (644)	79.1 (644)
FDA	69.9 (11)	69.7(13)	80.9 (15)	79.3 (11)	88.7 (15)	87.8 (19)
RFDA	73.8 (11)	72.6 (11)	83.7(11)	83.6 (13)	89.9 (45)	89.5 (71)
SLPP	65.3 (19)	66.9 (19)	76.3 (19)	75.7 (19)	83.7 (19)	82.7 (19)
SDA	75.3 (11)	74.7 (13)	84.6 (15)	84.1 (19)	89.8 (19)	89.9 (5)
RGLDA	**78.7** (15)	**77.9** (15)	**86.3** (13)	**87.0** (19)	**93.1** (21)	**92.3**(21)

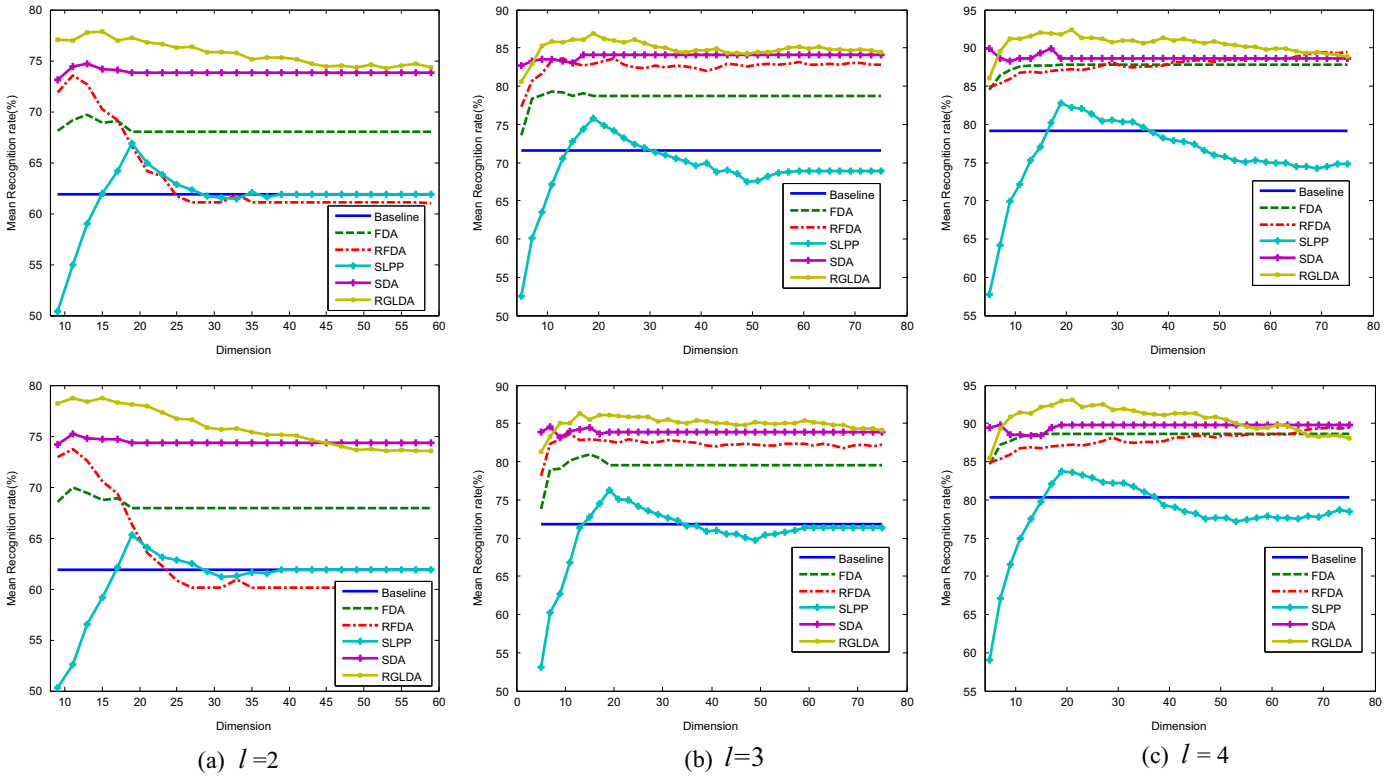

Fig. 2. Recognition rate variation with the dimension. The three columns show the results on the unseen test dataset and the unlabeled dataset, respectively. Two, three and four labeled samples are used in UMIST. (a) $l=2$, (b) $l=3$, and (c) $l=4$.

dataset with respect to the dimensions. From Fig. 3 and Table 4, we can observe that the performance superiority of RGLDA is very evident, compared with SDA and other supervised algorithms. We can also see from Fig. 3 first that the recognition rate of RGLDA decreases slowly when the dimension is over 19, and second that RGLDA

Table 4
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA Over 10 random runs on COLI20 database. For each row, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses.

Method	$l=2$		$l=3$		$l=4$	
	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)
Baseline	72.0 (1024)	72.4 (1024)	77.4 (1024)	77.5 (1024)	80.2 (1024)	79.9 (1024)
FDA	57.1 (17)	56.7 (19)	77.4 (9)	77.3 (9)	80.8 (15)	80.5 (15)
RFDA	73.9 (15)	72.9 (19)	79.9 (15)	79.6 (15)	83.4 (13)	82.9 (13)
SLPP	66.7 (19)	67.0 (19)	71.8 (19)	71.4 (21)	75.5 (21)	75.6 (21)
SDA	78.4 (7)	77.9 (7)	83.6 (7)	83.2(7)	86.0 (7)	85.0 (7)
RGLDA	**80.1** (13)	**80.4** (15)	**85.9** (13)	**86.0** (13)	**88.1** (19)	**88.0** (23)

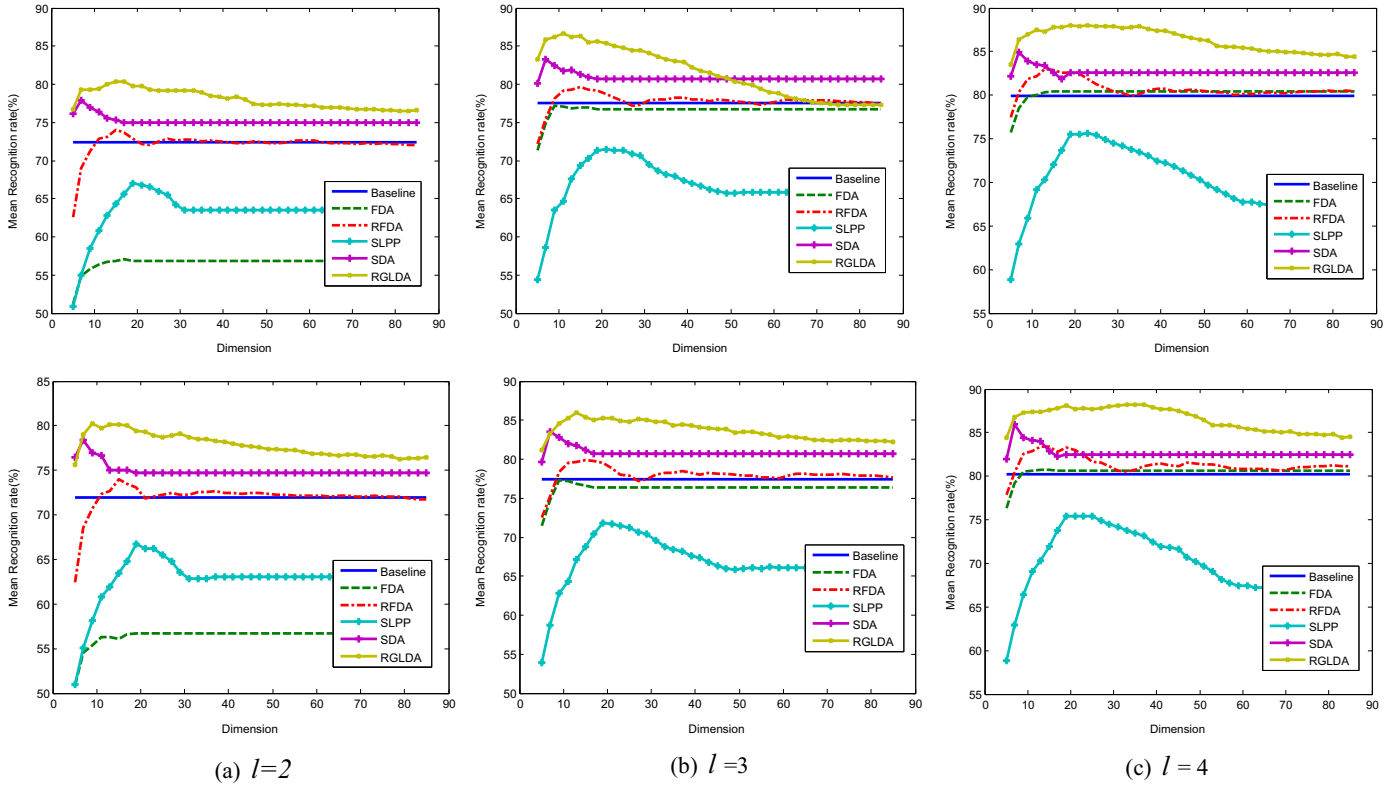

Fig. 3. Recognition rate variation with the dimension. The three columns show the results on the unseen test dataset and the unlabeled dataset, respectively. Two, three and four labeled samples are used in COIL20. (a) $l=2$, (b) $l=3$, and (c) $l=4$.

significantly outperforms other algorithms at any dimension.

4) Handwritten digit dataset: The database [29] contains 390 digit images of 10 classes. Each image is represented by a 20×16 dimensional vector in image space. Fig. 7 shows some images of two digits. We randomly selected 20 samples per class as the training set, and then randomly labeled l ($=2$, 3, 5) samples from this constructed training set per class. For each l, we averaged the results over 10 splits. Table 5 reports the results of all the algorithms. From Table 5, we can see that RGLDA performs the best, and that SDA has the effectiveness over FDA.

5) Dynamic Textures Database [3]: The D1ynTex database is a diverse collection of high-equality dynamic texture videos. We use this database to define 9 different categories of dynamic textures images: waterfall, sea_anemone, candle, lab, leaf, ripple, wave, cloud, and flower. For each class of images, we took 60 frames from two sequences. We resized each image to 50×46 pixels, and then extracted 2300-dimensional gray-level feature for each image. The gray-level values were normalized to unit. Fig. 4 shows two images of each category. Half of the images per class, i.e., 30 images, were randomly selected as the training set, and among the training set, we randomly labeled l ($=2$, 3, 4) samples. For each l, we average the results over 10 splits and report the best results in Table 6. Fig. 5 gives a plot of recognition rates with the dimensions. As can be seen, RGLDA is also very effective on the DynTex database. SDA still outperforms FDA.

6) Extended YALEB dataset: The YALE-B database [34] used in this experiment contains 38 subjects, with each person having around 64 near frontal images under different illuminations. The images are cropped and then resized to 32×32 pixels. In the experiment, we randomly selected half of the images per class as the training set and the remaining images are used as the testing set. Among the training set, we randomly labeled l ($=4$, 5, 6) samples. For each l, we average the results over 10 splits and report the best results in Table 7. As can be seen, RGLDA outperforms FDA, RFDA, SLPP, and SDA, in the terms of recognition rate. SDA is generally better than the supervised methods such as FDA and RFDA.

Table 5
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA over 10 random runs on digit database. For each row, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses.

Method	$l=2$		$l=3$		$l=5$	
	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)
Baseline	65.2 (320)	63.2 (320)	71.1 (320)	70.1 (320)	75.8 (320)	76.0 (320)
FDA	60.6 (9)	59.7 (9)	66.5 (9)	64.3 (9)	72.8 (9)	72.9 (9)
RFDA	65.5 (13)	61.8 (17)	71.9 (21)	70.7 (23)	78.2 (51)	77.7 (45)
SLPP	58.9 (19)	59.8 (19)	64.5 (29)	65.0 (9)	70.9 (17)	71.3 (9)
SDA	63.3 (9)	59.9 (9)	68.9 (9)	64.5 (9)	74.5 (9)	73.8 (9)
RGLDA	**73.1** (9)	**68.9** (23)	**75.1** (29)	**72.5** (29)	**80.9** (15)	**79.6** (35)

Fig. 4. Illustration of 2 images of each of 9 categories from DynTex Database. (a) $l=2$ (b) $l=3$ (c) $l=4$.

Table 6
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA over 10 random runs on DynTex database. For each row, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses.

Method	$l=2$		$l=3$		$l=4$	
	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)
Baseline	67.5 (2300)	68.1 (2300)	77.8 (2300)	78.7 (2300)	81.4 (2300)	82.2 (2300)
FDA	64.1(8)	66.0 (8)	77.1 (8)	78.3 (8)	78.2 (8)	77.8 (8)
RFDA	65.5 (8)	67.6(8)	78.9 (8)	80.6 (10)	85.3 (14)	83.3 (34)
SLPP	70.1 (8)	71.3 (8)	82.5 (8)	82.5 (8)	86.4 (8)	86.9 (8)
SDA	69.5 (8)	71.8 (8)	79.8 (8)	81.3 (8)	82.9 (8)	84.4 (8)
RGLDA	**74.8** (20)	**76.11** (14)	**86.1** (14)	**86.9** (14)	**89.5** (20)	**90.5** (12)

7) ALOI dataset: The ALOI data set contains 72,000 images of 1000 general object categories taken at different illumination directions, illumination temperatures, object orientations, wide-baseline stereo angles. We selected the images of 80 classes under different illumination directions in the experiment. We resized each image to 32×32 pixels. Fig. 6 plots some images of the dataset. Half images of each class were randomly selected for training. We randomly labeled 5 samples. We average the results over 10 splits and report the best results in Table 8. As shown, our RGLDA still performs the best. The same analysis can be found in the previous experiments.

6.2. Experiments on URGLDA

We compare unsupervised RGLDA (URGLDA) and the unsupervised learning algorithms LPP [8] and UDP [9] on ORL, UMIST, Handwritten Digit and DynTex database. Five images per class are randomly chosen as the training set. LPP is directly performed using the codes from http://www.zjucadcg.cn/dengcai. The nearest neighbor classifier is used again for classification after dimension reduction. The neighbor size k in LPP, UDP, and URGLDA is also searched from 1 to $m-1$, and the optimal parameter v in URGLDA is also selected between 10^{-4} and 10^{4} at sampled intervals of 10. We average the results over 10 splits, and report the best results from the optimal parameters in Table 9. We report the best average results over 10 splits from the optimal parameters in Table 9. Fig. 7 plots the recognition rates versus the dimensions. As shown in Table 9 and Fig. 7, URGLDA outperforms LPP and UDP. It is worthwhile to note that LPP works worse than UDP. This result is consistent to the observation in [9].

6.3. Discussion

The experiments have been systematically conducted on several image databases. It is worthwhile to highlight the following some interesting points from the experiments:

1) In all the experiments, RGLDA consistently performs better than Baseline, FDA, RFDA, SLPP, and SDA. Our experiments also show that our algorithm is especially suitable for images classification.
2) When comparing to FDA, SDA, and RFDA, both SDA and RFDA are consistent winners. The result is consistent to the observations in [8,19]. The experiments also demonstrate that for the different labeled sample size, SDA generally performs better than RFDA and SLPP.
3) Our unsupervised method RGLDA has the performance advantages over LPP and UDP. In fact, RGLDA is a reformulation of UDA, which casts the problem of UDP as the related SVM-type one. Therefore, the results again disclose that this trick of reformulating an eigenvalue based technique as a SVM-type problem can improve the performance. It can be also applied to most of eigenvalue based dimension reduction techniques.

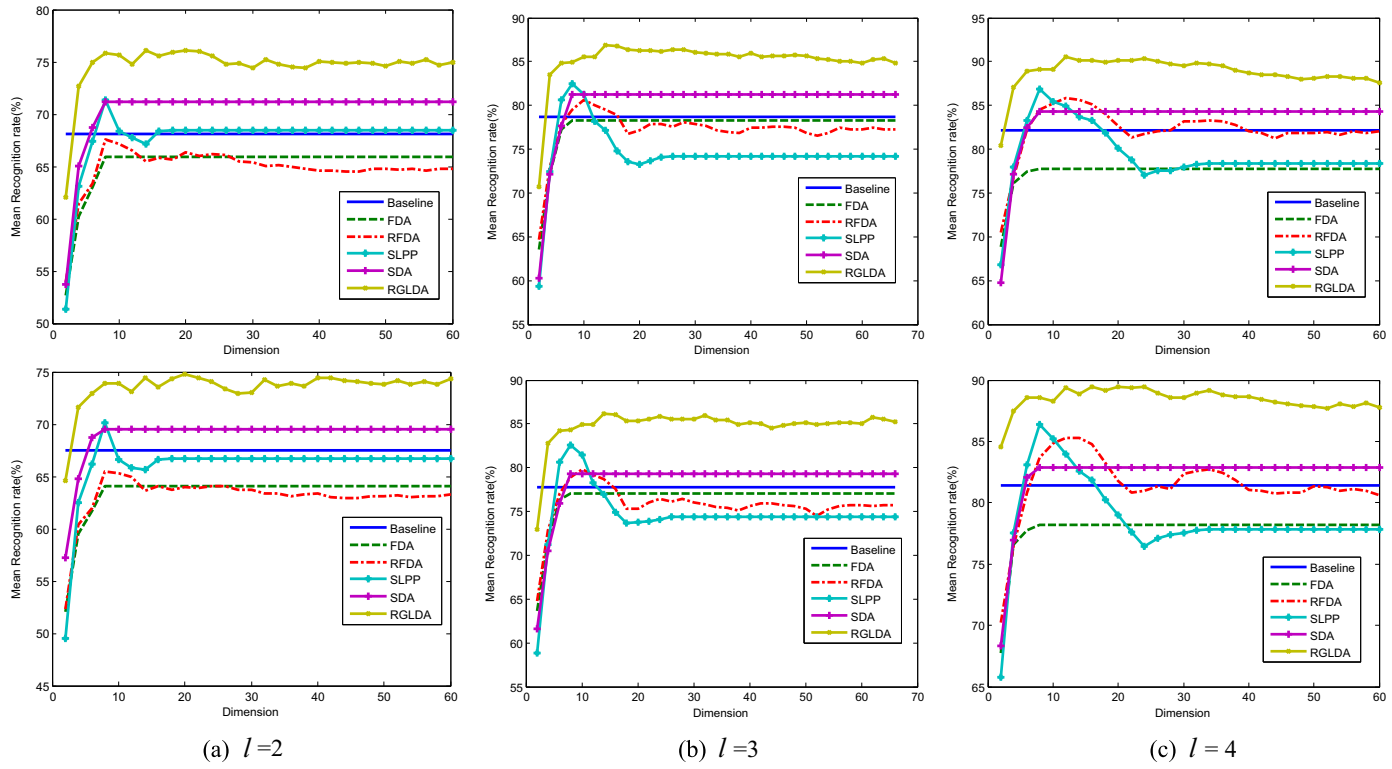

Fig. 5. Recognition rate variation with the dimension. The three rows show the results on the unseen test dataset and the unlabeled dataset, respectively. Two, three and four labeled samples are used in DynTex. (a).

Table 7
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA over 10 random runs on extended YALEB database. For each row, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses.

Method	l=4		l=5		l=6	
	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)	Unlabeled Acc. (%)	Test Acc. (%)
Baseline	76.0(1024)	80.6 (1024)	86.3 (1024)	88.5 (1024)	90.2(1024)	92.1 (1024)
FDA	86.0(37)	85.6 (37)	92.4 (37)	91.5 (37)	94.2(37)	94.1(37)
RFDA	87.5(43)	86.5 (39)	93.5 (47)	93.1 (51)	95.4 (37)	94.5 (37)
SLPP	87.7(37)	88.5 (41)	93.7 (43)	92.9 (39)	95.7 (51)	94.2 (51)
SDA	90.5(37)	90.5 (37)	93.5 (37)	93.5 (37)	95.6 (37)	95.5 (37)
RGLDA	**92.5**(39)	**92.9** (43)	**94.5** (37)	**95.5** (45)	**97.6** (51)	**97.7** (49)

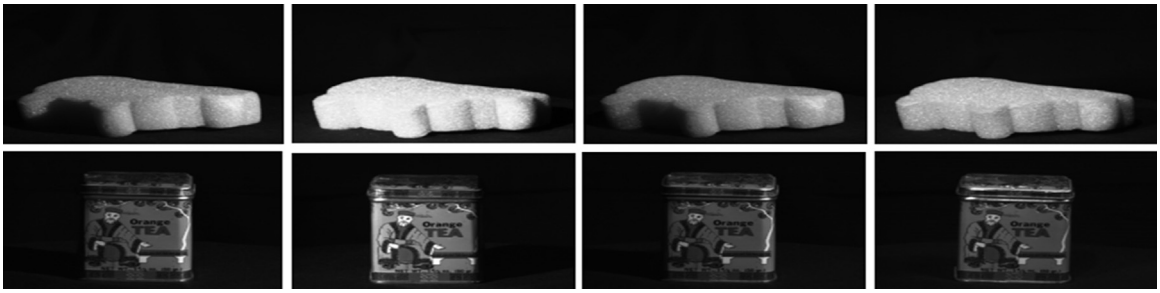

Fig. 6. Illustration of 4 images of two categories from ALOI Database.

Table 8
Comparative best recognition rates of baseline, FDA, RFDA, SLPP, SDA, and RGLDA over 10 random runs on ALOI database. For each row, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses.

Method	Baseline	FDA	RFDA	SLPP	SDA	RGLDA
Unlabeled Acc. (%)	84.4 (1024)	88.7 (79)	89.6 (89)	89.1 (89)	91.5 (79)	**93.5** (84)
Test Acc. (%)	82.7 (1024)	87.5 (79)	88.2 (79)	89.3 (84)	92.5 (79)	**84.5** (89)

Table 9
Comparative best recognition rates of LPP, UDP, and URGLDA over 10 runs on ORL, UMIST, Digit, DynTex, and COIL20. For each column, the results shown in bold are the best among all the results. The optimal dimensions are shown in parentheses, and parameters are shown in parentheses (k in SDA, and k and v RGLDA).

Method	ORL	UMIST	DIGIT	DynTex	COIL20
LPP	88.2 (47)	82.4 (19)	66.3 (25)	83.5 (14)	79.4 (29)
UDP	89.9 (59)	84.7 (43)	73.6 (34)	84.0 (10)	79.2 (31)
URGLDA	**90.2**(51)	**87.1** (43)	**76.3** (46)	**87.1** (26)	**82.6** (65)

7. Extension: sparse dimension reduction technique

To date, there are few techniques proposed to construct the sparse models for eigenvalue-based dimension reduction methods in multiclass setting. This is due to the fact that there is no obvious way to introduce sparsity. Existing sparse techniques, such as Sparse PCA (SPCA) [15], are to impose L_1-penalized regression on projection vectors. With the similar technique, any supervised or semi-supervised dimension reduction method, such as SDA, can obtain the sparse solution. However, such a technique is a two-stage approach. Intuitively, it is too naive to believe that there is no any information loss in the two-stage processing. We believe that sparsity should be introduced into the original objective for the guarantee of obtaining an optimal performance. Here, we only give our sparse RGLDA. To obtain the sparse RGLDA, we incorporate a L_1-penalized term into (17), and then we get

$$\begin{aligned}&\min_{\mathbf{w},\xi}\ \mathbf{w}^{\mathrm{T}}\mathbf{S}_D\mathbf{w}+\mathrm{v}\sum_{i}^{l+M}\xi_i+\omega\|\mathbf{w}\|_1\\&\text{s.t.}\ \ \left|\mathbf{w}^{\mathrm{T}}\mathbf{H}_i\right|+\xi_i\geq 1,\ \ \xi_i>0,\end{aligned}\tag{23}$$

where ω is a trade-off coefficient. Setting: $\mathbf{w}=\mathbf{p}-\mathbf{q}$, $\mathbf{p}\geq 0$, $\mathbf{q}\geq 0$, the problem (23) becomes

$$\begin{aligned}&\min_{\mathbf{w},\xi}\ (\mathbf{p}-\mathbf{q})^{\mathrm{T}}\mathbf{S}_D(\mathbf{p}-\mathbf{q})+\mathrm{v}\sum_{i}\xi_i+\omega\mathbf{e}^{\mathrm{T}}(\mathbf{p}+\mathbf{q})\\&\text{s.t.}\ \ \left|(\mathbf{p}-\mathbf{q})^{\mathrm{T}}\mathbf{H}_i\right|+\xi_i\geq 1,\ \xi_i\geq 0,\ i=1,2,\ldots,l+M.\end{aligned}\tag{24}$$

where $\mathbf{e}$ is a column vector of ones of appropriate dimensions. Letting $\boldsymbol{\tau}=\begin{bmatrix}\mathbf{p}\\ \mathbf{q}\end{bmatrix}$, (24) is simplified as

$$\begin{aligned}&\min_{\mathbf{w},\xi}\ \boldsymbol{\tau}^{\mathrm{T}}\big([\mathbf{I}\quad -\mathbf{I}]^{\mathrm{T}}\mathbf{S}_D[\mathbf{I}\quad -\mathbf{I}]\big)\boldsymbol{\tau}+\mathrm{v}\sum_{i}\xi_i+\omega\mathbf{e}^{\mathrm{T}}[\mathbf{I}\quad -\mathbf{I}]\boldsymbol{\tau}\\&\text{s.t.}\ \ \left|\boldsymbol{\tau}^{\mathrm{T}}[\mathbf{I}\quad -\mathbf{I}]^{\mathrm{T}}\mathbf{H}_i\right|+\xi_i\geq 1,\ \xi_i\geq 0,\ i=1,2,\ldots,l+M.\end{aligned}\tag{25}$$

It is not difficult to see that the optimization problem in (25) is similar as (17) in formulation that can be solved with the CCP. Detailed experimental study is left for future work.

8. Conclusion

In this paper, we proposed a recursively SVM-type dimension reduction framework. Then, a global and local discriminant method for semi-supervised and unsupervised dimension reduction was developed. The semi-supervised method RGLDA focuses on not only local but also global geometries, making the manifold regularization in RGLDA have a direct connection to classification. Different from existing eigenvalue based dimension reduction methods, RGLDA seeks for the projection axes by solving the SVM-type problems recursively. RGLDA is very flexible. Specifically, the basic idea behind RGLDA is very novel and significant, which can also be applied to most of existing methods. Based on the reformulation of RGLDA, it is easy to construct a new sparse model without any assumption and need of the multi-stage processing. When the labeled

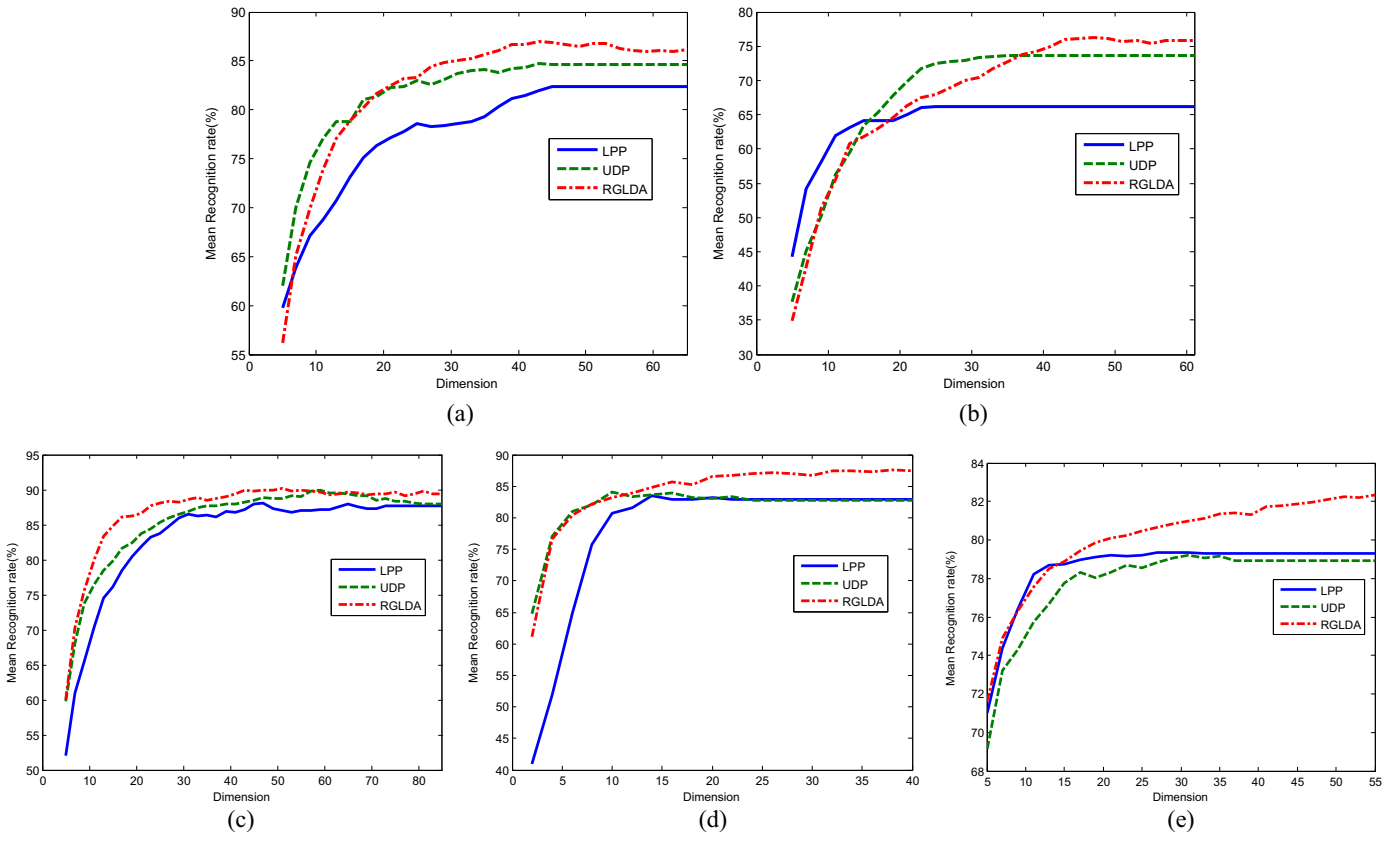

Fig. 7. Recognition rates with different dimensions on the ORL, UMIST, Digit, DynTex, and COIL20. (a) ORL, (b) UMIST, (c) Digit, (d) DynTex, and (e) COIL20.

information is not used, we extended RGLDA to URGLDA. The comprehensive experiments on five image databases demonstrate the effectiveness of RGLDA and URGLDA. Certainly, just as RFDA, our methods are more expensive computationally than SDA. However, once the dimensionality is reduced and ready to be used by any classifier, there will be no extra computational cost. Both RGLDA and URGLDA are linear. In the future, we will extend them to the nonlinear cases in the reproducing kernel Hilbert space as well as examine how to determine the optimal parameters. We will also extend RGLDA and URGLDA to new supervised, semi-supervised, and unsupervised dimension reduction algorithms.

Acknowledgments

This work was supported in part by the National Foundation for Distinguished Young Scientists under Grant 31125008, the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology) under Grant no. 30916014107, the Scientific Research Foundation for Advanced Talents and Returned Overseas Scholars of Nanjing Forestry University under Grant 163070679, the Open Fund of Jiangsu Provincial Key Laboratory for Advanced Manufacturing Technology (HGAMTL-1401), and the National Science Foundation of China under Grants 61402192, 61101197, 61272220, and 61401214, Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 14KJB580002, Six Talent Peaks Project in Jiangsu Province.

References

[1] M. Turk, and A.P. Pentland, Face Recognition Using Eigenfaces, in: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, 1991.
[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[3] K. Liu, Y.Q. Cheng, J.Y. Yang, X. Liu, An efficient algorithm for Foley-Sammon optimal set of discriminant vectors by algebraic method, Int'l J. Pattern Recognit. Artif. Intell. 6 (5) (1992) 817–829.
[4] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.
[5] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[6] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[7] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[8] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacian faces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[9] J. Yang, D. Zhang, J.-Y. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 650–664.
[10] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from examples, J. Mach. Learn. Res. 7 (Nov. 2006) 2399–2434.
[11] V. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.
[12] V. Sindhwani, P. Niyogi, and M. Belkin, Linear manifold regularization for large scale semi-supervised learning, in: Proceedings of the 22nd ICML Workshop on Learning With Partially Classified Training Data, Bonn, Germany, 2005.
[13] D. Cai, X.F. He, and J.W. Han, Semi-Supervised Discriminant Analysis, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, Oct. 2007.
[14] S. Mika, G. Rätsch, K.R. Müller, A mathematical programming approach to the Kernel Fisher algorithm, Adv. Neural Inf. Process. Syst. (2001).
[15] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Statist. 15 (2) (2006) 262–286.
[16] R. Jayadeva, Khemchandani, S. Chandra, Twin Support Vector Machines for pattern classification, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 905–910.
[17] X.B. Chen, J. Yang, Q.L. Ye, Jun Liang, Recursive projection twin support vector machine via within-class variance minimization, Pattern Recognit. 44 (44) (2011) 2643–2655.
[18] I.T. Jolliffe, Principal component, Analysis, Springer-Verlag, New York, 1986.
[19] C. Xiang, X.A. Fan, T.H. Lee, Face recognition using recursive fisher linear discriminant, IEEE Trans. Image Process. 15 (8) (2006) 2097–2105.
[20] O.L. Mangasarian, E.W. Wild, Multisurface proximal support vector classification via generalized eigenvalues, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (Jan. 2006) 69–74.
[21] Q.L. Ye, C.X. Zhao, N. Ye, Y.N. Chen, Multi-weight vector projection support vector machines, Pattern Recognit. Lett. vol. 31 (3) (2010) 2006–2011.
[22] A.J. Smola, S. Vishwanathan, T. Hofmann, Kernel methods for missing variables, AISTATS (2005).
[23] B. Zhao, F. Wang, C.S. Zhang, Efficient Maximum Margin Clustering via Cutting Plane Algorithm, in: Proceedings of SDM'2008. pp.751–762.
[24] T.H. Zhang, K.Q. Huang, et al., Discriminative orthogonal neighborhood-preserving projections for classification (Feb.), IEEE Trans. Syst., Man, Cybern.-Part B: Cybern. 40 (1) (2010) 253–263.
[25] D. Cai, X.-F. He, and J.W. Han, Using Graph Model for Face Analysis, Department of Computer Science Technical Report No. 2636, University of Illinois at Urbana-Champaign (UIUCDCS-R-2005–2636), Sept. 2005.
[26] O. Face Database. [Online]. Available: ⟨http://www.uk.research.att.com/facedatabase.html⟩.
[27] D.B. Graham, N.M. Allinson, Characterizing virtual eigensignatures for general purpose face recognition, in face recognition: from theory to applications ser, in: H. Wechsler, P.J. Phillips, V. Bruce, F. FogelmanSoulie, T.S. Huang (Eds.), NATO ASI Series F, Computer and Systems Science, 163, Springer-Verlag, New York, 1998, pp. 446–456.
[28] S.A. Nene, S.K. Nayar, and H. Murase, Columbia object image library (COIL-20), Columbia Univ., New York, Tech. Rep. CUCS-005-096, 1996.
[29] H. Digit Dataset. [Online]. ⟨http://www.cs.nyu.edu/~roweis/data.html⟩.
[30] DynTex. ⟨http://old-www.cwi.nl/projects/dyntex/database.html⟩.
[31] Q.L. Ye, C.X. Zhao, X.B. Chen, Recursive "concave-convex", Fish. Linear Discrim. Appl. Face, Handwrit. Digit Terrain Recognition", Pattern Recogn. 45 (1) (2012) 521–534.
[32] Y. Huang, D. Xu, F. Nie, Semi-supervised dimension reduction using trace ratio criterion, EEE Trans. Neural Netw. Learn. Syst. 23 (3) (2012) 519–526.
[33] F. Dornaika, Y. Traboulsi, Learning flexible graph-based semi-supervised embedding, IEEE Trans. Cybern. 45 (6) (2015) 206–218.
[34] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[35] Q.L. Ye, N. Ye, T.M. Yin, Fast orthogonal linear discriminant analysis with application to image classification, Neurocomputing 158 (2015) 216–224.
[36] J. Lu, Y. Tan, Improved discriminant locality preserving projections for face and palmprint recognition, Neurocomputing 74 (18) (2011) 3760–3767.
[37] J. Yu, Y. Guo, D. Tao, Human pose recovery by supervised spectral embedding, Neurocomputing 166 (2015) 301–308.
[38] Q.L. Ye, N. Ye, et al., Flexible orthogonal semisupervised learning for dimension reduction with image classification, Neurocomputing 144 (2014) 417–426.
[39] F.P. Nie, D. Xu, X.L. Li, S.M. Xiang, Semisupervised dimensionality reduction and classification through virtual label regression (Jun), IEEE Trans. Syst., Man, Cyber. B, Cyber. 41 (3) (2011) 675–685.
[40] F.P. Nie, D. Xu, I.W.H. Tsang, C.S. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction (Jul), IEEE Trans. Image Process. 19 (11) (2010) 1921–1932.
[41] T.L. Liu, D.C. Tao, and D. Xu, Dimensionality-Dependent Generalization Bounds for k-Dimensional Coding Schemes, arXiv preprint arXiv:1601.00238, 2016.
[42] T.L. Liu, and D.C. Tao, On the performance of manhattan nonnegative matrix factorization, IEEE Transactions on Neural Networks and Learning Systems, 2015.
[43] T.L. Liu, D.C. Tao, Classification with noisy labels by importance reweighting, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 447–461.
[44] B. Wei, D.C. Tao, Asymptotic generalization bound of fisher's linear discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 36 (12) (2014) 2325–2337.
[45] J. Yu, D.C. Tao, et al., Learning to rank using user clicks and visual features for image retrieval, IEEE Trans. Cybern. 45 (4) (2015) 767–779.
[46] J. Yu, D.C. Tao, High-order distance based multiview stochastic learning in image classification, IEEE Trans. Cybern. 44 (12) (2014) 2431–2442.
[47] J. Yu, M. Wang, D.C. Tao, Semi-supervised multiview distance metric learning for cartoon synthesis, IEEE Trans. Image Process. 21 (11) (2012) 4636–4648.

Shangbing Gao received the BS degree in mathematics from the Northwestern Polytechnical University in 2003. He received the MS degree in applied mathematics from the Nanjing University of Information and Science and Technology in 2006. He is now working at Huaiyin institute of technology as an assistant lecturer. He is currently pursuing the pH.D. degree with School of Computer Science and Technology, Nanjing University of Science and Technology (NUST). He is on the subject of pattern recognition and intelligence systems. His current research interests include pattern recognition and computer vision.

Qiaolin Ye received the BS degree in Computer Science from Nanjing Institute of Technology, Nanjing, China, in 2007, the MS degree in Computer Science and Technology from Nanjing Forestry University, Jiangsu, China, in 2009, and the pH.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology, Jiangsu, China, in 2013. He is currently an associate professor with the computer science department at the Nanjing Forestry University, Nanjing, China. He has authored more than 50 scientific papers. Some of them are published in IEEE TNNLS, IEEE TIFS, and IEEE TCSVT. His research interests include machine learning, data mining, and pattern recognition.