Locality-Preserved Maximum Information Projection

Haixian Wang, Sibao Chen, Zilan Hu, and Wenming Zheng

***Abstract*—Dimensionality reduction is usually involved in the domains of artificial intelligence and machine learning. Linear projection of features is of particular interest for dimensionality reduction since it is simple to calculate and analytically analyze. In this paper, we propose an essentially linear projection technique, called locality-preserved maximum information projection (LPMIP), to identify the underlying manifold structure of a data set. LPMIP considers both the within-locality and the between-locality in the processing of manifold learning. Equivalently, the goal of LPMIP is to preserve the local structure while maximize the out-of-locality (global) information of the samples simultaneously. Different from principal component analysis (PCA) that aims to preserve the global information and locality-preserving projections (LPPs) that is in favor of preserving the local structure of the data set, LPMIP seeks a tradeoff between the global and local structures, which is adjusted by a parameter α, so as to find a subspace that detects the intrinsic manifold structure for classification tasks. Computationally, by constructing the adjacency matrix, LPMIP is formulated as an eigenvalue problem. LPMIP yields orthogonal basis functions, and completely avoids the singularity problem as it exists in LPP. Further, we develop an efficient and stable LPMIP/QR algorithm for implementing LPMIP, especially, on high-dimensional data set. Theoretical analysis shows that conventional linear projection methods such as (weighted) PCA, maximum margin criterion (MMC), linear discriminant analysis (LDA), and LPP could be derived from the LPMIP framework by setting different graph models and constraints. Extensive experiments on face, digit, and facial expression recognition show the effectiveness of the proposed LPMIP method.**

***Index Terms*—Dimensionality reduction, image recognition, linear projection, manifold learning, undersampled problem.**

I. Introduction

IN MANY areas of artificial intelligence, statistical pattern recognition, computer vision, and data mining, one is often confronted with the situation of excessive dimension of a pattern space. For example, a facial image with the resolution of 128 × 128 pixels is represented as a point in a 16 384-dimensional face space. To learn a classifier well, the number of training samples needed increases rapidly with respect to the dimension. It is a good practice to follow that at least ten times as many samples per class as the dimension are used [11]. In many real-world applications, such large number of samples, however, are fairly expensive to obtain due to limitations of sample availability, time, and cost. Therefore, one fundamental problem in many scientific subjects is dimensionality reduction, which seeks a meaningful low-dimensional representation of high-dimensional data. Dimensionality reduction could effectively avoid the "curse of dimensionality" [16], improve performance and computational efficiency of pattern classification, suppress noise, and alleviate storage requirement. In fact, the intrinsic dimension of the pattern space is low. Taking an example, the facial images varying in pose, expression, or rotation may reside on an intrinsically low-dimensional manifold embedded in high-dimensional space.

In the past few decades, many useful techniques for dimensionality reduction have been developed. Linear combination of features is of particular interest since it is simple to calculate and analytically analyze. That is, dimensionality reduction is realized via linear projection. The most well-known techniques for this purpose may be principal component analysis (PCA) [15], [17] and linear discriminant analysis (LDA) [10], [30]. PCA, also known as Karhunen–Loéve transformation, aims to find a set of mutually orthogonal bases that capture the global information of the data points in terms of variance. PCA has been successfully applied to discover the subspace of face space, which is termed eigenfaces method [27], [36]. By contrast with the unsupervised method of PCA, LDA, also called Fisher's linear discriminant, is a supervised learning approach. LDA seeks a subspace projected onto which the data points of different classes are far away while the data points of the same class are close to each other. The optimal transformation is computed by minimizing the within-class scatter and maximizing the between-class scatter simultaneously. LDA has been extensively used in face recognition creating the popular Fisherfaces method [3], [27].

However, one critical drawback of LDA is that it suffers from the small sample size (SSS) or undersampled problem [9], [20], [44]. To overcome this limitation, a large number of LDA extensions were proposed in literature [21], [45] (and references therein). Particularly, orthogonal LDA (OLDA) is one of such examples [43]. More recently, Li *et al.* [22] developed a maximum margin criterion (MMC) from another perspective as an efficient and robust feature extraction criterion instead of Fisher's criterion. The new criterion is general in the sense that it turns out to be LDA when imposed a suitable constraint. Although both LDA and MMC are two supervised subspace learning methods, MMC effectively avoids the inverse matrix

Manuscript received October 22, 2006; revised May 23, 2007 and July 16, 2007; accepted August 6, 2007. This work was supported in part by the National Natural Science Foundation of China under Grants 60503023 and 10571001, the Natural Science Foundation of Jiangsu Province under Grant BK2005407, the Program for New Century Excellent Talents in University under Grant NCET-05-0467, and the Program of Excellent Young Teachers in Southeast University.

H. Wang and W. Zheng are with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast Unversity, Nanjing, Jiangsu 210096, China (e-mail: hxwang@seu.edu.cn).

S. Chen is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, China.

Z. Hu is with the School of Mathematics and Physics, Anhui University of Technology, Maanshan, Anhui 243002, China.

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/TNN.2007.910733

operation and the SSS problem, which makes the implementation of MMC much easier. Yan *et al.* [39] gave the incremental version of MMC, and Zheng *et al.* [49] and Liu *et al.* [25] improved MMC by regularization. The computational issue of MMC was commented in [24].

Linear models, however, may fail to find the underlying nonlinear structure of a data set. To remedy this deficiency, a number of nonlinear dimensionality reduction techniques have been developed in the past few years, among which two received increasing attention: kernel-based counterparts of linear prototypes and manifold-learning-based approaches. The basic idea of kernel-based method is to first map the input data points into some higher or possibly infinite-dimensional feature space $\mathcal{F}$ typically via a nonlinear function ϕ and then carry out a linear algorithm in $\mathcal{F}$ using the mapped samples [34]. In implementation, the mapping ϕ and thus the space $\mathcal{F}$ are determined implicitly by the choice of *kernel function*. Kernel principal component analysis (KPCA) [33], generalized discriminant analysis (GDA) [2], [23], kernel discriminative common vector [7], and kernel (regularized) MMC [22], [25], [49] are the representative approaches, which are effectively applied to face recognition [40], [41]. However, the kernel-based techniques are computationally intensive, and do not explicitly consider the local structure of a data set, which is important for classification purpose.

By contrast, manifold learning takes the local information of data structure into account, aiming to directly discover the globally nonlinear data structure. The desired manifold is an intrinsically low-dimensional space hidden in the input space. The most well-known manifold learning algorithms include isometric feature mapping (Isomap) [35], locally linear embedding (LLE) [32], and Laplacian eigenmaps [4]. Recently, Yan *et al.* [38] introduced a general framework for dimensionality reduction, called graph embedding, where a large number of popular dimensionality reduction algorithms, e.g., PCA, LDA, Isomap, LLE, and Laplacian eigenmaps, could be considered as special cases within the framework. Also, the out-of-sample problem of Isomap, LLE, and Laplacian eigenmaps is addressed in [5] and [38]. Particularly, locality preserving projections (LPPs) [13] is the optimal linear approximation to the eigenfunctions of Laplace Beltrami operator on the manifold derived from Laplacian eigenmaps. LPP is essentially linear while considering manifold structure via adjacency graph. Based on LPP, the Laplacianfaces was further developed for face recognition [14], giving encouraging performance.

However, there exists one common problem with current manifold learning algorithms; that is, they might not necessarily discover the most important manifold for pattern discrimination tasks. Manifold learning, which is to pursue *locality* characterization of the data, is not originally and essentially designed for discrimination purpose. If the patterns needed to be classified take on multimanifolds (corresponding to different classes) and two or more modes have common chief axis, then the locality-preserving algorithms of manifold learning may result in overlapped embeddings belonging to different classes, which deteriorates the discrimination performance. This thus raises a problem that we refer to as "overlearning of locality" for current manifold learning algorithms and some supervised extensions [18], [37], since they all model data structure on the basis of locality. In other words, for classification problem, the locality quantity itself is not sufficient. We, therefore, consider introducing a *between-locality* quantity to enlarge the distances between embeddings belonging to different classes. For terminology clarity, we refer to conventional *locality* as *within-locality*, and they are interchangeably used here.

In this paper, we propose a new approach, called locality-preserved maximum information projection (LPMIP), to perform linear projection. LPMIP takes within-locality and between-locality into account simultaneously in the modeling of manifold, and motivated by the idea of MMC [22], [25], we present an objective function that seeks to maximize the difference, rather than the ratio, between the *between-locality* and *within-locality*. We expect that, after maximizing the criterion, the embeddings corresponding to the same manifold are close to each other while the embeddings corresponding to different manifolds are far away from each other. Similar with LPP, the characterization of within-locality in LPMIP is based on the adjacency graph that incorporates the neighborhood information of the data points, and by contrast, the characterization of between-locality is based on a dissimilarity graph that embodies the interneighborhood information of the data points. The projections are then obtained by employing such two graphs, as discussed in Section III. LPMIP could be applied for discrimination in unsupervised or supervised way, since we could build the adjacency matrix by using the label information of samples indicating class memberships or by the nearest-neighborhood. It is worthwhile to highlight some properties of LPMIP from a number of perspectives.

- LPMIP shares some similar good properties with LPP. Like LPP, LPMIP is linear and defined on both the training and the testing data sets. Thus, it is straightforward to evaluate the image of any new data point in the reduced-dimensional space. However, their objective functions are totally different. The basis functions of LPP are not necessarily orthogonal while LPMIP produces orthogonal basis functions. Although orthogonal LPP [6] was proposed. It is, however, rather computationally complex. LPP only considers the within-locality while LPMIP considers the additional between-locality as well, so unlike LPP, LPMIP is a directly oriented classification. Although there are variants of LPP for classification such as marginal Fisher analysis (MFA) [38], local discriminant embedding (LDE) [8], discriminant neighborhood embedding (DNE) [47], and discriminant locality preserving projections (DLPP) [46], they are supervised and similar in formulations, since they all aim to characterize within-class local compactness and between-class local separability by using the class-label information. In other words, these methods introduce locality into within-class and between-class, respectively, and thus could be viewed as localized versions of LDA. By contrast, LPMIP combines locality and out-of-locality information and could be performed in unsupervised or supervised way for discrimination task. It could be noted that LPMIP is somewhat similar with these methods, for example, [38], in the sense that they all use two locality-related graphs to characterize the data set. However, the formulations and interpretations of the objective functions of LPMIP and these methods are completely and essentially different. LPMIP has an additional regularization parameter α that balances the within-locality and the between-locality. In many real-world classification problems, both the locality and out-of-locality manifold structures are of importance. It really reflects the intrinsic geometry of the data set.

- Computationally, LPMIP completely avoids the singularity problem as it exists in LPP [14] and MFA [38], since it does not involve any inverse matrix operation. Further, we develop an efficient and stable algorithm for performing LPMIP, namely, LPMIP/QR, which couples the QR-decomposition into the LPMIP framework. LPMIP/QR significantly alleviates the computational burden especially on high-dimensional data set, and we justify LPMIP/QR by showing the equivalence between LPMIP/QR and LPMIP theoretically. The incremental property of LPMIP/QR makes it desirable in high-dimensional and dynamic databases.
- As a connection to LPP, we could derive LPP from the LPMIP framework by imposing some constraints. Likewise, the conventional linear projections such as (weighted) PCA, MMC, and LDA could also be induced from the LPMIP framework by setting different graph models and constraints.

The rest of this paper is organized as follows. Section II briefly reviews the conventional linear projection methods, i.e., PCA, LDA, MMC, and LPP. The LPMIP is developed in Section III. In Section IV, we give a theoretical analysis of LPMIP and discuss its relations with (weighted) PCA, MMC, LDA, and LPP. The experimental results are presented in Section V. We conclude this paper in Section VI.

II. Brief Reviews of Conventional Linear Projection Methods

Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a set of p-dimensional samples of size n. The generic problem of linear dimensionality reduction is to find a linear transformation $\mathbf{V} \in \mathbb{R}^{p\times q}$ that maps each vector $\mathbf{x}_i$ $(i = 1, \ldots, n)$ in the p-dimensional space to a vector $\mathbf{y}_i$ in the lower q-dimensional space by $\mathbf{y}_i = \mathbf{V}^T\mathbf{x}_i$ such that $\mathbf{y}_i$ "represents" $\mathbf{x}_i$ well in terms of some optimal criterion. For description simplicity and without loss of generality, we particularly consider finding a linear mapping from the p-dimensional space to a line, i.e., $y_i = \mathbf{v}^T\mathbf{x}_i$, where the transformation vector is denoted by $\mathbf{v}$. Since the magnitude of $\mathbf{v}$ is of no real interest, we constraint it to have unitary norm. Different criterion functions pursue different goals, resulting in different algorithms.

A. Principal Component Analysis

PCA seeks a subspace, in which the projected global variance reaches maximization. The variance could be equivalently computed by the sum of all squared pairwise Euclidean distances between the projected data points [19], [42]. The criterion function of PCA is

$$\max \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \propto \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 \qquad (1)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$. The directions $\mathbf{v}_1, \ldots, \mathbf{v}_q$ (called principal axes) are given by the q leading orthogonal eigenvectors (corresponding to the q largest eigenvalues) of the sample covariance matrix. PCA is an orthogonal transformation of the coordinate axes, in which we describe the observed data, resulting in uncorrelated principal components. From the perspective of reconstruction, PCA finds the principal axes that are useful for the representation of the data in the sense of minimum reconstruction error.

B. Linear Discriminant Analysis

LDA seeks optimal directions in the sense that they are efficient for classification. Suppose that the n samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ belong to c classes. The number of samples in the ith class is n_i satisfying $\sum_{i=1}^{c} n_i = n$. We use $\mathbf{x}_i^j$ to denote the ith sample in the jth class for $i = 1, \ldots, n_j$, $j = 1, \ldots, c$. The between-class scatter matrix $\mathbf{S}_b$ and within-class scatter matrix $\mathbf{S}_w$ are, respectively, defined as

$$\mathbf{S}_b = \frac{1}{n}\sum_{j=1}^{c} n_j(\mu_j - \mu)(\mu_j - \mu)^T \qquad (2)$$

$$\mathbf{S}_w = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\mathbf{x}_i^j - u_j\right)\left(\mathbf{x}_i^j - u_j\right)^T \qquad (3)$$

where μ is the mean of the entire samples, i.e., $\mu = (1/n)\sum_{i=1}^{n}\mathbf{x}_i$, and $\mu_j = (1/n_j)\sum_{i=1}^{n_j}\mathbf{x}_i^j$ is the mean of the jth class. Then, LDA seeks to maximize the Fisher criterion, which is given by

$$\max \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}}. \qquad (4)$$

The projection directions $\mathbf{v}$ are the generalized eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_q$ solving $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$ associated with the first q largest eigenvalues. These generalized eigenvectors, which are not necessarily orthogonal, form the basis functions of LDA.

C. MMC and Its Regularization

MMC aims to find the optimal transformation $\mathbf{v}$ such that the class structure (similarity/dissimilarity) information of the original high-dimensional space is well preserved in the transformed lower dimensional space. Geometrically, MMC maximizes the (average) margins between different classes. The objective function is given by

$$\max\ \mathbf{v}^T(\mathbf{S}_b - \mathbf{S}_w)\mathbf{v} \qquad (5)$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are, respectively, the between-class scatter matrix and the within-class scatter matrix as defined previously. The optimal transformations $\mathbf{v}_1, \ldots, \mathbf{v}_q$ are immediately computed as the q leading eigenvectors of $\mathbf{S}_b - \mathbf{S}_w$. The regularized

MMC (RMMC) is to maximize $\mathbf{v}^T(\mathbf{S}_b - \gamma\mathbf{S}_w)\mathbf{v}$ instead of (5) with the nonnegative regularized parameter γ.

D. Locality-Preserving Projection

LPP seeks a subspace that preserves the local structure of the data set. LPP models the manifold structure explicitly by constructing the nearest-neighbor graph that reveals neighborhood relationship between data points. In the process of projection, LPPs try to maintain this graph structure. The objective function of LPP is given by

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2\mathbf{A}(i,j) \tag{6}$$

where $\mathbf{A}(i,j)$ denotes the similarity between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$. The criterion (6) could be minimized subject to different constraints. One way is to assume that $\mathbf{v}$ are of unitary norms as in PCA. Then, solving the optimization problem (6) under these constraints yields orthogonal basis functions $\mathbf{v}_1,\ldots,\mathbf{v}_q$, which are the leading eigenvectors of $\mathbf{XLX}^T$. Here, $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n]$, the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, and $\mathbf{D}$ is a diagonal matrix whose entries on diagonal are column sum of $\mathbf{A}$, i.e., $\mathbf{D}(i,i) = \sum_{j=1}^{n}\mathbf{A}(i,j)$. Another constraints are to assume $\mathbf{v}^T\mathbf{XDX}^T\mathbf{v} = 1$, which, in fact, are adopted in original LPP [13]. Using these constraints to minimize the criterion (6), we obtain conjugately orthogonal projection axes that are the leading generalized eigenvectors of $\mathbf{XLX}^T\mathbf{v} = \lambda\mathbf{XDX}^T\mathbf{v}$.

III. Locality-Preserved Maximum Information Projection

Given the samples $\mathbf{x}_1,\ldots,\mathbf{x}_n$, they could be represented by a weighted undirected graph $G = (\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes that correspond to all the data points, and $\mathcal{E}$ denotes the edges that connect pairwise data points with weights. We consider the problem of linearly mapping the graph G to a line. In the process of mapping, we try to preserve the local structure of the data set and simultaneously retain the between-locality information contained in the data set. We refer to such linear dimensionality reduction method as LPMIP. Generally speaking, LPMIP aims to find an optimal transformation that maintains the intrinsic geometry of the data set. Let $y_1,\ldots,y_n$ be the projected 1-D data set. The objective function of LPMIP is proposed as follows:

$$\begin{aligned}\max\ J(\mathbf{v}) &= \alpha\sum_{i=1}^{n}\sum_{j\notin O(i;\varepsilon)}(y_i - y_j)^2\mathbf{W}(i,j)\\ &\quad - (1-\alpha)\sum_{i=1}^{n}\sum_{j\in O(i;\varepsilon)}(y_i - y_j)^2\mathbf{W}(i,j)\\ &= \alpha J_b(\mathbf{v}) - (1-\alpha)J_w(\mathbf{v})\end{aligned} \tag{7}$$

where $J_b(\mathbf{v}) = \sum_{i=1}^{n}\sum_{j\notin O(i;\varepsilon)}(y_i - y_j)^2\mathbf{W}(i,j)$, $J_w(\mathbf{v}) = \sum_{i=1}^{n}\sum_{j\in O(i;\varepsilon)}(y_i - y_j)^2\mathbf{W}(i,j)$, $\alpha \in [0,1]$ is a regularized parameter, and $\mathbf{W}(i,j)$ is the weight imposed on the edge that connects the data points $\mathbf{x}_i$ and $\mathbf{x}_j$. As in the Laplacian eigenmaps [4] and LPP [13], the weight could be realized by the heat kernel (Gaussian kernel) that is defined as

$$\mathbf{W}(i,j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma\right), \qquad \text{for } i,j = 1,\ldots,n \tag{8}$$

where σ is a positive parameter, and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^p$. Clearly, $\mathbf{W}(i,j)$ falls into the interval $[0,1]$ for any pair of $\mathbf{x}_i$ and $\mathbf{x}_j$, and parameter σ. Also, $\mathbf{W}(i,j)$ is monotonously decreasing with respect to the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Let $O(\mathbf{x}_i;\varepsilon)$ denote the neighborhood points of $\mathbf{x}_i$ with a positive and appropriately small radius ε, i.e., $O(\mathbf{x}_i;\varepsilon) = \{\mathbf{x}\,|\,\|\mathbf{x} - \mathbf{x}_i\|^2 < \varepsilon\}$. Then, in (7), we let $O(i;\varepsilon)$ denote the indexes (subscripts) of the data points falling into the ε-neighborhood of $\mathbf{x}_i$, i.e., $O(i;\varepsilon) = \{j|\mathbf{x}_j \in O(\mathbf{x}_i;\varepsilon)\}$. Maximizing the objective function $J(\mathbf{v})$ in (7) is equal to maximizing the first term $\alpha J_b(\mathbf{v})$ and minimizing the second term $(1-\alpha)J_w(\mathbf{v})$ simultaneously. As will be discussed, $J_b(\mathbf{v})$ and $J_w(\mathbf{v})$, respectively, reflect the between-locality and the within-locality structure of the data set.

A. Minimize $J_w(\mathbf{v})$

First, we consider $J_w(\mathbf{v})$. For any data point $\mathbf{x}_i$, $\sum_{j\in O(i;\varepsilon)}(y_i - y_j)^2\mathbf{W}(i,j)$ computes the weighted sum of all squared pairwise Euclidean distances between $\mathbf{x}_i$ and the data points $\mathbf{x}_j$ that are within the ε-neighborhood of $\mathbf{x}_i$. The weighting of heat kernel is designed to indicate the *closeness degree* of $\mathbf{x}_j$ with $\mathbf{x}_i$ within the ε-neighborhood. The closer $\mathbf{x}_j$ and $\mathbf{x}_i$ are, the larger the *closeness degree* will be. $J_w(\mathbf{v})$ reveals the *within-locality* of the data structure, since the data points beyond the ε-neighborhood are not taken into account. Minimizing $J_w(\mathbf{v})$ is an attempt to make samples within the ε-neighborhood as compact as possible, thus preserving the local characterization of the samples. Besides, the nearer the data points are, the nearer their projected images are. Otherwise, a heavy penalty will be incurred due to the weighting of heat kernel.

Denote the adjacency matrix by $\mathbf{A}$, whose entries are defined as

$$\mathbf{A}(i,j) = \begin{cases}\mathbf{W}(i,j), & \mathbf{x}_j \in O(\mathbf{x}_i;\varepsilon)\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

so $J_w(\mathbf{v})$ could be rewritten as

$$\begin{aligned}J_w(\mathbf{v}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2\mathbf{A}(i,j)\\ &= \mathbf{v}^T\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}(i,j)\right)\mathbf{v}\\ &= \mathbf{v}^T\mathbf{H}_w\mathbf{v}\end{aligned} \tag{10}$$

where, on account of the symmetry of $\mathbf{A}$

$$\begin{aligned}\mathbf{H}_w &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}(i,j)\\ &= 2\mathbf{XLX}^T.\end{aligned} \tag{11}$$

We call $\mathbf{H}_w$ the *weighted within-locality matrix*. Clearly, it is symmetric and positive semidefinite.

We see that the within-locality $J_w(\mathbf{v})$ is actually the same with the objective function of LPP. As in LPP, minimizing $J_w(\mathbf{v})$ could produce a subspace that preserves the local structure of the data set. That is, if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close, then their projections y_i and y_j are close as well. However, on the other hand, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are far away, then LPP cannot necessarily guarantee that their projections y_i and y_j are far away as well. This implies that LPP may happen to project mutually distant data points that belong to different manifolds into close images, which obviously is unwanted from the perspective of classification. We, therefore, introduce $J_w(\mathbf{v})$ to serve as this purpose, by ensuring that two mutually distant samples are projected as apart as possible.

B. Maximize $J_b(\mathbf{v})$

In $J_b(\mathbf{v})$, for any data point $\mathbf{x}_i$, $\sum_{j \notin O(i;\varepsilon)} (y_i - y_j)^2 \mathbf{W}(i,j)$ computes the weighted sum of all squared pairwise Euclidean distances between $\mathbf{x}_i$ and the data points $\mathbf{x}_j$ that are beyond the ε-neighborhood of $\mathbf{x}_i$. $J_b(\mathbf{v})$ reveals the *between-locality* of the data structure, since the data points within the ε-neighborhood are not taken into account. Maximizing it is an attempt to make samples beyond the ε-neighborhood as dispersive as possible, thus preserving the interlocality characterization of the samples. The weighting of heat kernel introduced in $J_b(\mathbf{v})$ also plays an important role. Samples that are beyond the ε-neighborhood are not processed equally. The nearer the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are (satisfying $\|\mathbf{x}_i - \mathbf{x}_j\|^2 > \varepsilon$), the heavier the weight that is put between them. To avoid incurring large penalty, it is thus expected that the projected images of close points are far away. As a result, it is easy to classify different manifolds, since even the originally close manifolds are projected far away. On the other hand, a light weight is put between mutually distant samples, which is useful in deemphasizing atypical samples and, therefore, makes LPMIP robust to outliers.

Likewise, denote the dissimilarity matrix by $\bar{\mathbf{A}}$, whose entries are defined as

$$\bar{\mathbf{A}}(i,j) = \begin{cases} \mathbf{W}(i,j), & \mathbf{x}_j \notin O(\mathbf{x}_i;\varepsilon) \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

Clearly, $\bar{\mathbf{A}}$ is symmetric and $\bar{\mathbf{A}} + \mathbf{A} = \mathbf{W}$. Therefore, it follows that

$$\begin{aligned} J_b(\mathbf{v}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 \bar{\mathbf{A}}(i,j) \\ &= \mathbf{v}^T \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \bar{\mathbf{A}}(i,j) \right) \mathbf{v} \\ &= \mathbf{v}^T \mathbf{H}_b \mathbf{v} \end{aligned} \tag{13}$$

where

$$\mathbf{H}_b = \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \bar{\mathbf{A}}(i,j) \tag{14}$$

which we refer to as the *weighted between-locality matrix*. Similarly, it could be shown that $\mathbf{H}_b = 2\mathbf{X}\bar{\mathbf{L}}\mathbf{X}^T$, where the Laplacian matrix $\bar{\mathbf{L}} = \bar{\mathbf{D}} - \bar{\mathbf{A}}$, and $\bar{\mathbf{D}}$ is a diagonal matrix whose entries on diagonal are column sum of $\bar{\mathbf{A}}$, i.e., $\bar{\mathbf{D}}(i,i) = \sum_{j=1}^{n} \bar{\mathbf{A}}(i,j)$. Also, $\mathbf{H}_b$ is symmetric and positive semidefinite. Alternatively, by using the equation $\bar{\mathbf{A}} + \mathbf{A} = \mathbf{W}$, $J_b(\mathbf{v})$ in (13) could be computed as

$$\begin{aligned} J_b(\mathbf{v}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 \mathbf{W}(i,j) - \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 \mathbf{A}(i,j) \\ &= \mathbf{v}^T \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}(i,j) - \mathbf{H}_w \right) \mathbf{v} \\ &= \mathbf{v}^T (\mathbf{H}_t - \mathbf{H}_w) \mathbf{v} \end{aligned} \tag{15}$$

where

$$\mathbf{H}_t = \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}(i,j). \tag{16}$$

Also, we have that $\mathbf{H}_t = 2\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T$, where the Laplacian matrix $\tilde{\mathbf{L}} = \tilde{\mathbf{D}} - \mathbf{W}$, and $\tilde{\mathbf{D}}$ is a diagonal matrix with $\tilde{\mathbf{D}}(i,i) = \sum_{j=1}^{n} \mathbf{W}(i,j)$. Clearly, $\mathbf{v}^T\mathbf{H}_t\mathbf{v}$, in fact, is the weighted sum of all squared Euclidean distances between all pairwise data points, in which if the weights $\mathbf{W}(i,j)$ are dropped, then PCA is actually recovered, so $\mathbf{v}^T\mathbf{H}_t\mathbf{v}$ reflects the global scatter of the data points. We thus call $\mathbf{H}_t$ as the *weighted globality matrix*. Combining (13) and (15), we have that $\mathbf{H}_t = \mathbf{H}_w + \mathbf{H}_b$, which says that the between-locality scatter and the within-locality scatter information completely make up the global scatter information.

C. Formulation of LPMIP

Substituting $J_b(\mathbf{v})$ and $J_w(\mathbf{v})$ into the objective function (7), we have

$$\begin{aligned} J(\mathbf{v}) &= \alpha J_b(\mathbf{v}) - (1-\alpha) J_w(\mathbf{v}) \\ &= \mathbf{v}^T \left(\alpha \mathbf{H}_b - (1-\alpha)\mathbf{H}_w \right) \mathbf{v} \\ &= 2\mathbf{v}^T \mathbf{X} \left(\alpha \bar{\mathbf{L}} - (1-\alpha)\mathbf{L} \right) \mathbf{X}^T \mathbf{v}. \end{aligned} \tag{17}$$

Maximizing $J(\mathbf{v})$ is to find projections such that the close data points are attracted closer (minimizing the within-locality scatter) while mutually distant data points are simultaneously pulled farther away (maximizing the between-locality scatter). The parameter α is to control the tradeoff between $J_b(\mathbf{v})$ and $J_w(\mathbf{v})$. We can imagine α as a cursor moving in $[0,1]$. The larger the α is, the more favorable the between-locality is to win. Usually, α is determined by using cross-validation strategy. Since $\mathbf{H}_b = \mathbf{H}_t - \mathbf{H}_w$, $J(\mathbf{v})$ can be computed as

$$\begin{aligned} J(\mathbf{v}) &= \mathbf{v}^T (\alpha \mathbf{H}_t - \mathbf{H}_w) \mathbf{v} \\ &= 2\mathbf{v}^T \mathbf{X} (\alpha \tilde{\mathbf{L}} - \mathbf{L}) \mathbf{X}^T \mathbf{v} \end{aligned} \tag{18}$$

so maximizing $J(\mathbf{v})$ could be equivalently interpreted as minimizing the within-locality scatter while simultaneously maximizing the global scatter, again regularized by α. The objective function in (17) or (18) is formally similar to that of MMC in the sense that they are both the difference between the two quantities. However, their meanings are different. Note that all the quantities in (17) and (18) can be constructed with or without using the class labels of observed samples, while $\mathbf{S}_b$ and $\mathbf{S}_w$ in MMC have to use the class labels. This means LPMIP could be unsupervised or supervised, while MMC is supervised.

If we write

$$\mathbf{M} = \alpha\tilde{\mathbf{L}} - \mathbf{L} \tag{19}$$

the criterion in (18) can be maximized by solving

$$\arg\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{v} \quad \text{subject to} \quad \mathbf{v}^T\mathbf{v} - 1 = 0. \tag{20}$$

Therefore, we introduce the Lagrangian multiplier technique

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{v} - \lambda(\mathbf{v}^T\mathbf{v} - 1) \tag{21}$$

with the multiplier λ. Differentiating $\mathcal{L}$ with respect to $\mathbf{v}$ and λ and then setting to zero yield $\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{v} = \lambda\mathbf{v}$, which means that $\mathbf{v}$ is the unitary eigenvector of $\mathbf{X}\mathbf{M}\mathbf{X}^T$, associated with the eigenvalue λ, and λ is just the value of the objective function. Thus, the transformation vector $\mathbf{v}$ maximizing the objective function is given by the largest eigenvector of $\mathbf{X}\mathbf{M}\mathbf{X}^T$, corresponding to the largest eigenvalue. More generally, the q columns of the transformation matrix $\mathbf{V}$ are the first q largest eigenvectors of $\mathbf{X}\mathbf{M}\mathbf{X}^T$. Note that the matrix $\mathbf{X}\mathbf{M}\mathbf{X}^T$ is symmetric (but not necessarily positive semidefinite). Thus, the matrix obtained $\mathbf{V}$ has the orthogonal columns.

D. *Efficient Algorithm for LPMIP via QR-Decomposition*

In real-world applications of such image, gene expression, and web document recognition, the dimension p of the vector samples is usually large, so performing LPMIP by directly solving the eigenvectors of the $p \times p$ matrix $\mathbf{X}\mathbf{M}\mathbf{X}^T$ is still computationally intensive. Besides, there is still the attendant problem of numerical accuracy when diagonalizing large matrix directly [31]. To reduce the computational demand, in this section, we present an efficient and stable algorithm for performing LPMIP via QR-decomposition, which we refer to as LPMIP/QR.

From matrix computation knowledge, we know that the data matrix $\mathbf{X}$ could be QR-decomposed as $\mathbf{X} = \mathbf{Q}\mathbf{R}$ by using the incomplete Cholesky decomposition [1], where $\mathbf{Q} \in \mathbb{R}^{p\times t}$ has the orthonormal columns, $\mathbf{R} \in \mathbb{R}^{t\times n}$ is an upper triangular matrix, and $t = \mathrm{rank}(\mathbf{X})$ is the rank of $\mathbf{X}$. If we suppose, for the time being, that the optimal transformation matrix $\mathbf{V}$ can be expressed as $\mathbf{V} = \mathbf{Q}\mathbf{T}$ for some $\mathbf{T} \in \mathbb{R}^{t\times q}$ satisfying $\mathbf{T}^T\mathbf{T} = I_q$ (since $\mathbf{V}$ has the orthonormal columns), which implies that $\mathbf{V}$ can be spanned by $\mathbf{Q}$, or equivalently $\mathbf{X}$, then the original problem of computing $\mathbf{V}$ is converted into computing $\mathbf{T}$ such that

$$\mathbf{T} = \arg\max_{\mathbf{T}^T\mathbf{T}=I_q} \mathrm{tr}\left(\mathbf{T}^T(\mathbf{Q}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{Q})\mathbf{T}\right) \tag{22}$$

where I_q is the q-dimensional identity matrix and "tr" denotes the trace operator. On the other hand, on account of $\mathbf{Q}^T\mathbf{Q} = I_t$, we have

$$\mathbf{Q}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{Q} = \mathbf{Q}^T\mathbf{Q}\mathbf{R}\mathbf{M}\mathbf{R}^T\mathbf{Q}^T\mathbf{Q} = \mathbf{R}\mathbf{M}\mathbf{R}^T. \tag{23}$$

Note that $\mathbf{R}\mathbf{M}\mathbf{R}^T$ is of size $t \times t$, which has much smaller size than that of $\mathbf{X}\mathbf{M}\mathbf{X}^T$, since usually $t \ll p$. The optimization problem in (22) can be solved using the similar method of Lagrangian multiplier. That is, we compute the optimal $\mathbf{T}$ by applying the eigendecomposition to $\mathbf{R}\mathbf{M}\mathbf{R}^T$. The optimal solutions (i.e., columns of $\mathbf{T}$) are the q leading eigenvectors of $\mathbf{R}\mathbf{M}\mathbf{R}^T$, associated with the q largest eigenvalues. Consequently, the optimal solution of $\mathbf{V}$ is given by

$$\mathbf{V} = \mathbf{Q}\mathbf{T}. \tag{24}$$

The equivalence between LPMIP/QR and LPMIP is theoretically established as formally stated in the following.

Theorem 1: Let $\mathbf{X} = \mathbf{Q}\mathbf{R}$ be the QR-decomposition of $\mathbf{X}$, and $\mathbf{T}$ be the matrix whose columns are the unitary eigenvectors of $\mathbf{R}\mathbf{M}\mathbf{R}^T$. Then, the columns of the optimal transformation matrix $\mathbf{V} = \mathbf{Q}\mathbf{T}$ obtained from the LPMIP/QR algorithm are just the unitary eigenvectors of $\mathbf{X}\mathbf{M}\mathbf{X}^T$ with the same eigenvalues.

The proof of Theorem 1 is given in the Appendix. This theorem shows that LPMIP/QR is computationally equivalent to the standard LPMIP. The LPMIP/QR algorithm, however, provides a computationally efficient and stable way for performing LPMIP. The QR-decomposition for computing $\mathbf{R}$ is of time complexity $O(t^2n)$, and solving the eigenvalue problem of $\mathbf{R}\mathbf{M}\mathbf{R}^T$ has time complexity $O(t^3)$. By contrast, if we perform the standard LPMIP, the time complexity of directly diagonalizing $\mathbf{X}\mathbf{M}\mathbf{X}^T$ is $O(p^3)$. The computational complexity of LPMIP/QR compares favorably with that of LPMIP in the situation involving high-dimensional data set.

E. *Learning LPMIP/QR for Discrimination*

In the previous description, ε-neighborhood is used to characterize the within-locality and the between locality. It is intuitive in geometry but unsuitable in implementation, since it is usually difficult to determine an appropriate value of ε in practice. We thus instead adopt the k-nearest-neighbors to characterize the locality. Then, the adjacency matrix $\mathbf{A}$ could be redefined, using the k-nearest-neighbors method, as

$$\mathbf{A}(i,j) = \begin{cases} \mathbf{W}(i,j), & \text{if } \mathbf{x}_i \text{ is among the } k\text{-nearest-neighbors} \\ & \text{of } \mathbf{x}_j \text{ or } \mathbf{x}_j \text{ is among the} \\ & k\text{-nearest-neighbors of } \mathbf{x}_i \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Particularly, when choosing the parameter $\sigma = +\infty$, then $\mathbf{W}(i,j) = 1$ always holds. In this case, the elements in the adjacency matrix $\mathbf{A}$ are 1 or 0, and we still have $\bar{\mathbf{A}} = \mathbf{W} - \mathbf{A}$. Now, the algorithmic procedure of LPMIP is formally summarized as follows.

1) **Compute the matrix $\mathbf{W}$**. Given the training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, compute $\mathbf{W}$ according to (8), or simply let $\mathbf{W} = \mathbf{1}_n\mathbf{1}_n^T$, where $\mathbf{1}_n$ is an n-dimensional vector with all entries being one.
2) **Construct nearest-neighbors graph**. Construct a graph G having n nodes, where the ith node corresponds to the data point $\mathbf{x}_i$. An edge is placed between nodes i and j if $\mathbf{x}_i$ is among the k-nearest-neighbors of $\mathbf{x}_j$, or $\mathbf{x}_j$ is among the k-nearest-neighbors of $\mathbf{x}_i$.
3) **Compute the matrix $\mathbf{M}$**. If nodes i and j are linked, then set $\mathbf{A}(i,j) = \mathbf{W}(i,j)$. Otherwise, let $\mathbf{A}(i,j) = 0$. As a result, we get an $n \times n$ sparse symmetric matrix $\mathbf{A}$. Using the matrices $\mathbf{W}$ and $\mathbf{A}$, compute the matrix $\mathbf{M}$ according to (19).

4) **QR-decomposition**. Decompose the data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ using the incomplete Cholesky decomposition technique as $\mathbf{X} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{p \times t}$ has orthonormal columns, the upper triangular matrix $\mathbf{R} \in \mathbb{R}^{t \times n}$ has full row rank, and $t = \mathrm{rank}(\mathbf{X})$.
5) **Eigenvalue decomposition**. Solve the eigenvalue problem $\mathbf{RMR}^T\mathbf{t} = \lambda\mathbf{t}$. Let $\lambda_1 \geq \cdots \geq \lambda_q$ be the q largest eigenvalues of $\mathbf{RMR}^T$ and $\mathbf{t}_1, \ldots, \mathbf{t}_q$ be the associated orthonormal eigenvectors.
6) **Compute projection matrix**. The optimal projection matrix is given by $\mathbf{V} = \mathbf{QT}$, where $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_q]$.
7) **Obtain low-dimensional embeddings**. Using the projection matrix $\mathbf{V}$, we could carry out the following linear projection for the training samples $\mathbf{X}$:

$$(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \mathbf{V}^T\mathbf{X} = (\mathbf{QT})^T\mathbf{QR} = \mathbf{T}^T\mathbf{R}. \tag{26}$$

For a testing point $\mathbf{x}$, its image in the lower dimensional space is given by

$$\mathbf{x} \mapsto \mathbf{y} = \mathbf{V}^T\mathbf{x}. \tag{27}$$

Now, let $\mathbf{X}_{\text{test}}$ be the testing data matrix. By using the columns of $\mathbf{Q}$ as bases, $\mathbf{X}_{\text{test}}$ could be represented as $\mathbf{X}_{\text{test}} = \mathbf{QR}_{\text{test}}$. Likewise, the low-dimensional embeddings of the testing data are $\mathbf{T}^T\mathbf{R}_{\text{test}}$. Interestingly, it could be seen that the basis matrix $\mathbf{Q}$ needs, in fact, not to be computed in implementation.

As could be seen, the implementation of the LPMIP/QR algorithm is fairly straightforward. The algorithm does not involve any inverse matrix, and thus completely avoids the SSS problem. It is efficient and stable.

IV. Theoretical Analysis of LPMIP, (Weighted) PCA, MMC, LDA, and LPP

In this section, we will show the theoretical relationship between LPMIP and classical linear projection methods: (weighted) PCA, MMC, LDA, and LPP. Thus, a deeply theoretical perspective of LPMIP is revealed.

A. Relation With (Weighted) PCA

It is worthwhile to point out that (weighted) PCA is a particular example of LPMIP. Specifically, we have the following theorem.

Theorem 2: In the LPMIP's objective function (18), if we let ε (or k) be sufficiently small such that there is no neighbor for each data point, then LPMIP is reduced to the weighted PCA [19]. If we further let the weighting matrix $\mathbf{W} = \mathbf{1}_n\mathbf{1}_n^T$, then the conventional PCA is recovered.

It suffices to show that, under the condition of Theorem 2, the adjacency matrix $\mathbf{A}$ is a diagonal matrix, and the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A} = \mathbf{0}$, so in (18), the objective function $J(\mathbf{v}) \propto \sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^2 \mathbf{W}(i,j)$, which exactly is the weighted PCA [19]. Besides, if we let $\mathbf{W}(i,j) = 1$, then $J(\mathbf{v}) \propto \sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^2 \propto \sum_{i=1}^n (y_i - \bar{y})^2$, which is the conventional PCA. Theorem 2 demonstrates that PCA seeks to retain the global structure to the maximum extent, and totally ignores the within-locality structure of the data set, since it does not consider any neighborhood for each data point. PCA could be viewed as a limiting case of LPMIP when taking infinitesimal within-locality tuned by ε (or k).

In [14], PCA is interpreted under the LPP framework. The explanation, however, is not sound. First, LPP, which is a minimization problem, does not appropriately explain why PCA maximizes the objective function. Second, the projection axes obtained by LPP are not necessarily orthogonal while PCA always has orthogonal axes. Finally, to explain PCA, the neighborhood measure ε (or k) in LPP is taken to be infinite. This does not agree with the meaning of ε (or k), which is originally assumed to be sufficiently small. Compared with LPP, LPMIP has a more straightforward interpretation for PCA. Simply eliminating the locality characterization in LPMIP will result in (weighted) PCA, which is completely consistent with the global property of PCA.

B. Relations With MMC and LDA

It will be seen that MMC and LDA are also special cases of LPMIP, as stated in the following theorem.

Theorem 3: Suppose that the n training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ belong to c classes, and each class has the same number of samples, say n_0. Let $\mathbf{W} = \mathbf{1}_n\mathbf{1}_n^T$ and $\alpha = n_0/2n$. In the ideal clustering case that each local neighborhood contains exactly the training samples belonging to the same class, the LPMIP turns out to be conventional MMC. If we further require the optimal direction $\mathbf{v}$ to satisfy $\mathbf{v}^T\mathbf{H}_w\mathbf{v} = 2n_0n$ instead of $\mathbf{v}^T\mathbf{v} = 1$, then LPMIP becomes the conventional LDA.

In the conditions of Theorem 3, the adjacency matrix $\mathbf{A}$ is

$$\mathbf{A}(i,j) = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class} \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

With this adjacency matrix, the entire data set is divided into c local neighborhoods, each of which corresponds to a cluster of n_0 samples of one class, where $n_0 = n/c$. Let $\mathbf{X}_j = [\mathbf{x}_1^j, \ldots, \mathbf{x}_{n_0}^j]$ be the samples from the jth class. Without loss of generality, we assume that in the total data matrix $\mathbf{X}$, the samples are arranged from the first class to the cth class, i.e., $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_c]$. Since the adjacency matrix $\mathbf{A}$ is a block diagonal matrix with each block being $\mathbf{1}_{n_0}\mathbf{1}_{n_0}^T$, i.e., $\mathbf{A} = \mathrm{diag}(\underbrace{\mathbf{1}_{n_0}\mathbf{1}_{n_0}^T, \ldots, \mathbf{1}_{n_0}\mathbf{1}_{n_0}^T}_{c})$, the weighted within-locality matrix

$$\begin{aligned}\mathbf{H}_w &= 2\mathbf{X}\left(n_0 I_n - \mathrm{diag}\left(\mathbf{1}_{n_0}\mathbf{1}_{n_0}^T, \ldots, \mathbf{1}_{n_0}\mathbf{1}_{n_0}^T\right)\right)\mathbf{X}^T \\ &= 2n_0 \sum_{j=1}^{c} \mathbf{X}_j \left(I_{n_0} - \mathbf{1}_{n_0}\mathbf{1}_{n_0}^T/n_0\right)\mathbf{X}_j^T \\ &= 2n_0 \sum_{j=1}^{c}\sum_{i=1}^{n_0} \left(\mathbf{x}_i^j - u_j\right)\left(\mathbf{x}_i^j - u_j\right)^T \\ &= 2n_0 n \mathbf{S}_w \end{aligned} \tag{29}$$

where $\mathbf{S}_w$ is the within-class scatter matrix. Likewise, the weighted globality matrix

$$\begin{aligned}\mathbf{H}_t &= 2n\mathbf{X}\left(I_n - \mathbf{1}_n\mathbf{1}_n^T/n\right)\mathbf{X} \\ &= 2n\sum_{i=1}^{n}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^T \\ &= 2n^2\mathbf{S}_t \end{aligned} \tag{30}$$

where $\mathbf{S}_t = (1/n)\sum_{i=1}^{n}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^T$ is the total scatter matrix. Consequently, in (18), we have

$$\begin{aligned}J(\mathbf{v}) &= \mathbf{v}^T(\alpha\mathbf{H}_t - \mathbf{H}_w)\mathbf{v} \\ &= 2\mathbf{v}^T(\alpha n^2\mathbf{S}_t - n_0 n\mathbf{S}_w)\mathbf{v}.\end{aligned} \tag{31}$$

Substituting $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ into (31) reads as

$$\begin{aligned}J(\mathbf{v}) &= 2\mathbf{v}^T\left(\alpha n^2\mathbf{S}_b - (n_0 n - \alpha n^2)\mathbf{S}_w\right)\mathbf{v} \\ &\propto \mathbf{v}^T\left(\mathbf{S}_b - \left(\frac{n_0}{\alpha n} - 1\right)\mathbf{S}_w\right)\mathbf{v}\end{aligned} \tag{32}$$

which is the objective of RMMC (up to a constant) with the parameter $\gamma = (n_0/\alpha n) - 1$, and when $\alpha = n_0/2n$, we have $\gamma = 1$, so MMC is recovered. Theorem 3 shows that MMC could be formulated in the LPMIP framework, in which each local neighborhood covers the n_0 training samples that belong to the same class and all the entries in the weighting matrix $\mathbf{W}$ are one.

If we replace the constraints $\mathbf{v}^T\mathbf{v} = 1$ by $\mathbf{v}^T\mathbf{H}_w\mathbf{v} = 2n_0 n$, then LPMIP will seek the optimal direction that maximizes

$$\begin{aligned}J(\mathbf{v}) &= \alpha\mathbf{v}^T\mathbf{H}_t\mathbf{v} - 2n_0 n \\ &\propto \frac{1}{2n^2}\mathbf{v}^T\mathbf{H}_t\mathbf{v} - \frac{1}{2n_0 n}\mathbf{v}^T\mathbf{H}_w\mathbf{v} \\ &= \mathbf{v}^T\mathbf{S}_b\mathbf{v}\end{aligned} \tag{33}$$

subject to $\mathbf{v}^T\mathbf{S}_w\mathbf{v} = 1$. By using the Lagrangian multiplier technique, we know that $\mathbf{v}$ is the maximum eigenvector of the generalized eigenvalue problem $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$, which is just the LDA projection axis. Therefore, LDA can be induced from the LPMIP perspective by incorporating some constraints. Again, LDA is a supervised projection method while LPMIP is not necessarily supervised. The adoption of using the constraints $\mathbf{v}^T\mathbf{v} = 1$ in LPMIP is that it allows us to avoid calculating the inverse matrix of $\mathbf{S}_w$ and thus effectively circumvents the SSS problem.

C. Relation With LPP

As we have pointed out, LPP pursues the within-locality structure of the data set, while LPMIP considers both the within-locality and the between-locality properties of data structure. This could be formally clarified from the following theorem.

Theorem 4: In the objective function (17) of LPMIP, if we move the regularization parameter α to the minimum value 0, then the objective function of LPP arises, so as in LPP, the optimal projection axes could be obtained by imposing some constraints.

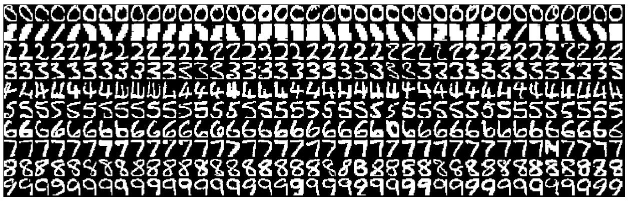


Fig. 1. Examples of handwritten digits from number 0 to 9.

The parameter α tunes the between-locality and the within-locality. When $\alpha = 0$, LPMIP actually is to maximize $-J_w(\mathbf{v})$, which is equivalent to minimizing $J_w(\mathbf{v})$. This is exactly the LPP problem. This minimization problem could be solved in different ways subject to different constraint conditions as in LPP

$$\arg\min_{\mathbf{v}} \mathbf{v}^T\mathbf{XLX}^T\mathbf{v} \tag{34}$$

$$\text{subject to } \mathbf{v}^T\mathbf{XDX}^T\mathbf{v} = 1 \text{ or } \mathbf{v}^T\mathbf{v} = 1 \tag{35}$$

so LPP substantially minimizes the within-locality quantity $J_w(\mathbf{v})$. The projection axes obtained are (conjugately) orthogonal. Although both LPP and LPMIP are linear projection methods and could be carried out in unsupervised or supervised way, their starting points and objective functions are totally different. In short, LPMIP considers not only the within-locality but also the between-locality quantity, and the resultant objective function thus has an immediate purpose for classification. LPP, in contrast, only involves one side of the problem, i.e., the within-locality quantity. In this sense, LPP could be viewed as a special case of LPMIP. The intention of the constraints in LPP is simply to remove an arbitrary scaling factor, which is not a directly oriented classification. Also, if we use the conjugately orthogonal constraints in LPP, then the optimization model of LPP is converted into the Rayleigh quotient problem $\arg\max_{\mathbf{v}}(\mathbf{v}^T\mathbf{XLX}^T\mathbf{v}/\mathbf{v}^T\mathbf{XDX}^T\mathbf{v})$, which suffers from the singularity of $\mathbf{XDX}^T$ in the undersampled situation and yields nonorthogonal basis functions. By contrast, LPMIP effectively circumvents these deficiencies.

V. Experiments

In this section, we carry out several experiments to investigate the performance of the proposed LPMIP method for data visualization, face recognition, character recognition, and facial expression recognition.

A. Data 2-D-Visualization

The experiment is conducted on a digit database,[1] which contains 390 binary images of handwritten digits of ten numeral classes with 39 samples per class. Each image, which has the size 20 × 16 pixels, is represented as a 320-dimensional vector by scanning the digit image row by row. Fig. 1 shows all the examples of the ten digits. As could be seen, the handwritings of

[1] The database is publicly available from http://www.cs.toronto.edu/roweis/data.html

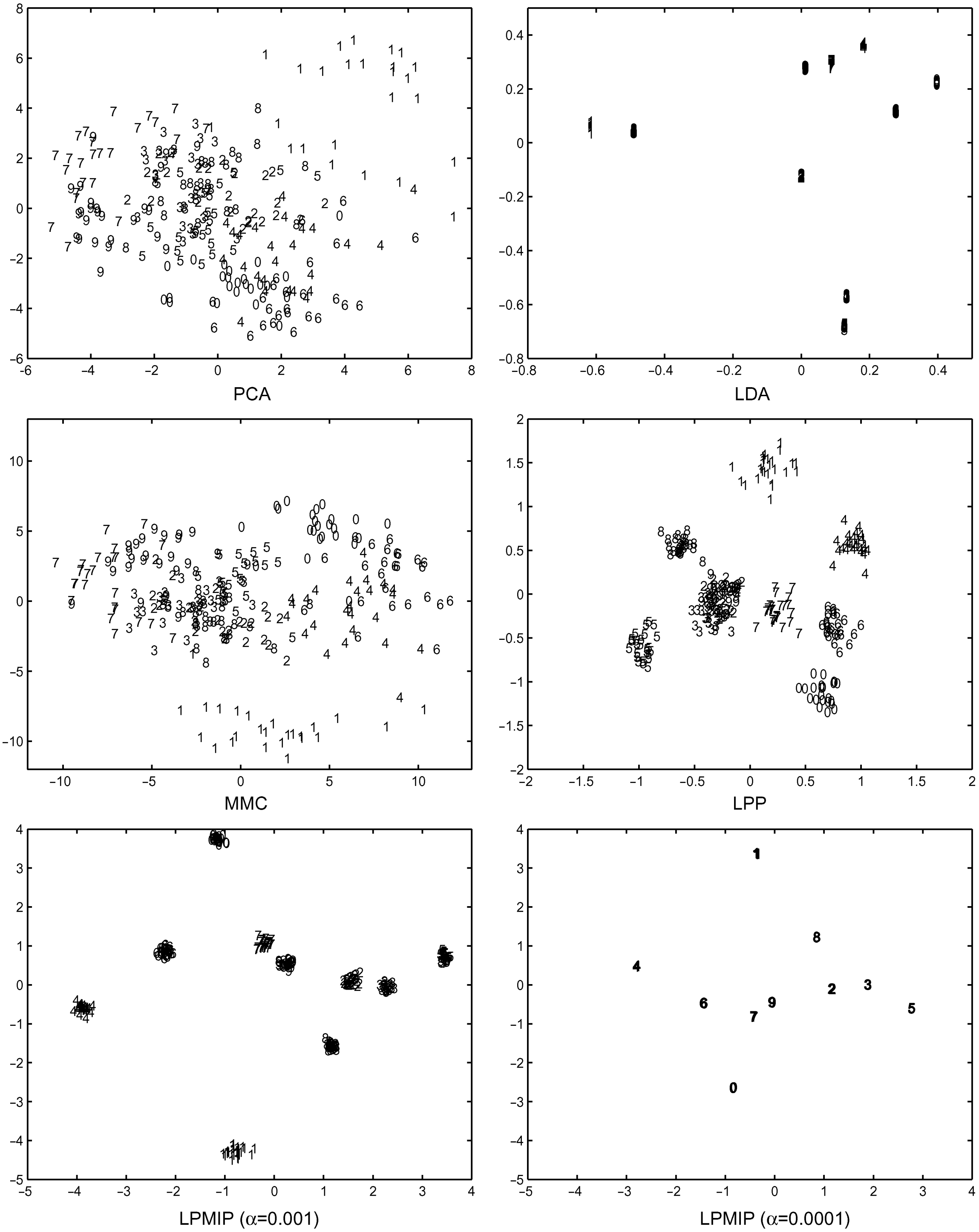


Fig. 2. Handwritten digits from number 0 to 9 are mapped onto a 2-D space by, respectively, using five linear projection techniques: PCA, LDA, MMC, LPP, and LPMIP.

some digits are somewhat illegible. In this experiment, we use 250 samples (25 patterns per digit) to train PCA, LDA, MMC, LPP, and LPMIP, respectively. This is an undersampled situation. Both LDA and LPP involve a preceding PCA stage to avoid the singularity problem. In using LDA, we perform PCA such that the dimension of the resulting samples is $n - c = 250 - 10 = 240$ [3]. These reduced-dimensional samples are then input into LDA, and in using LPP, we keep 98% data energy in the PCA stage as in [14]. The adjacency matrix in LPP and LPMIP is constructed based on the class labels of the samples, and we let the parameter $\sigma = +\infty$. The 250 data points are projected onto a 2-D space by using the five subspace learning methods mentioned previously. The experimental results are illustrated in Fig. 2. As can be seen, the projections of PCA are scattered everywhere since PCA tries to maintain the maximal energy. LDA projects the samples belonging to the same digit class into nearly one single point, since LDA seeks to maximize the Fisher's criterion. The mapping obtained by LDA works well for the training samples, but it may have a poor generalization ability, noting that the projected data points are very close to each other. MMC has a slightly better performance than PCA in this situation. By contrast, LPP yields more meaningful clusters than both PCA and MMC. Clearly, the proposed LPMIP method with $\alpha = 0.001$ achieves the best clustering performance, since the digits of the same class appear relatively compact while different digit classes are relatively far away. By comparing the results of MMC and LPMIP with $\alpha = 0.001$ and $\alpha = 0.0001$, we can visually see the role of α. By Theorem 3, MMC in this experiment in fact is equivalent to LPMIP with $\alpha = 0.05$, so when α varies from 0.05 to 0.001 and to 0.0001, the within-locality is more and more emphasized, and when $\alpha = 0.001$, the within-locality and globality are appropriately balanced. Clearly, LPMIP with $\alpha = 0.0001$, as well as LDA, overlearn the samples of the same class. The former, however, is better than the latter, since different classes are farther away (noting the limits of axes).

B. Face Recognition

In this section, we investigate the performance of the proposed LPMIP algorithm for face recognition on three benchmark databases: the Olivetti Research Laboratory (ORL),[2] UMIST,[3] and color FERET face databases. The ORL database contains 400 images grouped into 40 distinct subjects with ten different images for each. The image were captured at different times, and for some subjects, the images may vary in facial expressions and facial details. All the images were taken against a dark homogeneous background with the tolerance for some side movement of about 20°. The original images are all sized 112 × 92 pixels with 256 gray levels per pixel, which are further downsampled into 28 × 23 pixels in our experiment. The UMIST database contains 20 persons with totally 564 images [12]. There are variations of race, sex, and appearance with different subjects. The size of each image is approximately 220 × 220 pixels, with 256 gray levels per pixel. Precropped versions (with a size of 112 × 92) of the images may be also made available from the database, which are also downsampled into 28 × 23 pixels in experiment. The color FERET database contains a total of 11 338 facial images obtained by photographing 994 subjects [28], [29]. The images are of size 512 × 768 pixels. We select 984 images of 246 subjects who have "fa" images in both the gallery "fa" and prob "dup1" sets and "fb" images in both the "fb" and "dup1" sets. For each individual, the four frontal images (two "fa" images plus two "fb" images) are used in the experiment. There are variations of facial expression, aging, and wearing glasses in this subset. In our experiment, all of these color images are converted into gray ones. Then, the gray images are automatically cropped proportionally based on the distance between eyes and the distance between eyes and mouth. Finally, the cropped images are resized to 72 × 64 pixels.

Some sample images after preprocessing of the three databases are shown in Fig. 3. In our experiment, the grayscale is linearly normalized to be located within $[0, 1]$. To perform the face recognition, we first obtain the face subspaces by dimensionality reduction techniques. Then, facial images are projected onto the face subspaces. Finally, the nearest-neighbor classifier is adopted to identify new facial images, where the Euclidean metric is used as the distance measure.

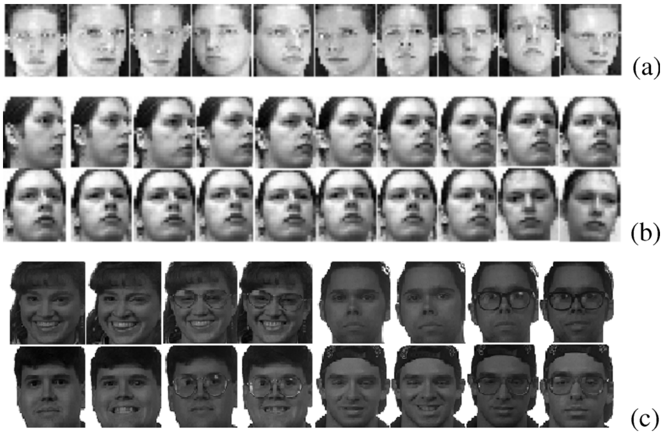


Fig. 3. Face examples from three face databases: (a) ORL, (b) UMIST, and (c) color FERET.

1) ORL Database: We randomly select six images of each individual to construct the training set and use the remainder images of the database to form the testing set. Thus, the numbers of the training samples and testing samples are, respectively, 240 and 160. For each evaluation, 30 rounds of experiments are repeated with random selection of the training data, and the average result is recorded as final recognition accuracy. We test the performance of LPMIP in comparison with Laplacianfaces, and for a comprehensive comparison, we also perform the baseline methods: eigenfaces, Fisherfaces, OLDA, and MMC for face recognition.

In using both the Fisherfaces and Laplacianfaces methods, preprocessing by PCA is needed. In this PCA stage, we retain $n - c$ dimensions before implementing LDA [3], where c is the number of subjects, and keep 98% data information to determine the number of principal components before carrying out LPP [14], and for fair comparisons with Fisherfaces and Laplacianfaces, we also investigate all the PCA dimensions keeping energies of 85%–99% before performing LDA and LPP, which are referred to as PCA + LDA and PCA + LPP, respectively, in our experiments. Further, the RMMC method is also considered. For each method, we report the best result with the optimal reduced dimensions, as well as the PCA dimensions. Noting that in the LDA stage when using the Fisherfaces and PCA + LDA methods, there are at most $c - 1$ nonzero generalized eigenvalues, we take the dimension of the reduced space being at most $c - 1$. We use these settings throughout the experiments.

In using the proposed LPMIP algorithm, we compute the weighting matrix $\mathbf{W}$ by using the heat kernel $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma)$. The parameter σ is set as $\sigma = 2^m\sigma_0$, where σ_0 is the standard deviation of the squared norms of the training samples, and $m \in \{-10, -9, \ldots, 0, \ldots, 9, 10\}$. The regularized parameter α is set as $\alpha = 2^{a/4.5}\lambda_{\max}(\mathbf{RLR}^T)/\lambda_{\max}(\mathbf{R\tilde{L}R}^T)$, where $\lambda_{\max}(\mathbf{A})$ denotes the largest eigenvalue of a matrix $\mathbf{A}$, and $a \in \{-20, -19, \ldots, 0, \ldots, 19, 20\}$. This setting is

[2] http://www.uk.research.att.com/facedatabase.html

[3] http://www.images.ee.umist.ac.uk/danny/database.html

TABLE I

COMPARISON OF MAXIMAL AVERAGE RECOGNITION RATES (IN PERCENT) AS WELL AS THE CORRESPONDING STANDARD DEVIATIONS AND THE OPTIMAL REDUCED DIMENSIONS (FOR PCA + LDA AND PCA + LPP, THE FIRST NUMBER IN PARENTHESES IS THE OPTIMAL PERCENTAGE OF ENERGY KEEPING IN THE PCA PHASE) ON THE ORL DATABASE

Methods	Eigenfaces	Fisherfaces	PCA+LDA	OLDA	MMC	RMMC	Laplacianfaces	PCA+LPP	LPMIP
						$(r=9)^a$	$(m=-1)$	$(m=-1)$	$\begin{pmatrix} m{=}0 \\ a{=}4 \end{pmatrix}$
Recog.	90.4	93.6	94.1	94.2	93.1	94.0	94.3	94.7	**97.9**
Std.	1.9	3.4	2.1	2.1	1.8	1.9	1.7	1.5	1.9
Dims	65	39	(95, 35)	39	40	40	35	(93, 30)	20

[a] Here, the regularized parameter γ is set as $\gamma = 2^{r/4.5}\lambda_{\max}(\mathbf{S}_b)/\lambda_{\max}(\mathbf{S}_w)$.

obviously similar to that of γ in [25]. Since α essentially is to balance the energy variations between, for example, $\mathbf{X\tilde{L}X}^T$ and $\mathbf{XLX}^T$ from (18), or equivalently, $\mathbf{R\hat{L}R}^T$ and $\mathbf{RLR}^T$, and the largest eigenvalue of respective matrix is a reasonable, but simple, measurement of the energy variation, we thus consider the ratio of the two largest eigenvalues, multiplied by a constant, as an appropriate estimate of α. Noting that $\mathbf{R\tilde{L}R}^T - \mathbf{RLR}^T$ is a positive–semidefinite matrix, it follows that $0 \leq \lambda_{\max}(\mathbf{RLR}^T)/\lambda_{\max}(\mathbf{R\tilde{L}R}^T) \leq 1$. In all the experiments, the values m and a are experimentally chosen to obtain the maximal recognition accuracy. Based on $\mathbf{W}$, the adjacency matrix $\mathbf{A}$ is constructed via five nearest-neighbors. Since there are six samples per class, we expect that each sample and the remaining five samples of the same class form nearest-neighborhoods. For fair comparison, the LPP uses the same adjacency matrix $\mathbf{A}$ with LPMIP, where only the optimal value m may not necessarily be identical.

In general, the performances of these methods mentioned previously vary with the reduced dimensions and some relevant parameters involved. Table I reports the maximal average recognition rates, as well as the corresponding standard deviations and optimal reduced dimensions, across 30 runs of each configuration with the values of parameter γ in RMMC, m in LPP and LPMIP, and a in LPMIP. Fig. 4 illustrates the recognition rates versus the variation of reduced dimensions. The recognition rates of LPMIP versus the values of parameter α are shown in Fig. 5. From Table I, it can be found that the LPMIP method outperforms all the other methods with fewer features. The eigenfaces method gives relatively poor recognition accuracy, while the rest of the methods have comparative performances. Note that PCA+LDA and PCA+LPP can improve the performances of Fisherfaces and Laplacianfaces, respectively. This improvement may be due to the extraction of principal features and denoising of PCA. LPMIP and LPP have different optimal m in constructing the adjacency matrix. Fig. 4 indicates that increased dimension will improve the recognition rate until the dimension reaches some optimal value. From Fig. 5, we observe that too small or too large α will not lead to good recognition accuracy in this experiment. It is worthwhile to note that, in the case $\alpha = 0$ where the global information is completely ignored, LPMIP works rather inefficiently. Recall that α compromises between the global and local structures of the data set. However, this in turn will depend on the data set at hand.

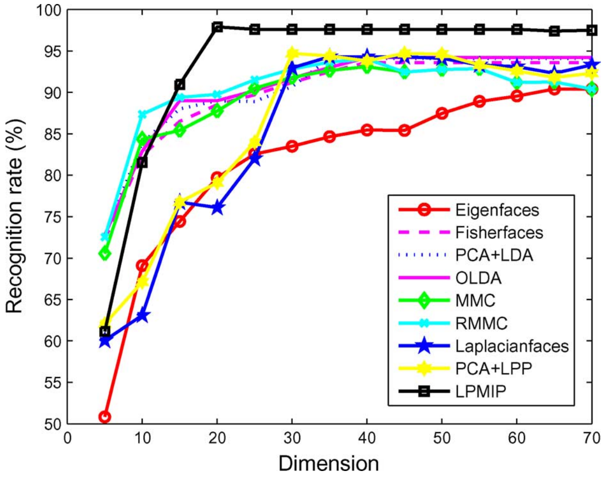


Fig. 4. Maximal average recognition rates of various linear projection methods versus reduced dimensions on the ORL database.

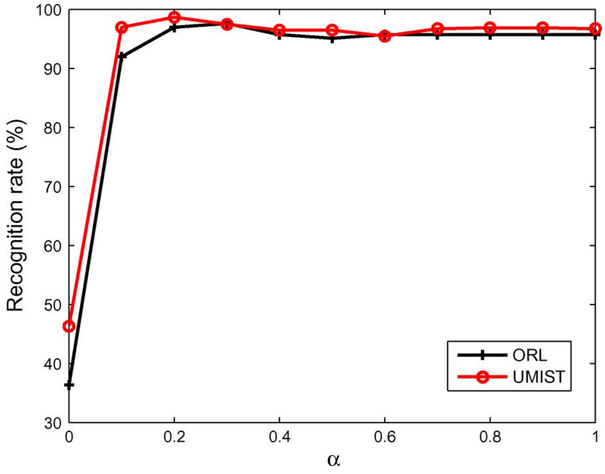


Fig. 5. Maximal average recognition rates of LPMIP versus the parameter α on the ORL and UMIST databases, respectively.

2) UMIST Database: The training set is a randomly selected subset with ten images per individual, and the remaining images of the database are used as the testing set. Likewise, we average the results over 30 random splits of the database as a final recognition accuracy. The experimental configurations are set as the same with those applied in the ORL database, except that the adjacency matrix is based on nine nearest-neighbors. Table II lists the experimental results of maximal average recognition rates with the standard deviations and optimal dimensions, as well as the parameters γ, m, and a. The recognition rates versus the variation of reduced dimensions by using various methods are shown in Fig. 6. The recognition rate curve of LPMIP with

TABLE II
COMPARISON OF MAXIMAL AVERAGE RECOGNITION RATES (IN PERCENT) AS WELL AS THE CORRESPONDING STANDARD DEVIATIONS AND THE OPTIMAL REDUCED DIMENSIONS ON THE UMIST DATABASE

Methods	Eigenfaces	Fisherfaces	PCA+LDA	OLDA	MMC	RMMC	Laplacianfaces	PCA+LPP	LPMIP
						$(r = 12)$	$(m = -3)$	$(m = -3)$	$\binom{m=0}{a=11}$
Recog.	91.8	93.0	94.4	93.5	93.8	95.0	94.3	95.1	**98.7**
Std.	5.2	7.7	7.0	7.4	5.3	4.9	4.8	4.9	3.1
Dims	32	19	(92, 19)	19	22	24	22	(99, 28)	30

TABLE III
COMPARISON OF MAXIMAL AVERAGE RECOGNITION RATES (IN PERCENT) AS WELL AS THE CORRESPONDING STANDARD DEVIATIONS AND THE OPTIMAL REDUCED-DIMENSIONS ON A SUBSET OF THE COLOR FERET DATABASE

Methods	Eigenfaces	Fisherfaces	PCA+LDA	OLDA	MMC	RMMC	Laplacianfaces	PCA+LPP	LPMIP
						$(r = 16)$	$(m = -1)$	$(m = -1)$	$\binom{m=0}{a=19}$
Recog.	77.6	65.9	82.1	68.3	80.1	80.5	67.5	79.7	**85.4**
Std.	2.3	1.4	3.2	3.9	5.5	3.1	4.9	2.4	1.8
Dims	30	60	(96, 35)	60	50	25	45	(94, 50)	35

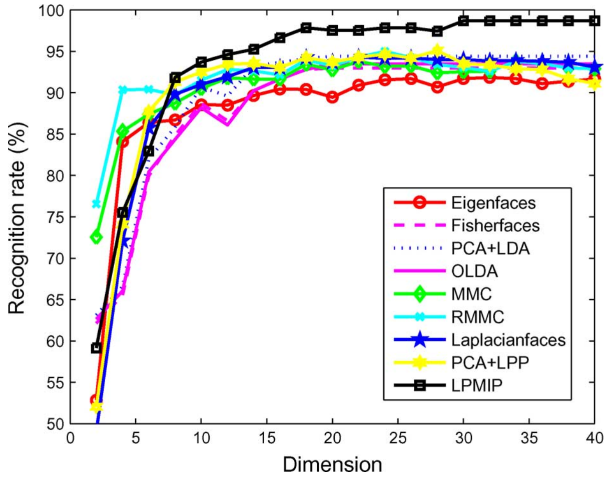


Fig. 6. Average recognition rates of various linear projection methods versus reduced dimensions on the UMIST database.

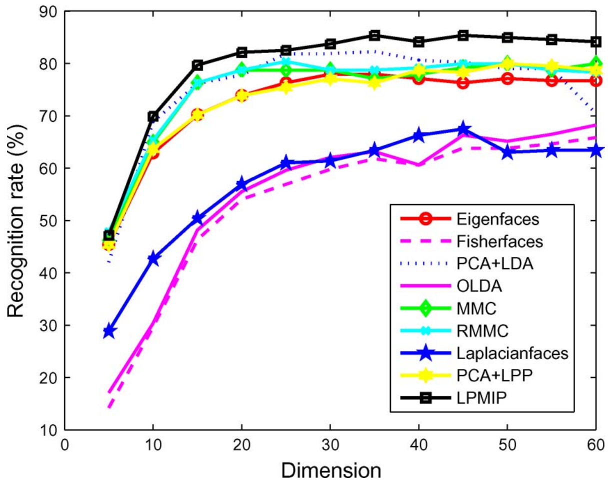


Fig. 7. Average recognition rates of various linear projection methods versus reduced dimensions on a subset of the color FERET database.

respect to the parameter α is plotted in Fig. 5, so we see that LPMIP has the best performance, and the parameter α again plays an important role in LPMIP. On this database, we could draw consistent conclusions with those of the ORL database.

3) Color FERET Database: We use three images per class for training and the remaining one image for testing. The final recognition rate is computed by averaging all four trials. The building of adjacency matrices is based on two nearest-neighbors, and the other settings are the same with previously. The maximal average recognition rates of various linear methods are presented in Table III. The recognition rates versus the reduced dimensions are shown in Fig. 7. We see again that LPMIP outperforms other linear methods.

C. Handwritten Digit Recognition

In this experiment, we evaluate the recognition performance of LPMIP on the United States Postal Service (USPS) database of handwritten digital character. The USPS database contains 7291 training samples and 2007 testing samples with dimensionality 256 of ten numeral classes. All these digits are collected from mail envelopes in Buffalo, NY. We choose 2000 training samples (200 samples per class) to train various linear methods, and then use all the 2007 testing points to compare their recognition rates, as shown in Table IV, where LPP and LPMIP use ten nearest-neighbors to build the adjacency matrix. The experimental protocol is the same as in Section V-B. From Table IV, we could see that in most cases, the LPMIP method achieves the best recognition accuracy. Although there does not exist singularity problem on this database, our experimental results still show that the best recognition rates of LDA and LPP are lower than those of LPMIP. The main reason could be attributed to the fact that the number of extracted features in LDA is limited to nine (which is referred to as the rank limitation problem), and LPP may not well consider the between-locality property of the data points. LPMIP, however, builds the adjacency/dissimilarity relationship in the data set and naturally partitions the whole data set (including the data points of the same

TABLE IV

COMPARISON OF MAXIMAL RECOGNITION RATES (IN PERCENT) ON THE USPS HANDWRITTEN DIGIT DATABASE. THE NUMBERS IN PARENTHESES OF PCA + LDA, RMMC, AND LPP ARE, RESPECTIVELY, THE OPTIMAL PERCENTAGE OF PCA ENERGY, r, AND m. FOR PCA + LPP, THE TWO NUMBERS ARE THE OPTIMAL PERCENTAGE OF PCA ENERGY, AND m AND FOR LPMIP, THE TWO NUMBERS ARE m AND a

	Number of features					
Methods	9	16	32	64	128	256
PCA	88.69	90.38	91.23	91.63	91.68	91.68
LDA	**91.78**	N.A.	N.A.	N.A.	N.A.	N.A.
PCA+LDA	**91.78** (99)	N.A.	N.A.	N.A.	N.A.	N.A.
OLDA	**91.78**	N.A.	N.A.	N.A.	N.A.	N.A.
MMC	80.07	87.84	87.84	87.89	88.04	**93.92**
RMMC	86.20 (12)	91.03 (9)	91.83 (9)	92.63 (0)	92.83 (0)	**93.92** (-3)
LPP	86.70 (2)	89.84 (2)	92.33 (2)	92.03 (-1)	91.58 (-1)	91.28 (0)
PCA+LPP	87.64 (92, 2)	90.23 (93, 2)	92.33 (99, 2)	92.03 (99, -1)	91.58 (99, -1)	91.58 (90, 0)
LPMIP	88.79 (-1, 9)	**93.72** (-2, 9)	**95.47** (-2, 7)	**94.77** (-2, 8)	**94.22** (-2, 10)	92.97 (-1, 8)

Fig. 8. Face examples from the JAFFE facial expression database.

class) into many small groups. Thus, a digit could be identified more accurately by matching several clusters corresponding to different handwritings. It could be observed that when the number of features is nine LDA achieves the best recognition rate. This shows the good discriminant ability of LDA when the number of training samples per class is relatively large.

D. Facial Expression Recognition

In this section, we use the Japanese female facial expression (JAFFE) database [26] for facial expression recognition to test the performance of LPMIP. The JAFFE database contains 213 images of ten Japanese women. The original facial images are of size 256 × 256 pixels, with 256 gray levels per pixel. Each image demonstrates one of seven expressions: neutral, happiness, sadness, surprise, anger, disgust, and fear. Some face examples from this database are shown in Fig. 8. In our experiment, we discard all the images of neutral expression, and only use the remainder 183 images that cover six basic facial expressions. As in [26], facial expression features are extracted by using the Gabor filter coefficients at 34 manually marked fiducial points on each face. Consequently, each image is represented as a 1020-dimensional labeled-graph (LG) vector, where the Gabor wavelet kernel takes five scales and six directions.

Generally speaking, facial expression recognition involves three steps. First, the facial expression features are extracted from the training set. Second, we learn a facial expression subspace. Then, the facial expression images are projected onto the subspace. Finally, we apply the nearest-neighbor classifier to identify new facial images into one of the six basic expression categories.

The recognition accuracy is evaluated by using the "leave-one-subject-out" cross-validation strategy. That is, in each trial, the facial images belonging to one person are chosen for testing, while the rest of the images are for training. We average the recognition results across all of the possible trials such that each subject is used once as the testing data. The average recognition rate for a novel expresser of the proposed LPMIP method is compared with the popular linear methods used previously, as well as linear canonical correlation analysis (CCA) [48]. Since CCA is one popular method for facial expression recognition, we include it here for comparison. Using LDA for facial expression recognition was first proposed in [26]. Although this experiment is for facial expression recognition, we still use the terms Fisherfaces and Laplacianfaces, the meanings of which are consistent with Section V-B. In the experiment, we use class labels of the training samples to design the adjacency matrix in LPP and LPMIP, and set $\sigma = +\infty$. The recognition results with the corresponding standard deviations and optimal reduced dimensions, as well as other parameter involved, are shown in Table V. We see that the LPMIP and RMMC methods significantly outperform the other methods. The PCA and CCA methods performs rather inefficiently. PCA+LDA, PCA+LPP, and RMMC greatly improve their corresponding counterparts: Fisherfaces, Laplacianfaces, and MMC. According to Theorem 3, RMMC in this experiment is approximately equivalent to LPMIP, so they have competitive performances.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new linear subspace learning method, called LPMIP, to identify the underlying manifold structure of the data set. LPMIP considers both the within-locality and the between-locality in the modeling of manifold learning. The objective of LPMIP is to preserve the local structure while maximize the nonlocal (global) information of the samples at the same time, so we present a criterion that is based on the difference of two terms reflecting the within-locality and the between-locality, respectively. The two localities are characterized and realized from the perspective graph theory. By constructing the adjacency matrix, the basis

TABLE V

COMPARISON OF MAXIMAL AVERAGE FACIAL EXPRESSION RECOGNITION RATES (IN PERCENT) AS WELL AS THE CORRESPONDING STANDARD DEVIATIONS AND OPTIMAL REDUCED DIMENSIONS USING "LEAVE-ONE-SUBJECT-OUT" CROSS VALIDATION ON THE JAFFE DATABASE

Methods	PCA	CCA	Fisherfaces	PCA+LDA	MMC	RMMC	Laplacianfaces	PCA+LPP	LPMIP
						$(r = 9)$			$(a = -4)$
Recog.	51.35	67.21	64.48	75.00	74.07	80.20	70.95	75.77	**83.18**
Std.	16.03	13.12	14.24	10.04	7.92	7.79	21.62	13.82	8.64
Dims	25	5	5	(95, 5)	5	5	4	(90, 4)	5

functions of LPMIP are solved as an eigenvalue problem, so LPMIP is computational feasible, and it effectively circumvents the singularity problem. The LPMIP/QR algorithm further alleviates the computational demand especially on high-dimensional data set. LPMIP actually gives a general framework in the sense that many popular linear subspace learning methods can be formulated under this umbrella. Extensive experimental results of data visualization and object recognition demonstrate the effectiveness of our method.

LPMIP is essentially linear. One possible extension of our work is to perform LPMIP in the reproducing kernel Hilbert $\mathcal{F}$ space induced by a nonlinear function ϕ. In implementation, we may resort to the kernel trick. The performance of kernel-based LPMIP needs to be further investigated. Another two questions are how to choose the value of α and how to determine the intrinsic dimensionality of the manifold theoretically. As we have seen, both of these two parameters are important for discrimination, but they may depend on the data set at hand and are not easy to be analytically determined. We are currently studying these problems from several perspectives.

APPENDIX
PROOF OF THEOREM 1

Suppose $\mathbf{t}$ is one column of the solution matrix $\mathbf{T}$ obtained from the LPMIP/QR algorithm. That is to say, there exists a λ such that

$$\mathbf{RMR}^T\mathbf{t} = \lambda\mathbf{t}.$$

Noting that $\mathbf{Q}^T\mathbf{Q} = I_t$, it then follows that

$$\mathbf{QRMR}^T\mathbf{Q}^T\mathbf{Qt} = \lambda\mathbf{Qt}.$$

Substituting $\mathbf{QR} = \mathbf{X}$ into the previous equation reads

$$(\mathbf{XMX}^T)\mathbf{Qt} = \lambda\mathbf{Qt}$$

which implies that $\mathbf{Qt}$ is an eigenvector of $\mathbf{XMX}^T$. The theorem is thus established. ■

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their helpful recommendations. Portions of the research in this paper use the FERET database of facial images collected under the FERET program.

REFERENCES

[1] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.

[2] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.

[3] 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.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, 2003.

[5] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering," in *Proc. 16th Conf. Neural Inf. Process. Syst.*, 2004, pp. 177–184.

[6] 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.

[7] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1550–1565, Nov. 2006.

[8] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 846–853.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annu. Eugenics*, vol. 7, pp. 179–188, 1936.

[11] D. H. Foley, "Considerations of sample and feature size," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 5, pp. 618–626, Sep. 1972.

[12] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences*, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, Eds. New York: Springer-Verlag, 1998, vol. 163, pp. 446–456.

[13] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2003.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacian faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[15] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educat. Psychol.*, vol. 24, pp. 417–441, 1933.

[16] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[17] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[18] E. Kokiopoulou and Y. Saad, "Face recognition using OPRA-faces," in *Proc. 4th Int. Conf. Mach. Learn. Appl.*, 2005, pp. 15–17.

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

[20] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data," *Appl. Statist.*, vol. 44, pp. 101–115, 1995.

[21] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise LDA for solving the small sample size problem in face verification," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 506–519, Mar. 2007.

[22] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.

[23] Y.-H. Liu and Y.-T. Chen, "Face recognition using total margin-based adaptive fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 178–192, Jan. 2007.

[24] J. Liu, S. Chen, X. Tan, and D. Zhang, "Comments on 'efficient and robust feature extraction by maximum margin criterion'," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1862–1864, Nov. 2007.

[25] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face recognition using kernel scatter-difference-based discriminant analysis," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 1081–1085, Jul. 2006.

[26] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[27] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[28] 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.

[29] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image Vis. Comput.*, vol. 16, pp. 295–306, 1998.

[30] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *J. Roy. Statist. Soc.*, ser. B, vol. 10, pp. 159–203, 1948.

[31] R. Rosipal, M. Girolami, L. J. Trejo, and A. Cichocki, "Kernel PCA for feature extraction and de-noising in nonlinear regression," *Neural Comput. Appl.*, vol. 10, pp. 231–243, 2001.

[32] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[33] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[34] J. Shawe-Taylor and N. Cristianini*, Kernel Methods for Pattern Analysis*. London, U.K.: Cambridge Univ. Press, 2004.

[35] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[36] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, pp. 71–86, 1991.

[37] N. Vlassis, Y. Motomura, and B. Krose, "Supervised dimension reduction of intrinsically low dimensional data," *Neural Comput.*, vol. 14, pp. 191–215, 2002.

[38] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extension: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[39] J. Yan, B. Zhang, S. Yan, Q. Yang, H. Li, Z. Chen, W. Xi, W. Fan, W. Ma, and Q. Cheng, "IMMC: Incremental maximum margin criterion," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2004, pp. 725–730.

[40] M. H. Yang, "Kernel Eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 215–220.

[41] J. Yang, A. F. Frangi, D. Zhang, J.-Y. Yang, and J. Zhong, "KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[42] J. Yang, D. Zhang, and J.-Y. Yang, "Locally principal component learning for face representation and recognition," *Neurocomputing*, vol. 69, pp. 1697–1701, 2006.

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

[44] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on under-sampled problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 982–994, Aug. 2004.

[45] J. Ye and Q. Li, "A two-stage linear discriminant analysis via QR-decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 929–941, Jun. 2005.

[46] W. Yu, X. Teng, and C. Liu, "Discriminant locality preserving projections: A new method to face representation and recognition," in *Proc. 2nd Joint IEEE Int. Workshop Vis. Surveillance—Perform. Eval. Track. Surveillance*, 2005, pp. 201–207.

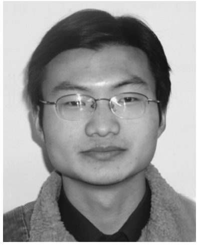
[47] W. Zhang, X. Xue, H. Lu, and Y.-F. Guo, "Discriminant neighborhood embedding for classification," *Pattern Recognit.*, vol. 39, pp. 2240–2243, 2006.

[48] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.

[49] W. Zheng, C. Zou, and L. Zhao, "Weighted maximum margin discriminant analysis with kernels," *Neurocomputing*, vol. 67, pp. 357–362, 2005.

Haixian Wang received the B.S. and M.S. degrees in statistics and the Ph.D. degree in computer science from Anhui University, Hefei, Anhui, China, in 1999, 2002, and 2005, respectively.

During 2002–2005, he was with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education of China. Currently, he is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu, China. His research interests focus on statistical pattern recognition, image processing, computer vision, and machine learning.



Sibao Chen received the B.S. and M.S. degrees in probability and statistics and the Ph.D. degree in computer science from Anhui University, Hefei, Anhui, China, in 2000, 2003, and 2006, respectively.

Currently, he is a Postdoctoral Researcher at the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui, China. His research interests include image processing, pattern recognition, machine learning, and computer vision.

Zilan Hu received the B.S. degree in mathematics from Anqing University, Anqing, Anhui, China, in 2003, and the M.S. degree in statistics from Anhui University, Hefei, Anhui, China, in 2006.

Currently, she is with the School of Mathematics and Physics, Anhui University of Technology, Maanshan, Anhui, China. Her research interests include statistical computing and applied statistics.

Wenming Zheng received the B.S. degree in computer science from Fuzhou University, Fuzhou, Fujian, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, Fujian, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, Jiangsu, in 2004.

Since 2004, he has been with the Research Center for Leaning Science (RCLS), Southeast University, Nanjing, Jiangsu, China. His research interests include neural computation, pattern recognition, machine learning, and computer vision.