Local and Weighted Maximum Margin Discriminant Analysis

Haixian Wang, Wenming Zheng
Key Laboratory of Child Development and Learning Science of Ministry of Education
Southeast University, Nanjing, Jiangsu 210096, China
hxwang@seu.edu.cn

Zilan Hu
School of Mathematics and Physics
Anhui University of Technology
Maanshan, Anhui 243002, China

Sibao Chen
Department of Electronic Engineering
and Information Science
University of Science and Technology of China
Hefei, Anhui 230027, China

Abstract

In this paper, we propose a new approach, called local and weighted maximum margin discriminant analysis (LWMMDA), to performing object discrimination. LWMMDA is a subspace learning method that identifies the underlying nonlinear *manifold for discrimination. The goal of LWMMDA is to seek a transformation such that data points of different classes are projected as far as possible while points within a same class are as compact as possible. The projections are obtained by maximizing a new discriminant criterion, called local and weighted maximum margin criterion (LWMMC). Different from previous maximum margin criterion (MMC) which seeks only the globally Euclidean structure of data points, LWMMC takes the local property into account, which makes LWMMC more accurate in finding discriminant information. LWMMC has an additional weighted parameter β that further broadens the average margin between different classes. Computationally, LWMMDA completely avoids the singularity problem. Besides, LWMMDA couples the QR-decomposition into its framework, which makes LWMMDA very efficient and stable in implementation. Finally, LWMMDA framework is straightforwardly extended into the reproducing kernel Hilbert space induced by a nonlinear function ϕ. Experiments on digit visualization, face recognition, and facial expression recognition are presented to show the effectiveness of the proposed method.*

1. Introduction

Automatical object recognition has been extensively used in a wide range of military, commercial and law enforcement applications. A few recognition subjects have been developed over the past few decades. Statistical pattern recognition is one of the most successful and well-studied subject. In statistical pattern recognition, a pattern is represented by a point in a high dimensional space. For example, a pattern of facial image with a resolution of 100×100 pixels yields a 10000-dimensional data point in the face space. The high dimensionality problem thus arises due to the fact the number of samples available is relatively small when compared with the sample features. This causes the "curse of dimensionality". The suggestion that at least ten times as many training samples per class as the dimensionality are used is a good practice [5]. However, in many practical applications, it is fairly expensive to obtain so large number of samples due to limitations of sample availability, time and cost. Learning a meaningful subspace in which the patterns possibly reside so as to reduce the dimensionality is an effective method for performing discrimination. In fact, the intrinsic dimensionality of the object space is much lower. In other words, the data points have an underlying structure of manifold embedded in the high-dimensional space. They substantially have a much more compact characterization.

In the past few decades, an increasing interest in unravelling the manifold of perceptual observation has been witnessed. Principal component analysis (PCA) [9] and linear discriminant analysis (LDA) [4] are two most classical techniques of learning a manifold, in which different properties are pursued by formulating different objective functions. PCA, also known as Karhunen-Loéve transformation, aims at preserving the global structure of the data set and seeks a set of mutually orthogonal basis of maximum variance. PCA has been successfully applied to discover the manifold of face space [10], which is called Eigenfaces

1-4244-1180-7/07/$25.00 ©2007 IEEE

method. LDA, also called Fisher's linear discriminant, is a supervised learning approach. The goal of LDA is to find a subspace projected onto which the data points of different classes are far from each other while the data points within a same class are close to each other. The optimal transformation is obtained by maximizing the ratio between the between-class scatter and the within-class scatter. LDA has been widely used in face recognition producing the well-known Fisherfaces method [2].

More recently, the maximum margin criterion (MMC) was developed by Li *et al*. [11] 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, when a suitable constraint is imposed, it actually gives rise to be LDA. Although both LDA and MMC are two supervised manifold learning methods, the implementation of MMC is much easier than that of LDA since MMC completely circumvents the inverse matrix operation and thus the small sample size (SSS) problem as exists in LDA. As we know, in under-sampled situation, the singularity makes LDA cannot be used directly. Meanwhile, it has been shown that MMC could achieve competitive (or better) recognition rate when compared with LDA and its variants [11]. Zheng *et al*. [19] gave a weighted version of MMC. Another non-parametric margin maximum criterion (NMMC) was also developed in literature [13]. This method has a reasonable interpretation for classification problem. However, its computational demand is intractable.

PCA, LDA, MMC and NMMC are all linear methods. They fail to discover the underlying nonlinear structure of the manifold. To overcome this drawback, the kernel-based counterparts are developed via the so-called kernel trick [15]. However, none of the kernel-based methods and their original prototypes explicitly consider the *local* structure of the manifold in which the patterns possibly reside. To explore the local structure, some nonlinear techniques have been developed, such as Isomap [16], LLE [14], and Laplacian eigenmaps [3]. They do give impressive performance on some benchmark data set. The transformation, however, is always implicit and is defined only on the training data set. So, it is unclear how to *analytically* evaluate the images of data points from the testing data set, which prohibits them from being applied in some pattern recognition problems.

Based on Laplacian eigenmaps, the locality preserving projections (LPP) was recently proposed to model the local manifold structure [7]. The manifold structure is modelled by using a nearest-neighborhood graph that keeps the neighborhood relationship. In projection process, LPP faithfully considers this graph structure. LPP inherits the locality preserving property of Laplacian eigenmaps via the adjacency graph. Further, LPP avoids the out-of-sample problem. When LPP is applied to face recognition, the method is called Laplacianfaces [8]. However, when performing LPP in under-sampled situation, the singularity problem arises. At this time, LPP can't be applied directly. To cope with this singularity problem, the common method is to perform an intermediate dimensionality reduction procedure using PCA [8]. However, the problem is that after performing PCA, can LPP still preserve the local information of original data set? Note that PCA is a global projection method, which may blend the local structure of data points. Moreover, the optimal value of the reduced dimensionality of PCA is difficult to determine. Besides, LPP is not inherently designed for discrimination. Since, even if the neighborhood is preserved in the lower-dimensional space, it is not easy to separate classes that have large spread and overlap with each other, LPP will not necessarily discover the most important manifold for discrimination problems. So, the discriminant locality preserving projections was also proposed by combining Fisher's criterion with LPP [17]. However, this method still suffers from the SSS problem.

In this paper, we propose a new approach, called local and weighted maximum margin discriminant analysis (LWMMDA), to performing discrimination. LWMMDA is to find projections that maximize the distances between different classes while minimize the distances between the data points within a same class. Unlike MMC and NMMC, the local information between different classes and the neighborhood information of the data points within a same class are incorporated into LWMMDA, which make LWMMDA more accurate in finding the discriminative information. The locality is modelled by constructing adjacency matrices, which is motivated by the idea of LPP. So, LWMMDA shares some similar properties with LPP such as explicitly considering the nonlinear manifold structure of the data set, although LWMMDA is essentially a linear projection method. LWMMDA is also defined on both the training and the testing data set, and not sensitive to outliers. However, the objective functions of LWMMDA and LPP are totally different. LWMMDA produces orthogonal basis while the basis of LPP is non-orthogonal. The non-orthogonality makes LPP difficult to reconstruct a data point. Computationally, LWMMDA completely avoids the SSS problem. Further, we develop an efficient and stable algorithm for performing LWMMDA by coupling the QR decomposition into LWMMDA framework. The algorithm for performing LWMMDA is theoretically established. The incremental property of the QR technique makes LWMMDA desirable in high-dimensional and dynamic databases, since performing the QR decomposition can work well when the training data are incremental. Finally, the LWMMDA algorithm can be transformed into the nonlinearly-related feature space $\mathcal{F}$ straightforwardly.

2. Local and weighted maximum margin discriminant analysis

Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a set of p-dimensional samples of size n, which belong to C classes. The number of samples in class c is n_c satisfying $\sum_{c=1}^{C} n_c = n$. We use $\mathbf{x}_i^c$ to denote the ith sample in class c for $i = 1, \ldots, n_c$, $c = 1, \ldots, C$. In the generic problem of subspace learning, we wish to find a linear transformation $\mathbf{V} \in \mathbb{R}^{p \times q}$ that maps each vector $\mathbf{x}_i$ in the p-dimensional space to a vector $\mathbf{y}_i$ in the lower q-dimensional space by $\mathbf{y}_i = \mathbf{V}^{\mathsf{T}}\mathbf{x}_i$. Given some optimal criterion, a good $\mathbf{V}$ will make $\mathbf{y}_i$ "represent" $\mathbf{x}_i$ well. We first consider a particular case that the p-dimensional data points are mapped to a line, i.e., $y_i = \mathbf{v}^{\mathsf{T}}\mathbf{x}_i$, where the transformation $\mathbf{v}$ is a vector. Since the magnitude of $\mathbf{v}$ is of no real interest, we let it be unitary norm.

The local and weighted maximum margin criterion (LWMMC) is proposed as follows:

$$\begin{aligned} J &= \beta \sum_{c=1}^{C}\sum_{d=1}^{C}(m_c - m_d)^2 B_{cd} \\ &\quad - (1-\beta)\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}(y_i^c - y_j^c)^2 W_{ij}^c, \end{aligned} \tag{1}$$

where $\beta \in [0, 1]$ is a parameter controlling the tradeoff between the first and the second term, y_i^c denotes the projection of $\mathbf{x}_i^c$, m_c is the mean of the projections of the samples belonging to class c, i.e., $m_c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_i^c$, W_{ij}^c is the weight of the edge that connects data points $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$, and B_{cd} is the weight between two mean vectors of classes c and d: $\mu_c (= \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{x}_i^c)$ and $\mu_d (= \frac{1}{n_d}\sum_{i=1}^{n_d}\mathbf{x}_i^d)$. Here, we are motivated by the utilization of graph as introduced in LPP. Specifically, from the graph perspective we represent the n_c samples of class c, i.e., $\mathbf{x}_1^c, \ldots, \mathbf{x}_{n_c}^c$, by a weighted undirected graph $G^c = (\mathcal{V}^c, \mathcal{E}^c)$, where $\mathcal{V}^c$ denotes a set of nodes that correspond to all the n_c data points, and $\mathcal{E}^c$ denotes the edges that connect pairwise points with the weight W_{ij}^c. Likewise, the C mean vectors μ_c, $c = 1, \ldots, C$, can also be represented by a graph; and the weights are given by B_{cd}. One reasonable definition of the weight matrix W^c is given by

$$W_{ij}^c = \exp(-\|\mathbf{x}_i^c - \mathbf{x}_j^c\|^2/\tau), \tag{2}$$

where τ is a positive parameter, and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^p$. In this paper, the parameter τ is set to be the maximal distance between pairwise data points within each class. Similarly, the weight matrix B can be defined as

$$B_{cd} = \exp(-\|\mu_c - \mu_d\|^2/\tau). \tag{3}$$

Maximize the objective function (1) equals to maximize the first term and minimize the second term simultaneously. The first term is a distance measure between different classes. Maximize it is an attempt to make different classes as far as possible. One interesting trick introduced here is that different classes are not processed equally. In other words, we pay special attention to the two mean vectors, say μ_c and μ_d, that are close each other. A heavy weight is put between μ_c and μ_d to ensure that their projections m_c and m_d are far away. As a result, it is easier to classify different classes, since even the originally close classes are transformed far away. The second term of the objective (1) is a distance measure within each class. Minimizing it is an attempt to make the data points of a same class as close as possible and meanwhile preserve the local structure of each class. That is, if $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ within a same class are close then y_i^c and y_j^c are close as well, since it will incur a heavy penalty if $(y_i^c - y_j^c)^2$ is large. The weight W_{ij}^c also deemphasize the atypical samples of a class, which make LWMMC robust to outliers. In a word, the first term of the objective function reflects between-class distance, while the second term reflects within-class distance. Maximizing the objective function seeks to maximize the "average margin" between different classes and meanwhile preserves the structure of each class. The property of preserving the structure of each class may be beneficial when using a nearest-neighbor classifier for discrimination. The parameter β is adjusted to balance the between-class distance and the within-class distance. The larger the β is, the more score the between-class get. β can be chosen by cross-validation.

By some algebraic operations, it follows that the first term of the objective function

$$\begin{aligned} &\sum_{c=1}^{C}\sum_{d=1}^{C}(m_c - m_d)^2 B_{cd} \\ =\ & \sum_{c=1}^{C}\sum_{d=1}^{C}\left(\frac{1}{n_c}\sum_{i=1}^{n_c} y_i^c - \frac{1}{n_d}\sum_{i=1}^{n_d} y_i^d\right)^2 B_{cd} \\ =\ & \sum_{c=1}^{C}\sum_{d=1}^{C}\left(\frac{1}{n_c}\sum_{i=1}^{n_c} \mathbf{v}^{\mathsf{T}}\mathbf{x}_i^c - \frac{1}{n_d}\sum_{i=1}^{n_d} \mathbf{v}^{\mathsf{T}}\mathbf{x}_i^d\right)^2 B_{cd} \\ =\ & \sum_{c=1}^{C}\sum_{d=1}^{C}\left(\mathbf{v}^{\mathsf{T}}\mu_c - \mathbf{v}^{\mathsf{T}}\mu_d\right)^2 B_{cd} \\ =\ & 2\sum_{c=1}^{C}\sum_{d=1}^{C}\mathbf{v}^{\mathsf{T}}\mu_c B_{cd}\mu_c^{\mathsf{T}}\mathbf{v} - 2\sum_{c=1}^{C}\sum_{d=1}^{C}\mathbf{v}^{\mathsf{T}}\mu_c B_{cd}\mu_d^{\mathsf{T}}\mathbf{v} \\ =\ & 2\mathbf{v}^{\mathsf{T}}\mathbf{M}D\mathbf{M}^{\mathsf{T}}\mathbf{v} - 2\mathbf{v}^{\mathsf{T}}\mathbf{M}B\mathbf{M}^{\mathsf{T}}\mathbf{v} \\ =\ & 2\mathbf{v}^{\mathsf{T}}\mathbf{M}(D-B)\mathbf{M}^{\mathsf{T}}\mathbf{v}, \end{aligned} \tag{4}$$

where the $p \times C$ matrix $\mathbf{M} = [\mu_1, \ldots, \mu_C]$, and D is a diagonal matrix whose entries are row (or column, notice the symmetry of B) sums of B, i.e., $D_{cc} = \sum_{d=1}^{C} B_{cd}$. Likewise, the second term of the objective function can be

reduced as

$$\begin{aligned}&\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{n_c}(y_i^c-y_j^c)^2W_{ij}^c\\ =\;& 2\sum_{c=1}^{C}\mathbf{v}^\mathsf{T}\mathbf{X}_c(E^c-W^c)\mathbf{X}_c^\mathsf{T}\mathbf{v},\end{aligned}\tag{5}$$

where the $p\times n_c$ matrix $\mathbf{X}_c=[\mathbf{x}_1^c,\ldots,\mathbf{x}_{n_c}^c]$, and E^c is a diagonal matrix with the entries $E_{ii}^c=\sum_{j=1}^{n_c}W_{ij}^c$. If write $\mathbf{X}=[\mathbf{X}_1,\ldots,\mathbf{X}_C]$, $E=\mathrm{diag}(E^1,\ldots,E^C)$, and $W=\mathrm{diag}(W^1,\ldots,W^C)$, then (5) can be rewritten as $2\mathbf{v}^\mathsf{T}\mathbf{X}(E-W)\mathbf{X}^\mathsf{T}\mathbf{v}$. Now, maximizing the objective (1) is converted to solve:

$$\arg\max_{\mathbf{v}}\quad \mathbf{v}^\mathsf{T}\mathbf{X}H\mathbf{X}^\mathsf{T}\mathbf{v}\tag{6}$$

$$\text{subject to } \mathbf{v}^\mathsf{T}\mathbf{v}-1=0,\tag{7}$$

where $H=\beta A(D-B)A^\mathsf{T}-(1-\beta)(E-W)$, and A is a block diagonal matrix diag $\left(\frac{1}{n_1}\mathbf{1}_{n_1},\ldots,\frac{1}{n_C}\mathbf{1}_{n_C}\right)$, where $\mathbf{1}_{n_c}$ is an n_c-dimensional vector with all entries being one. Here, we use the equation $\mathbf{M}=\mathbf{X}A$. Now, $\mathbf{v}$ can be immediately solved by using the Lagrangian multiplier technique. That is, the transformation vector $\mathbf{v}$ is given by the leading eigenvector of $\mathbf{X}H\mathbf{X}^\mathsf{T}$, associated with the largest eigenvalue. More generally, the q columns of the transformation matrix $\mathbf{V}$ are the first q largest eigenvectors of $\mathbf{X}H\mathbf{X}^\mathsf{T}$. Note that H is a symmetric matrix. So, the matrix $\mathbf{V}$ obtained is an orthogonal transformation.

2.1. LWMMC/ QR: an efficient algorithm for LWMMC via QR-decomposition

Maximizing LWMMC by diagonalizing the $p\times p$ matrix $\mathbf{X}H\mathbf{X}^\mathsf{T}$ is still time consuming in real world applications, since the dimensionality of the samples p is usually large. This occurs frequently in pattern classification, for example face recognition, gene expression data, and web document classification, in which the dimensionality can be up to several thousand. Besides, diagonalizing large matrix directly may give rise to attendant problem of numerical accuracy. To alleviate the computational demand, an efficient and effective algorithm for solving LWMMC, namely LWMMC/QR, is presented in this subsection.

By using the incomplete Cholesky decomposition [15], we know that the data matrix $\mathbf{X}$ can be QR-decomposed as $\mathbf{X}=\mathbf{QR}$ where $\mathbf{Q}\in\mathbb{R}^{p\times t}$ has 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, for the time being, suppose that the optimal transformation matrix $\mathbf{V}$ can be expressed as $\mathbf{V}=\mathbf{QT}$ for some $\mathbf{T}\in\mathbb{R}^{t\times q}$ having $\mathbf{T}^\mathsf{T}\mathbf{T}=I_q$ (since $\mathbf{V}$ has unitary columns), then the original problem of computing $\mathbf{V}$ is converted into computing $\mathbf{T}$ such that

$$\mathbf{T}=\arg\max_{\mathbf{T}^\mathsf{T}\mathbf{T}=I_q}\mathrm{tr}\left(\mathbf{T}^\mathsf{T}(\mathbf{Q}^\mathsf{T}\mathbf{X}H\mathbf{X}^\mathsf{T}\mathbf{Q})\mathbf{T}\right),\tag{8}$$

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

$$\mathbf{Q}^\mathsf{T}\mathbf{X}H\mathbf{X}^\mathsf{T}\mathbf{Q}=\mathbf{Q}^\mathsf{T}\mathbf{Q}\mathbf{R}H\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{Q}=\mathbf{R}H\mathbf{R}^\mathsf{T}.\tag{9}$$

Note that $\mathbf{R}H\mathbf{R}^\mathsf{T}$ is of size $t\times t$, which has much smaller size than that of $\mathbf{X}H\mathbf{X}^\mathsf{T}$, since usually $t\ll p$. The $\mathbf{T}$ is thus computed as the q eigenvectors of $\mathbf{R}H\mathbf{R}^\mathsf{T}$, associated with the q largest eigenvalues. As a result, the optimal solution of $\mathbf{V}$ is $\mathbf{V}=\mathbf{QT}$.

The relationship between LWMMC/QR and LWMMC are stated in the following theorem.

Theorem 1. *Let $\boldsymbol{X}=\boldsymbol{QR}$ be the QR-decomposition of $\boldsymbol{X}$, and $\boldsymbol{T}$ be the matrix whose columns are the eigenvectors of $\boldsymbol{RHR}^\mathsf{T}$, with the corresponding eigenvalues sorted in decreasing order. Let $\boldsymbol{V}=\boldsymbol{QT}$; that is, $\boldsymbol{V}$ is the optimal transformation matrix obtained from LWMMC/QR algorithm. Then the columns of $\boldsymbol{V}$ are just the leading eigenvectors of $\boldsymbol{XHX}^\mathsf{T}$ with the same eigenvalues.*

The proof is omitted here because of limit of space. This theorem shows that LWMMC/QR is equivalent to the standard LWMMC. However, the LWMMC/QR provides a computationally efficient and effective way for performing LWMMC. The QR-decomposition for computing $\mathbf{R}$ is of time complexity $O(t^2n)$. And solving the eigenvalue problem of $\mathbf{R}H\mathbf{R}^\mathsf{T}$ has the complexity $O(t^3)$. The storage requirement of the QR-decomposition is $O(tn)$. By contrast, the time complexity of diagonalizing $\mathbf{X}H\mathbf{X}^\mathsf{T}$ is $O(p^3)$ in the standard LWMMC. Both the time and storage complexity of LWMMC/QR compare favorably with that of LWMMC in the situation involving high-dimensional data set.

2.2. Learning LWMMC/ QR for discrimination

We call the discriminant analysis based on LWMMC/QR as local and weighted maximum margin discriminant analysis (LWMMDA). The algorithmic procedure of LWMMDA is formally summarized as follows.

1. **Compute the matrix H.**

2. **QR-decomposition.** By using the incomplete Cholesky decomposition technique, we decompose the data matrix $\mathbf{X}=\mathbf{QR}$, 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 $t=\mathrm{rank}(\mathbf{X})$.

3. **Eigenvalue decomposition.** Solve the eigenvalue problem $\mathbf{R}H\mathbf{R}^\mathsf{T}\mathbf{t}=\lambda\mathbf{t}$.

4. **Compute transformation matrix.** Let $\lambda_1 \geq \cdots \geq \lambda_q$ be the eigenvalues of $\mathbf{R}H\mathbf{R}^\mathsf{T}$ and $\mathbf{t}_1, \ldots, \mathbf{t}_q$ the corresponding eigenvectors. Then the optimal transformation matrix $\mathbf{V} = \mathbf{QT}$, where $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_q]$.

5. **Compute projections.** For a data point $\mathbf{x}$, its image in the lower-dimensional space is given by

$$\mathbf{x} \mapsto \mathbf{y} = \mathbf{V}^\mathsf{T}\mathbf{x}. \tag{10}$$

Specifically, for the training data set $\mathbf{X}$, the lower-dimensional embedding is

$$(\mathbf{y}_1, \ldots, \mathbf{y}_n)^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{V} = (\mathbf{QR})^\mathsf{T}\mathbf{QT} = \mathbf{R}^\mathsf{T}\mathbf{T}. \tag{11}$$

Suppose that $\mathbf{X}_{\text{test}}$ is the testing data set. By using the columns of $\mathbf{Q}$ as a basis, $\mathbf{X}_{\text{test}}$ can be decomposed as $\mathbf{X}_{\text{test}} = \mathbf{QR}_{\text{test}}$. Likewise, the lower-dimensional representation of the testing data set is $\mathbf{R}_{\text{test}}^\mathsf{T}\mathbf{T}$. It can be seen that the basis matrix $\mathbf{Q}$ in fact needs not be computed in implementation.

As can be seen, the implementation of LWMMDA is fairly straightforward. There is no need to compute any inverse matrix. So, it is computationally efficient and stable.

2.3. Kernel LWMMDA

The idea of kernel LWMMDA is to first map the input data into some new feature space $\mathcal{F}$ typically via a nonlinear function $\phi : \mathbb{R}^p \to \mathcal{F}$ and then carry out a linear LWMMDA in $\mathcal{F}$ using the mapped samples $\phi(\mathbf{x}_i)$. In implementation, the mapping ϕ does not need to be computed explicitly, while it and thus the space $\mathcal{F}$ are determined implicitly by the choice of a *kernel function* k which calculates the dot product between two mapped samples $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ in $\mathcal{F}$ by

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) \,. \tag{12}$$

The commonly used kernel functions include dth-order polynomial kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$, and Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma)$ with width $\sigma > 0$. Since the QR-decomposition of $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]$ only depends on the Gram matrix defined as $\mathbf{K} = (\mathbf{K}_{ij})_{n\times n}$ with elements $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ [15], and so does the computation of distances $\|\mathbf{x}_i^c - \mathbf{x}_j^c\|$ and $\|\mu_c - \mu_d\|$, LWMMDA can be applied in the feature space $\mathcal{F}$ directly. The computational complexity does not increase at all. This further illustrates the powerful and flexible of the proposed method.

3. Experiments

In this section, we conduct several experiments to investigate the performance of LWMMDA for data visualization, face recognition, and facial expression recognition.

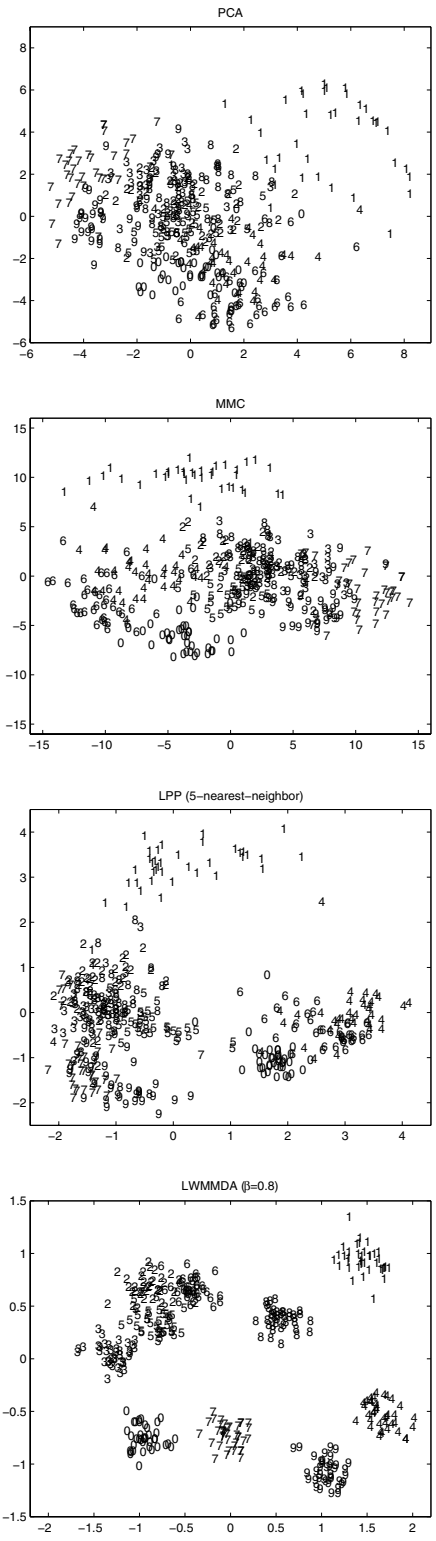

Figure 1. The handwritten digits from number 0 to 9 are mapped onto a 2-dimensional subspace using PCA, MMC, LPP and LWMMDA, respectively.

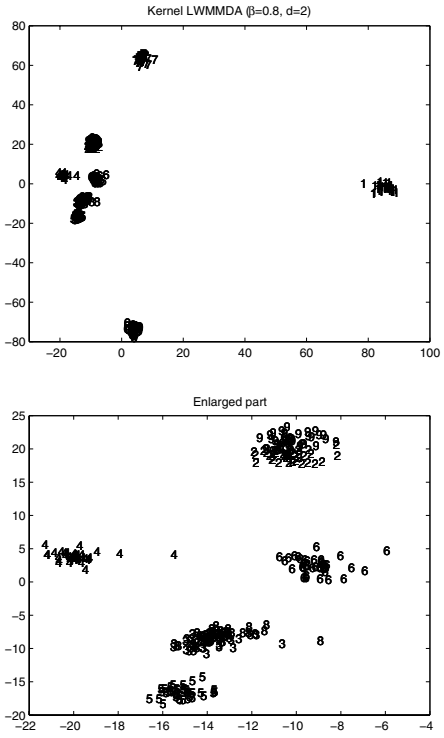

Figure 2. The handwritten digits from number 0 to 9 are mapped onto a 2-dimensional subspace using kernel LWMMDA. The left-center part of the top figure is enlarged in the bottom figure.

3.1. Data 2d-visualization

The experiment involves digit visualization. We use the digit database, which are publicly available from S. Roweis' web page (http://www.cs.toronto.edu/~roweis/data.html). This database contains 390 binary images of handwritten digits from "0" to "9", and each digit has 39 samples. The digit images are of size 20×16 pixels, and is represented by a 320-dimensional vector by lexicographic ordering of the pixel elements. The handwriting of some digits are somewhat illegible. These data points are mapped to a 2-dimensional subspace using four methods: PCA, MMC, LPP, and LWMMDA. The experimental results are depicted in Fig. 1. As can be seen, the projections of PCA are spread out since PCA aims at maximizing the variance. The classes of different digits have a heavy overlap. On the other hand, MMC and LPP yield more meaningful results. Clearly, the LWMMDA produces much better projections than PCA, MMC and LPP, since the clusters appears more compact. Finally, we illustrate the projections obtained by 2th-order polynomial kernel LWMMDA as shown in Fig. 2. The kernel LWMMDA gives a slightly better projections, by observing that different classes are far away except the numbers "2" and "9", and "3" and "8" are overlapped.

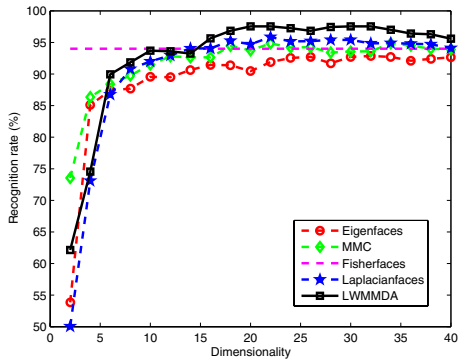

Figure 3. The average recognition accuracies vs. reduced-dimensionality on the UMIST database.

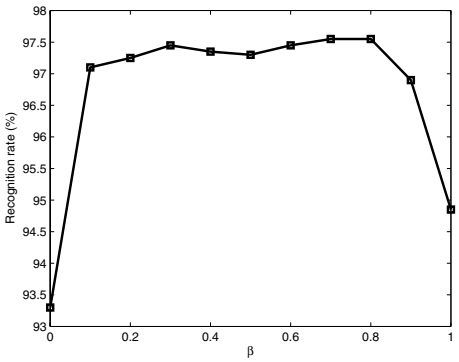

Figure 4. The average recognition accuracies vs. β using LWMMDA on the UMIST database.

3.2. Face recognition using LWMMDA

The system performance of the proposed LWMMDA method for face recognition is compared with Eigenfaces, MMC, Fisherfaces, and Laplacianfaces methods, four of the most popular methods in face recognition domain. To perform face recognition, we first obtain the face subspace from the training samples using these methods. Then facial images are represented by projecting onto the face subspace. Finally, we adopt the nearest-neighbor classifier to identify new facial images, where the Euclidean metric is used as the distance measure.

We conduct the experiment on the UMIST database. The UMIST database consists of 564 images of 20 subject [6]. The face images of each person cover a range of poses from profile to frontal views. Subjects vary with respect to race, sex and appearance. Precropped versions (with a size of 92×112 and 256 grey levels per pixel) of the images may also be made available from the same web site. In the experiments, the cropped images are further down-sampled into 23×28 pixels. As a result, each image is represented by a

644-dimensional vector in face space. We randomly select 10 images of each individual with labels as the training set, and the rest images in the database are for the testing set. We perform ten rounds of experiments with random selection of training data, and record the average results as the final recognition accuracy. Fig. 3 shows the recognition accuracy versus reduced-dimensionality by using various methods.

In using the Fisherfaces method, since there are at most $C-1$ nonzero generalized eigenvalues, we take the dimensionality of the reduced space being $C-1$. The graph structure used in Laplacianfaces is based on 5-nearest-neighbors. In general, the performance of LWMMDA method varies with the value of parameter β. We show the best results given by the optimal β. As can be seen, the LWMMDA method outperforms the other methods. In Fig. 4, we evaluate the influence of β on the classification accuracy. Ten rounds of experiments are tested, and the average result is summarized, where the reduced-dimensionality is fixed at 20. We observe that too small or too large β will not result in good recognition accuracy. So this, in turn, will depend to a large degree on the data set at hand.

3.3. Facial expression recognition using LWMMDA

This subsection tests the performance of LWMMDA for facial expression recognition. The experiment is performed on Japanese female facial expression (JAFFE) database [12]. There are 213 images posed by ten Japanese women in the JAFFE database. Each subject has two to four images for one of seven expressions: neutral, happiness, sadness, surprise, anger, disgust and fear. The original facial images are of the size 256×256 pixels, with 256 grey levels per pixel. In the experiment, only the images covering six basic facial expressions (excluding all neutral images) are chosen as experimental data. Thus, there are 183 facial images used.

We follow the work of Lyons *et al.* [12] to extract facial expression features. Specifically, 34 fiducial points are manually marked on each facial image, and then we compute the Gabor filter coefficients at the fiducial points. After Gabor filtering, we combine the amplitude values at the located fiducial points on each facial image into a labelled graph (LG) vector. Study shows that filter coefficients code facial expression better than geometric positions of the fiducial points [12]. In the experiment, we adopt the two-dimensional Gabor wavelet kernel:

$$g_{u,v}(\mathbf{z}) = \frac{\|\mathbf{k}_{u,v}\|^2}{\sigma^2} \exp\left(-\frac{\|\mathbf{k}_{u,v}\|^2\|\mathbf{z}\|^2}{2\sigma^2}\right) \times \left(\exp(i\mathbf{k}_{u,v}\mathbf{z}) - \exp\left(-\frac{\sigma^2}{2}\right)\right), \quad (13)$$

where $\mathbf{z}$ represents the location of a fiducial point and $\mathbf{k}_{u,v}$ is postulated as $\mathbf{k}_{u,v} = k_v \exp(i\phi_u)$. Here, u and v are given by $k_v = \pi/2^v$ ($v \in \{1,\ldots,5\}$) and $\phi_u = \pi u/6$ ($u \in \{0,\ldots,5\}$) respectively, which denote the orientation and scale respectively of the Gabor kernel. Hence, for each facial image, the LG vector is of dimensionality 1020 ($= 34 \times 5 \times 6$), where five scales and six directions for Gabor kernel are used. Using the extracted features, we then learn a facial expression subspace by the proposed method, onto which the facial expression images are projected.

Methods for FER	Recognition rate (%)
Gabor + LDA	64.48
Gabor + CCA	67.21
Gabor + LWMMDA ($\beta = 0.9$)	**78.35**
Gabor + GDA (polynomial kernel, $d = 3$)	68.85
Gabor + KCCA (Polynomial kernel, $d = 3$)	70.49
Gabor + LWMMDA ($\beta = 0.2$) (Polynomial kernel, $d = 7$)	**78.14**
Gabor + GDA (Gaussian kernel, $\sigma = 1e6$)	72.13
Gabor + KCCA (Gaussian kernel, $\sigma = 1e6$)	77.05
Gabor + LWMMDA ($\beta = 0.8$) (Gaussian kernel, $\sigma = 8e6$)	**78.10**

Table 1. Comparison of average facial expression recognition using "leave-one-subject-out" cross validation on JAFFE databases.

The recognition results of the proposed method and other two popular methods are listed in Table 1, where the reduced-dimensionality is 5 and we apply the nearest-neighbor classifier to identify new facial images into one of the six basic expression categories. The results are evaluated by using the "leave-one-subject-out" cross validation strategy. That is, the facial images that belong to one subject are chosen as testing data and the remainder facial images are used as training data. We repeat this procedure for all of the possible trials until every subject is used once as testing data. The final recognition rate is calculated by averaging all the recognition results. The system performance is compared with the generalized discriminant analysis (GDA) [1], kernel canonical correlation analysis (KCCA) [18] and linear discriminant analysis [12] respectively, three of the popular methods in facial expression classification. From Table 1, we see that the proposed method achieves much better performance than the other methods. For LWMMDA and kernel LWMMDA, the influence of β on average facial expression recognition is illustrated in Fig. 5. We observe that the kernel LWMMDA achieves only a slightly better results than linear LWMMDA. The reason could be attributed to that the linear LWMMDA has already considered the nonlinear structure of the data set.

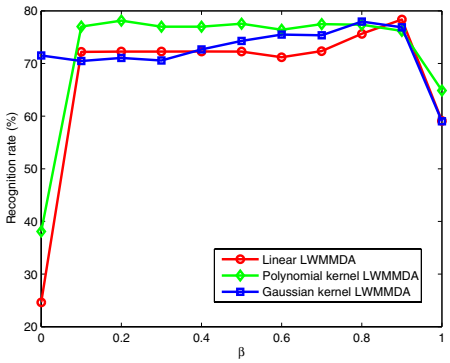

Figure 5. The influence of β on average facial expression recognition when using linear and kernel LWMMDA on JAFFE database.

4. Conclusions

A new linear subspace learning method called LWMMDA to identify the underlying nonlinear manifold for discrimination has been proposed in this paper. The projections are then obtained by maximize the average distance between different classes while minimize the distances between the data points within a same class. LWMMDA takes the local property of data points into account. LWMMDA is computationally easy. It totally circumvents the singularity problem. LWMMDA is straightforwardly extended into reproducing kernel Hilbert space induced by a nonlinear function ϕ. Two questions that need further investigate are how to choose the value of β and how to determine the intrinsic dimensionality of the manifold.

Acknowledgement

The authors gratefully acknowledge the partly financial supports of National Natural Science Foundation of China (Grant No. 60503023 and 10571001), Natural Science Foundation of Jiangsu Province (Grant No. BK2005407), Research Foundation of Southeast University (Grant No. 9250182304), Program of New Century Excellent Talents in University (Grant No. NCET-05-0467), Program of Excellent Young Teachers in Southeast University, and Key Laboratory of Image Processing and Image Communication of Jiangsu Province (Grant No. ZK205013).

References

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation* 12, 2385–2404, 2000. 7

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19, 711–720, 1997. 2

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Proceedings of the Advances in Neural Information Processing Systems* 14, pp. 585–591, 2002. 2

[4] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* Wiley, New York, 2001. 1

[5] D.H. Foley. Considerations of sample and feature size. *IEEE Trans. Information Theory* 18, 618–626, 1972. 1

[6] 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* 163, pp. 446–456, 1998. 6

[7] X. He and P. Niyogi. Locality preserving projections. *Proceedings of the Advances in Neural Information Processing Systems* 2003. 2

[8] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27, 328–340, 2005. 2

[9] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986. 1

[10] M. Kirby and L. Sirovich. Application of the Karhunen–Loéve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence* 12, 103–108, 1990. 1

[11] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans. Neural Networks* 17, 157–165, 2006. 2

[12] M.J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Trans. Pattern Analysis and Machine Intelligence* 21, 1357–1362, 1999. 7

[13] X. Qiu and L. Wu. Face recognition by stepwise nonparametric margin maximum criterion. *Proceedings of the Tenth IEEE International Conference on Computer Vision* pp. 1567–1572, 2005. 2

[14] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326, 2000. 2

[15] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, England, 2004. 2, 4, 5

[16] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323, 2000, 2

[17] W. Yu, X. Teng, and C. Liu. Discriminant locality preserving projections: a new method to face representation and recognition. *Proceedings 2nd Joint IEEE International Workshop on VS-PETS* pp. 201–207, 2005. 2

[18] W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE Trans. Neural Networks* 17, 233–238, 2006. 7

[19] W. Zheng, C. Zou, and L. Zhao. Weithted maximum margin discriminant analysis with kernels. *Neurocomputing* 67, 357–362, 2005. 2