Research Article

Semisupervised Kernel Marginal Fisher Analysis for Face Recognition

Ziqiang Wang, Xia Sun, Lijun Sun, and Yuchun Huang

School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China

Correspondence should be addressed to Ziqiang Wang; wangziqiang_a@163.com

Dimensionality reduction is a key problem in face recognition due to the high-dimensionality of face image. To effectively cope with this problem, a novel dimensionality reduction algorithm called semisupervised kernel marginal Fisher analysis (SKMFA) for face recognition is proposed in this paper. SKMFA can make use of both labelled and unlabeled samples to learn the projection matrix for nonlinear dimensionality reduction. Meanwhile, it can successfully avoid the singularity problem by not calculating the matrix inverse. In addition, in order to make the nonlinear structure captured by the data-dependent kernel consistent with the intrinsic manifold structure, a manifold adaptive nonparameter kernel is incorporated into the learning process of SKMFA. Experimental results on three face image databases demonstrate the effectiveness of our proposed algorithm.

1. Introduction

During the past decade, face recognition has been an active area of research in image processing and computer vision due to its extensive range of prospective applications, such as human-computer interface, information surveillance, and identity authentication. One of the most successful and well-studied techniques to face recognition is the appearance-based method. When using appearance-based methods, a face image of size $n_1 \times n_2$ pixels is usually represented by a vector in an $n_1 \times n_2$-dimensional space. Consequently, the face images are typically of very high dimensionality, ranging from several thousands to several hundreds of thousands. Due to the consideration of the curse of dimensionality, learning in such high dimensionality in many cases is computationally expensive and often leads to low recognition accuracy. One common response to address this problem is to apply dimensionality reduction techniques to generate a lower-dimensional equivalence of the original high-dimensional face image space for the given observations and targets. Once the high-dimensional face image data is projected into lower-dimensional feature subspace in which the semantic structure of the face image space becomes clear, traditional classification schemes can then be applied. To this end, principal component analysis (PCA) and linear discriminant analysis (LDA) [1] are the most well-known dimensionality reduction techniques.

PCA aims to find a set of mutually orthogonal basis vectors that capture the global information of the data points in terms of variance, and the orthogonal basis vectors are the leading eigenvectors of the data's total variance matrix associated with the leading eigenvalues. PCA is optimal in terms of representation and reconstruction, but not for discriminating one face class from others. Unlike PCA, which is unsupervised, LDA is a supervised dimensionality reduction algorithm. LDA aims to find an optimal transformation that maps the data into a lower-dimensional space that minimizes the within-class scatter and simultaneously maximizes the between-class scatter, thus achieving maximum discrimination. Both PCA and LDA have widely been applied to face recognition and image retrieval. It is generally believed that, when they come to solving problems of pattern classification, LDA-based algorithms outperform PCA-based algorithms since the former focuses on the most discriminant feature extraction while the latter achieves simply object reconstruction [2]. Independent component analysis (ICA) [3] is another linear subspace analysis method, which separates the high-order moments of the input data besides the second-order moments in PCA. However, previous researches reported that ICA gave the same recognition accuracy as

PCA, sometimes even a little worse than PCA [4]. In addition, NMF is also a subspace method which aims to find a parts-based representation of objects by imposing nonnegative constraints [5]. However, NMF is an unsupervised learning method and still focuses on the global geometrical structure of face image space. Moreover, the iterative update method for solving NMF problem is computationally expensive. In summary, the aforementioned algorithms see only the global Euclidean structure and can not discover the local manifold structure hidden in the high-dimensional data. In fact, a number of research efforts have shown that the face images possibly reside on a nonlinear submanifold hidden in the face image space [6–13]. Therefore, face representation is fundamentally related to the problem of manifold learning.

Manifold learning focuses on uncovering the compact, low-dimensional representations of the observed high-dimensional data that lie on or nearly on a manifold in an unsupervised manner. In order to detect the underlying manifold structure, many manifold learning algorithms have been proposed, such as isometric feature mapping (ISOMAP) [14], local linear embedding (LLE) [15], and Laplacian eigenmap (LE) [16]. ISOMAP, a variant of multidimensional scaling (MDS), aims to perverse global geodesic distances of all pairs of samples. LLE is based on the assumption that data lying on a nonlinear manifold can be viewed as linear in local areas, and it aims to discover the nonlinear structure via locally linear reconstructions. LE aims to preserve proximity relationships by manipulations on an undirected weighted graph, which indicates neighbour relations of pairwise data points. Thus, one of the key ideas of these manifold learning algorithms is the so-called locally invariant idea [17]; that is, the nearby points are likely to have the similar embedding/labels. Despite that these manifold learning algorithms have yielded impressive results on some benchmark artificial data set, they suffer from the out of sample problem; that is, they yield maps that are defined only on the training data points and how to evaluate the maps on novel test data points remains unclear. Therefore, these manifold learning algorithms might not be optimal in discriminating face images with different semantics, which is the ultimate goal of face recognition. To cope with the out of sample problem, He and Niyogi [18] applied a linearization procedure to construct explicit maps over new samples and proposed locality preserving projection (LPP) algorithm for manifold learning. LPP is a linearization of LE, aims to discover the local geometrical structure, and can be derived by finding the optimal linear approximations to the eigenfunctions of the Laplace-Beltrami operator on the manifold. As LPP is unsupervised, it is designed to best preserve data locality or similarity in the embedding space rather than good discriminating capability. As a result, the projected data points of different classes may still mix up after LPP embedding, which deteriorates the discrimination performance. In other words, for classification problem such as face recognition, the local manifold structure itself is not sufficient. A successful manifold learning algorithm should have the following two properties: (1) close intraclass pairs remain close after projection and (2) close but dissimilar pairs are kept separate after projection. Based on this consideration, Yan et al. [19] recently proposed the marginal

Fisher analysis (MFA) method for manifold learning by simultaneously utilizing the local manifold structure and the class label information. The empirical studies in [19] have shown that MFA is more competitive than LDA and LPP algorithms on face recognition.

MFA is a supervised learning method. It searches for the projection directions on which the marginal sample pairs of different classes are far away from each other while requiring data points of the same class to be close to each other. To obtain good generalization capability on testing samples, one needs a collection of labelled data points to train MFA. However, in the many practical applications of pattern classification (such as face recognition), one often faces a lack of sufficient labelled data, since labelling often requires expensive human labour and much time. Meanwhile, large numbers of unlabeled data can be far easier to obtain. Given the high cost in manually labelling face image data and at the same time abundant unlabeled face image is often easily accessible, it is desirable to develop dimensionality reduction methods that are capable of exploiting both labelled and unlabeled data. This motivates us to introduce semisupervised learning [20] into the dimensionality reduction process.

All the early semisupervised learning techniques mainly focus on semisupervised classifier design [21–25], which aims to employ a large number of unlabeled data to help build a better classifier from the labelled data. Recently, the semisupervised learning idea has been successfully applied to feature selection [26], clustering [27], distance metric learning [28], and matrix factorization [29]. Particularly, the semisupervised learning idea achieved great successes on various image analysis tasks. For example, semisupervised discriminant analysis (SDA) [30] used the consistency assumption; that is, nearby samples in the feature space or samples on the same manifold structure are likely to have the similar embedding/labels. All these approaches demonstrated that the learning performance can be significantly enhanced if the consistency assumption is exploited and the unlabeled data is considered. It is very natural that this idea should also be considered in semisupervised dimensionality reduction. However, most of the existing extension algorithms of MFA fail to take into account the intrinsic manifold structure revealed by unlabeled data points.

In this paper, we propose a novel semisupervised kernel MFA (SKMFA) algorithm, which takes advantage of both labelled and unlabeled data for face recognition. The main idea of our algorithm is to convert the traditional marginal Fisher analysis (MFA) into a semisupervised kernel counterpart, which still has no straightforward solution available in the literature. In addition, for semisupervised kernel MFA, the kernel function has an essential impact on the dimensionality reduction performance. Therefore, we propose to first induce a new manifold adaptive kernel by employing kernel deformation techniques to incorporate the manifold structure revealed by unlabeled data into the nonparameter kernel and then apply semisupervised kernel MFA to dimensionality reduction tasks by using the manifold adaptive kernel. Finally, extensive experiments on three face image databases demonstrate the effectiveness of the proposed SKMFA algorithm.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of marginal Fisher analysis (MFA) algorithm. Section 3 introduces our proposed semisupervised kernel MFA (SKMFA) algorithm for face recognition. The experimental results on face recognition are presented in Section 4. Finally, we provide the concluding remarks and suggestions for future work in Section 5.

2. Brief Review of MFA

Marginal Fisher analysis (MFA) [19] is a recently proposed manifold learning algorithm for dimensionality reduction; it is based on the graph embedding framework and can precisely model both intraclass compactness and interclass separability by jointly considering the local manifold structure and the label information, as well as characterizing the separability of different classes with the margin criterion. Meanwhile, MFA avoids the out of sample problem existing in traditional manifold learning algorithms by applying a linearization procedure to construct explicit maps over new samples.

Given a set of face images $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^p$, let $X = [x_1, x_2, \ldots, x_n]$; there are c classes; the ith face image x_i is associated with a class label $c_i \in \{1, 2, \ldots, c\}$. MFA aims to find a linear transformation $U \in \mathbb{R}^{p \times q}$ that maps each face image x_i $(i = 1, \ldots, n)$ in the p-dimensional space to a vector y_i in the lower q-dimensional space by $y_i = U^T x_i$ such that y_i represents x_i well in terms of maximizing the interclass separability and simultaneously minimizing the intraclass compactness. The optimal linear transformation of MFA can be obtained by solving the following maximization problem:

$$
\begin{aligned}
U_{\text{opt}} &= \arg\max_U \frac{\widetilde{S}_p}{\widetilde{S}_c} \\
&= \arg\max_U \frac{\sum_i \sum_{(i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j)} \left\| U^T x_i - U^T x_j \right\|^2 W_{ij}^p}{\sum_i \sum_{i \in N_{k_1}(j) \text{ or } j \in N_{k_1}(i)} \left\| W^T x_i - W^T x_j \right\|^2 W_{ij}} \\
&= \arg\max_U \frac{U^T X \left(D^p - W^p \right) X^T U}{U^T X \left(D - W \right) X^T U},
\end{aligned}
\tag{1}
$$

where $\widetilde{S}_p$ and $\widetilde{S}_c$ denote the interclass separability and intraclass compactness, respectively, and their definition, are as follows:

$$
\begin{aligned}
\widetilde{S}_p &= \sum_i \sum_{(i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j)} \left\| U^T x_i - U^T x_j \right\|^2 W_{ij}^p \\
&= 2 U^T X \left(D^p - W^p \right) X^T U \\
\widetilde{S}_c &= \sum_i \sum_{i \in N_{k_1}(j) \text{ or } j \in N_{k_1}(i)} \left\| W^T x_i - W^T x_j \right\|^2 W_{ij} \\
&= 2 U^T X \left(D - W \right) X^T U,
\end{aligned}
\tag{2}
$$

where W_{ij}^p and W_{ij} denote the weighting coefficients of penalty graph and intrinsic graph defined on the data points, respectively; W_{ij}^p and W_{ij} as well as their corresponding diagonal matrices D^p and D are defined as follows:

$$
\begin{aligned}
W_{ij}^p &= \begin{cases} 1, & \text{if } (i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j) \\ 0, & \text{otherwise} \end{cases} \\
W_{ij} &= \begin{cases} 1, & \text{if } i \in N_{k_1}(j) \text{ or } j \in N_{k_1}(i) \\ 0, & \text{otherwise} \end{cases} \\
D_{ii}^p &= \sum_j W_{ij}^p \\
D_{ii} &= \sum_j W_{ij},
\end{aligned}
\tag{3}
$$

where $P_{k_2}(c_i)$ denotes a set of data pairs that are the k_2 nearest pairs among the set $\{(i,j) \mid c_i \neq c_j\}$ and $N_{k_1}(i)$ denotes the index set of the k_1 nearest neighbours of sample x_i that are in the same class.

As can be seen from (1)-(2), the objective function of MFA is to look for an optimal transformation matrix U such that nearby data pairs in the same class are made close and the data pairs in different classes are separated from each other with the margin criterion. Therefore, maximizing it is an attempt to ensure both within-class compactness and between-class separability. Finally, the transformation matrices U of MFA are the eigenvectors associated with the largest eigenvalues of the following generalized eigenproblem:

$$
X \left(D^p - W^p \right) X^T U = \lambda X \left(D - W \right) X^T U.
\tag{4}
$$

$X(D - W)X^T$ is nonsingular after some preprocessing steps (such as PCA projection) on X; thus, the transformation matrix U of MFA can also be regarded as the eigenvectors of the matrix $(X(D - W)X^T)^{-1} X(D^p - W^p)X^T$ associated with the largest eigenvalues.

Despite the success of applying MFA to many fields, there are still some problems that are not properly addressed till now.

(1) MFA has a singular problem in face recognition, which stems from the fact that the number of training images is usually much smaller than the dimension of each image, a deficiency that is generally known as singular or small sample size (SSS) problem.

(2) MFA is a supervised learning method; it needs a collection of labelled data in order to guarantee good generalization capability on testing samples. However, for real-world face recognition, it is easy to obtain a large number of face images while only a few of them are labelled manually. In this case, purely supervised MFA cannot be well trained because of the lack of sufficient labelled data.

(3) MFA is still a linear technique in nature, so it is inadequate to describe the complexity of real face images because of illumination, facial expression, and

pose variations. Although the nonlinear extension of MFA through kernel trick has been proposed in [19], the most commonly adopted kernels are the data-independent kernels which may not be consistent with the intrinsic manifold structure revealed by unlabeled data.

To fully address the above issues, we propose a novel semisupervised kernel MFA (SKMFA) algorithm for face recognition in the following section.

3. Semisupervised Kernel MFA Algorithm for Face Recognition

In the following, we first propose the semisupervised MFA algorithm which can avoid the singular problem and consider the unlabeled samples to learn the projection matrix for dimensionality reduction, and then the nonlinear extension of semisupervised MFA through kernel trick is proposed. Finally, we discuss how to design manifold adaptive non-parameter kernel function which can reflect the underlying geometry of the data.

3.1. Semisupervised MFA. Although MFA can produce linear discriminating features, the matrices $X(D - W)X^T$ and $X(D^p - W^p)X^T$ in the generalized eigenproblem (4) are often singular because the number of available samples is smaller than the dimensionality of the samples. In order to avoid the numerical computational problem caused by matrix singularity, inspired by the scatter-difference-based discriminant analysis method [2, 31, 32], we modified the original objective function of MFA as

$$J(U)$$

$$= \arg \max \left(\widetilde{S}_p - \widetilde{S}_c \right)$$

$$= \arg \max \left(\begin{array}{c} \sum_i \sum_{(i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j)} \left\| U^T x_i - U^T x_j \right\|^2 W_{ij}^p \\ - \sum_i \sum_{i \in N_{k_1}(j) \text{ or } j \in N_{k_1}(i)} \left\| W^T x_i - W^T x_j \right\|^2 W_{ij} \end{array} \right)$$

$$= \arg \max \operatorname{Tr} \left(U^T X \left(D^p - W^p \right) X^T U - U^T X (D - W) X^T U \right)$$

$$= \arg \max \operatorname{Tr} \left(U^T \left(X \left(D^p - W^p \right) X^T - X (D - W) X^T \right) U \right).$$

$$(5)$$

Maximizing $J(U)$ is to find projections such that the close data points are attracted closer (minimizing the within-class compactness), while the data pairs in different classes are simultaneously separated from each other with the margin criterion (maximizing the between-class separability). So maximizing $J(U)$ can be equivalently interpreted as minimizing the within-class compactness while simultaneously maximizing the between-class separability, which is consistent with the optimal objective of MFA.

In the formulation defined in (5), we have the freedom to multiply U by some nonzero constant. Thus, we additionally require U to be orthonormal vectors, which may help preserve the shape of the data distribution. This means that we need to solve the following constrained optimization problem:

$$\arg \max_U \operatorname{Tr} \left(U^T \left(X \left(D^p - W^p \right) X^T - X (D - W) X^T \right) U \right)$$

$$(6)$$

$$\text{subject to } U^T U = I, \quad (7)$$

where I is the identity matrix.

It is worth noting that the only differences between the previous optimization problem and the original optimization problem of MFA lie in the following: the former involves a constrained optimization whereas the latter solves an unconstrained optimization. The motivation for using the constraint $U^T U = I$ is that it allows us not to calculate the inverse of the matrix $\widetilde{S}_p$ or $\widetilde{S}_c$, which successfully avoids the matrix singularity problem existing in the original MFA.

In addition, the original MFA is a supervised learning technique, which typically requires a large number of training samples in order to achieve satisfactory performance. However, for the practical large-scale applications such as face recognition, one often faces a lack of sufficient labelled face image data since labelling often requires expensive human labour and much time. Meanwhile, large numbers of unlabeled face data can be far easier to obtain due to the rapid advances of digital camera technology. In this case, purely supervised dimensionality reduction algorithms cannot be well trained because of the lack of sufficient labelled data, and purely unsupervised methods are usually unreliable because there is no supervision guidance. This motivates us to explore semisupervised learning [20] techniques for dimensionality reduction. Consequently, to leverage both the labelled and unlabeled data for dimensionality reduction, we propose the semisupervised MFA algorithm as follows.

In face recognition, since the number of labelled samples is small, it is important to consider the unlabeled samples to learn the projection matrix for dimensionality reduction. In fact, recent research has found that unlabeled samples may be helpful to improve the classification performance [33, 34]. In the following, we generalize MFA by introducing new reconstruction optimizations based on unlabeled samples and then incorporating them into the whole dimensionality reduction process, which leads to the semisupervised MFA algorithm.

Assume that unlabeled samples are attached on the original data set: $X = [x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{n+n_u}]$, where the first n samples are labelled and the remaining n_u samples are unlabeled. For each unlabeled sample data x_i ($n + 1 \leq i \leq n + n_u$), as in [12, 15, 33], we assume that all of its neighbourhoods are linear; that is, each data point can be optimally reconstructed using a linear combination of

its neighbours. Hence, our objective is to minimize the reconstruction error:

$$\arg\min_{V_{ij}} \sum_{i=n+1}^{n+n_u} \|\varepsilon_i\|^2$$

$$= \arg\min_{V_{ij}} \sum_{i=n+1}^{n+n_u} \left\| x_i - \sum_{n+1\leq j \neq i \leq n+n_u} V_{ij} x_j \right\|^2, \tag{8}$$

where ε_i is the reconstruction error for x_i and V_{ij} is the reconstruction coefficient which indicates the contribution of x_j to x_i. We further constrain $\sum_{n+1\leq j \neq i \leq n+n_u} V_{ij} = 1, V_{ij} \geq 0$ on (8). Obviously, the more similar x_j is to x_i, the larger V_{ij} will be. Hence, we can easily obtain $V_i = \sum_p C_{i,p}^{-1} / \sum_{p,q} C_{p,q}^{-1}$, $C_{p,q} = (x_i - x_p)^T (x_i - x_q)$ is the local Gram matrix, and V_{ij} is the jth elements of V_i, wherein $n+1 \leq p, q, i \leq n+n_u$.

Then, we can reconstruct $y_i (= U^T x_i)$ from $y_j (= U^T x_j)$ in the low-dimensional feature space by using the obtained reconstruction coefficient V_{ij}, wherein $n+1 \leq i, j \leq n+n_u$; that is,

$$J(U) = \arg\min_U \sum_{i=n+1}^{n+n_u} \left\| y_i - \sum_{n+1\leq j \neq i \leq n+n_u} V_{ij} y_j \right\|^2$$

$$= \arg\min_U \sum_{i=n+1}^{n+n_u} \left\| U^T x_i - \sum_{n+1\leq j \neq i \leq n+n_u} V_{ij} U^T x_j \right\|^2 \tag{9}$$

$$= \arg\min_U \operatorname{Tr}\left(U^T X_u \left(I - V^T\right) \left(I - V^T\right)^T X_u^T U \right),$$

where Tr() denotes the trace of matrix, I is the identity matrix, and $X_u = [x_{n+1}, x_{n+2}, \ldots, x_{n+n_u}]$ is the data matrix to represent all unlabeled high-dimensional data samples.

Considering all the samples including both labelled and unlabeled samples, we obtain the whole optimal objective function of semisupervised MFA by using (6) and (9):

$$\arg\max_U \operatorname{Tr}\left(\begin{array}{c} U^T \left(X\left(D^p - W^p\right)X^T - X\left(D - W\right)X^T\right)U \\ -\tau U^T X_u \left(I - V^T\right)\left(I - V^T\right)^T X_u^T U \end{array} \right)$$

$$= \arg\max_U \operatorname{Tr}$$

$$\times \left(U^T \left(\begin{array}{c} \left(X\left(D^p - W^p\right)X^T - X\left(D - W\right)X^T\right) \\ -\tau X_u \left(I - V^T\right)\left(I - V^T\right)^T X_u^T \end{array} \right) U \right)$$

$$= \arg\max_U \operatorname{Tr}\left(U^T M U \right) \tag{10}$$

$$\text{subject to } U^T U = I, \tag{11}$$

where

$$M = \left(\left(X\left(D^p - W^p\right)X^T - X\left(D - W\right)X^T \right) \right.$$

$$\left. -\tau X_u \left(I - V^T\right)\left(I - V^T\right)^T X_u^T \right) \tag{12}$$

and the parameter $\tau > 0$ is a scaling factor to balance contributions to optimize U of labelled samples against unlabeled samples.

Obviously, the optimal projection matrices U of (10) are the eigenvectors associated with the largest eigenvalues of the following standard eigenproblem:

$$MU = \lambda U. \tag{13}$$

Let the column vectors $U_1, U_2, \ldots, U_d$ be the solution of (13) ordered according to eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_d$. The optimal projection matrix U is given by $U = [U_1, U_2, \ldots, U_d]$. Then, the embedding of the proposed semisupervised MFA is as follows:

$$x \longrightarrow y = U^T x, \tag{14}$$

where y is a lower-dimensional representation of the face image x.

Since the proposed semisupervised MFA does not need to compute any matrix inverse for generating discriminating lower-dimensional features, it successfully avoids the singularity problem existing in the original MFA.

3.2. Nonlinear Generalization of Semisupervised MFA via Kernel Trick. In this section, we describe how to generalize our proposed semisupervised MFA to the nonlinear case by using kernel trick [35]. The main idea of kernel trick is to map the input data to a feature space through a nonlinear mapping, where the inner products in the feature space can be computed by a kernel function without knowing the nonlinear mapping explicitly. Kernel trick has demonstrated huge success in modelling real-world data with highly complex nonlinear structures, such as support vector machine (SVM) [35], kernel linear discriminant analysis (KLDA) [36], and kernel principal component analysis (KPCA) [37].

To extend semisupervised MFA to the nonlinear case, which leads to semisupervised kernel MFA, we consider the problem in a feature space F induced by some nonlinear mapping

$$\varphi : \mathbb{R}^n \longrightarrow F. \tag{15}$$

For a proper chosen φ, an inner product $\langle, \rangle$ can be defined in F, which makes for a so-called reproducing kernel Hilbert space (RKHS). More specifically,

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle. \tag{16}$$

Holds, where $K(,)$ is a positive semidefinite kernel function. The popular kernel functions include Gaussian kernel, polynomial kernel, and Sigmoid kernel.

Let $\widetilde{S}_p^\varphi$ and $\widetilde{S}_c^\varphi$ denote the interclass separability and intraclass compactness in the feature space F, respectively. We have

$$\widetilde{S}_p^\varphi = \sum_i \sum_{(i,j)\in P_{k_2}(c_i)\,or\,(i,j)\in P_{k_2}(c_j)} \left\| U^T \varphi(x_i) - U^T \varphi(x_j) \right\|^2 W_{ij}^p$$

$$= \mathrm{Tr}\left(U^T \varphi(X)\left(D^p - W^p\right)\varphi^T(X)U \right)$$

$$\widetilde{S}_c^\varphi = \sum_i \sum_{i\in N_{k_1}(j)\,or\,j\in N_{k_1}(i)} \left\| W^T x_i - W^T x_j \right\|^2 W_{ij}$$

$$= \mathrm{Tr}\left(U^T \varphi(X)(D-W)\varphi^T(X)U \right). \tag{17}$$

In addition, the reconstruction optimal function defined in (9) for unlabeled samples in the feature space can be transformed as follows:

$$\arg\min_U \sum_{i=n+1}^{n+n_u} \left\| y_i - \sum_{n+1\le j\ne i\le n+n_u} V_{ij} y_j \right\|^2$$

$$= \arg\min_U \sum_{i=n+1}^{n+n_u} \left\| U^T\varphi(x_i) - \sum_{n+1\le j\ne i\le n+n_u} V_{ij} U^T \varphi(x_j) \right\|^2$$

$$= \arg\min_U \mathrm{Tr}\left(U^T\varphi(X_u)\left(I-V^T\right)\left(I-V^T\right)^T \varphi^T(X_u)U \right). \tag{18}$$

By combining (17) and (18) together, both labelled and unlabeled samples will be considered in obtaining the projection matrix U in the feature space. Then, we obtain the following optimal objective function of semisupervised kernel MFA:

$$\arg\max_U \left(\widetilde{S}_p^\varphi - \widetilde{S}_c^\varphi - \tau \sum_{i=n+1}^{n+n_u} \left\| U^T\varphi(x_i) - \sum_{n+1\le j\ne i\le n+n_u} V_{ij} U^T \varphi(x_j) \right\|^2 \right)$$

$$= \arg\max_U \mathrm{Tr}\begin{pmatrix} U^T\varphi(X)(D^p-W^p)\varphi^T(X)U - U^T\varphi(X)(D-W)\varphi^T(X)U \\ -\tau U^T\varphi(X_u)\left(I-V^T\right)\left(I-V^T\right)^T\varphi^T(X_u)U \end{pmatrix}$$

$$= \arg\max_U \mathrm{Tr}\begin{pmatrix} U^T\left(\varphi(X)(D^p-W^p)\varphi^T(X)-\varphi(X)(D-W)\varphi^T(X)\right)U \\ -\tau U^T\varphi(X_u)\left(I-V^T\right)\left(I-V^T\right)^T\varphi^T(X_u)U \end{pmatrix}$$

$$= \arg\max_U \mathrm{Tr}\begin{pmatrix} U^T\varphi(X)\left((D^p-W^p)-(D-W)\right)\varphi^T(X)U \\ -\tau U^T\varphi(X_u)\left(I-V^T\right)\left(I-V^T\right)^T\varphi^T(X_u)U \end{pmatrix}. \tag{19}$$

Because any solution $U \in F$ must lie within the span of all the samples in F, there exist coefficients α_i ($i = 1, 2, \ldots, n, n+1, \ldots, n+n_u$) such that

$$U = \sum_{i=1}^{n+n_u} \alpha_i \varphi(x_i) = \varphi\alpha, \tag{20}$$

where $\varphi = \varphi(X) = [\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_{n+n_u})]$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{n+n_u})^T$.

Following some algebraic formulations, we can rewrite (19) as follows:

$$\arg\max_U \left(\widetilde{S}_p^\varphi - \widetilde{S}_c^\varphi - \tau \sum_{i=n+1}^{n+n_u} \left\| U^T\varphi(x_i) - \sum_{n+1\le j\ne i\le n+n_u} V_{ij} U^T \varphi(x_j) \right\|^2 \right)$$

$$= \arg\max_U \mathrm{Tr}\begin{pmatrix} U^T\varphi(X)\left((D^p-W^p)-(D-W)\right)\varphi^T(X)U \\ -\tau U^T\varphi(X_u)\left(I-V^T\right)\left(I-V^T\right)^T\varphi^T(X_u)U \end{pmatrix}$$

$$= \arg\max_\alpha \mathrm{Tr}\begin{pmatrix} \alpha^T K\left((D^p-W^p)-(D-W)\right)K\alpha \\ -\tau \alpha^T K_u\left(I-V^T\right)\left(I-V^T\right)^T K_u\alpha \end{pmatrix}$$

$$= \arg\max_\alpha \mathrm{Tr}\left(\alpha^T \begin{pmatrix} K\left((D^p-W^p)-(D-W)\right)K \\ -\tau K_u\left(I-V^T\right)\left(I-V^T\right)^T K_u \end{pmatrix}\alpha \right), \tag{21}$$

where K is the kernel matrix defined on the labelled samples and K_u is the kernel matrix defined on the unlabeled samples.

By imposing $\alpha^T\alpha = I$ on (21), the problem of semisupervised kernel MFA (SKMFA) is transformed into finding the leading eigenvectors of matrix $(K((D^p-W^p)-(D-W))K - \tau K_u(I-V^T)(I-V^T)^T K_u)$. Since no matrix inverse needs to be computed, SKMFA successfully avoids the singularity problem.

Thus, each eigenvector α gives a projection function U in the feature space. For a new sample data x, its projection onto U in the feature space F can be calculated by

$$\langle U, \varphi(x) \rangle = \sum_{i=1}^n \alpha_i \langle \varphi(x_i), \varphi(x) \rangle$$

$$= \sum_{i=1}^{n+n_u} \alpha_i K(x_i, x), \tag{22}$$

where $K(,)$ is a given kernel function.

3.3. Manifold Adaptive Nonparameter Kernel.

Similar to the other kernel-based methods, the kernel is also at the heart of semisupervised kernel MFA (SKMFA) algorithm. To achieve good performance, one has to define a good kernel presentation. However, the most commonly used kernels (such as Gaussian kernel, polynomial kernel, and Sigmoid kernel) are all data-independent kernels which may not be consistent with the intrinsic manifold structure revealed by unlabeled data points [38]. Meanwhile, these traditional kernels need complex operation to determine model parameters, which greatly limits their performance. To tackle the previous problems, a novel manifold adaptive nonparameter kernel function is proposed to improve the performance of SKMFA.

Let v be a linear space with a positive semidefinite inner product (quadratic from) and let $S : H \to v$ be a bounded linear operator. We define $\widetilde{H}$ to be the space of functions from H with the modified inner product

$$\langle f, g \rangle_{\widetilde{H}} = \langle f, g \rangle_H + \langle Sf, Sg \rangle_v. \tag{23}$$

It has been shown that $\widetilde{H}$ is still a reproducing kernel hilbert space (RKHS) [38].

Given the data points $x_1, x_2, \ldots, x_n$, let $S : H \to \mathbb{R}^n$ be the evaluation map $S(f) = (f(x_1), f(x_2), \ldots, f(x_n))$. Denote $f = (f(x_1), f(x_2), \ldots, f(x_n))$, $g = (g(x_1), g(x_2), \ldots, g(x_n))$. Note that $f, g \in v$; thus we can obtain

$$\langle Sf, Sg \rangle_v = \langle f, g \rangle = f^T M g, \tag{24}$$

where M is a symmetric positive semi-definite matrix. If we define

$$K_x = (K(x, x_1), \ldots, K(x, x_n)), \tag{25}$$

then it can be shown that the reproducing kernel $\widetilde{K}$ in $\widetilde{H}$ is of the following explicit form:

$$\widetilde{K}(x, z) = K(x, z) - K_x^T (I + MK)^{-1} M K_z, \tag{26}$$

where I is an identity matrix and K is the kernel matrix $K_{ij} = K(x_i, x_j)$ in H. The key issue now is the choice of M and the original kernel K, so that the deformation of kernel induced by the data-dependent norm is motivated with respect to the intrinsic geometry of the data.

In order to model the intrinsic manifold structure, as suggested in [16], the graph Laplacian implements a smoothness assumption with respect to an empirical estimate of the geometric structure of the data. Then we construct a nearest graph $G(V, W)$ to reflect the underlying manifold structure of the data. Each data point x_i corresponds to a node in V, and an edge is established between two nodes v_i and v_j if the corresponding two data points x_i and x_j are among k nearest neighbours of each other. Although there are many choices for the weight matrix on the graph, in order to eliminate the noise data on the manifold, we adopt the trick proposed in [39] to construct the nearest graph G. Let us define a distance function $h(x) = \|x - x^{(k)}\|$, where $x^{(k)}$ is the kth nearest neighbor of x in G. The weight matrix W associated with G is defined as follows:

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{\max\{h^2(x_i), h^2(x_j)\}} \right), \\ \qquad \text{if } \|x_i - x_j\| \le \max\{h(x_i), h(x_j)\} \\ 0, \\ \qquad \text{otherwise.} \end{cases} \tag{27}$$

The graph Laplacian is defined as $L = D - W$, where D is a diagonal degree matrix given by $D_{ii} = \sum_j W_{ij}$. The graph Laplacian provides the following smoothness penalty on the graph:

$$f^T L f = \sum_{i=1}^{n} (f(x_i) - f(x_j)) W_{ij}. \tag{28}$$

Thus, we set M in (26) to be $M = L$. Then, the next central issue is how to select an original input kernel function. The traditionally used kernel functions often assume certain parametric forms, but how to choose appropriate parameters of kernel function is an open problem, thus limiting their

capacity of fitting diverse patterns in real applications. To cope with that problem, we adopt the following nonparameter kernel function as the original input kernel function in (26).

Generally, a nonparameter kernel matrix K with respect to n patterns can be expressed as $K = X^T X \succ 0$, where $X = [x_1, x_2, \ldots, x_n]$ is the matrix of the embedding of data points [40]. The regularizer of the kernel matrix K, which captures the local dependency between the embedding of x_i and x_j, can be defined as

$$\sum_{i,j=1}^{n} W_{ij} \left\| \frac{x_i}{\sqrt{D_{ii}}} - \frac{x_j}{\sqrt{D_{jj}}} \right\|^2 \tag{29}$$

$$= \mathrm{Tr}\left(X \widetilde{L} X^T \right) = \mathrm{Tr}\left(\widetilde{L} K \right),$$

where $\widetilde{L}$ is the normalized graph Laplacian matrix defined as follows:

$$\widetilde{L} = I - D^{-1/2} W D^{-1/2}, \tag{30}$$

where W is defined in (27) and D is a diagonal degree matrix given by $D_{ii} = \sum_j W_{ij}$.

In addition, for the semisupervised learning algorithm, recent researches have pointed out that the class label information is not readily available, while it is easier to obtain some collection of similar pairwise constraints S (known as "must links," i.e., the data pairs share the same class label) and a collection of dissimilar pairwise DS (known as "cannot-links," i.e., the data pairs have different class label), which is often referred to as "side information." Given S and DS, we construct a similarity matrix $T \in \mathbb{R}^{n \times n}$ to represent the pairwise constraints; that is,

$$T_{ij} = \begin{cases} +1, & (x_i, x_j) \in S \\ -1, & (x_i, x_j) \in \mathrm{DS} \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

Then, an intuitive principle for kernel learning is that the kernel entry $K_{ij} = K(x_i, x_j)$ should be aligned with the side information T_{ij} as much as possible; that is, the alignment $T_{ij} K_{ij}$ of each kernel entry is maximized.

Therefore, following the suggestions in [40], by simultaneously considering the side information in (31) and the regularizer in (29), the nonparameter kernel learning can be formulated as follows:

$$\min_{K \succ 0} \mathrm{Tr}\left(\widetilde{L} K \right) + C \sum_{(x_i, x_j) \in S \cup DS} l\left(T_{ij} K_{ij} \right), \tag{32}$$

where C is the positive constant to control the tradeoff between the empirical loss $l()$ and the intrinsic data manifold and $l()$ is the square hinge loss function defined as follows:

$$l(f) = \frac{1}{2}(\max(0, 1 - f))^2. \tag{33}$$

It is worth noting that the previous optimization problem belongs to a semidefinite programming (SDP) problem,

which can be solved with standard SDP solver SeDuMi [41]. Once we obtain the optimal nonparameter kernel K, by substituting M $(= L)$ and K into (26), we eventually get the following manifold kernel function:

$$\widetilde{K}(x, z) = K(x, z) - K_x^T (I + LK)^{-1} LK_z. \quad (34)$$

3.4. The Semisupervised Kernel MFA Algorithm. We summarize our proposed semisupervised kernel MFA (SKMFA) algorithm as follows.

(1) Calculate the initial nonparameter kernel matrix K by solving the optimization problem (32).

(2) Compute the weight matrix W in terms of (27) and set $M = L$, where $L = D - W$ is graph Laplacian and D is a diagonal degree matrix given by $D_{ii} = \sum_j W_{ij}$.

(3) Obtain the manifold adaptive kernel $\widetilde{K}$ according to (34).

(4) Find the eigenvector α by solving the following eigenproblem:

$$\begin{pmatrix} \widetilde{K} ((D^p - W^p) - (D - W)) \widetilde{K} \\ -\tau \widetilde{K}_u (I - V^T) (I - V^T)^T \widetilde{K}_u \end{pmatrix} \alpha = \lambda \alpha. \quad (35)$$

(5) For a new sample data x, its lower-dimensionality representation y can be calculated by

$$x \longrightarrow y = \sum_{i=1}^{n+n_u} \alpha_i \widetilde{K}(x_i, x), \quad (36)$$

where $\widetilde{K}(,)$ is the manifold adaptive kernel defined in (34).

4. Experimental Results

In this section, we investigate the performance of our proposed semisupervised kernel MFA (SKMFA) algorithm and compare it with other representative dimensionality reduction algorithms for face recognition. All of our experiments have been performed on a P4 3.5 GHz Windows XP machine with 2 GB memory.

4.1. Face Databases and Experimental Settings. Three real-world face databases are used in our experimental study, including the Yale database, the Olivetti Research Laboratory (ORL) database, and the PIE (pose, illumination, and expression) database from CMU. In all experiments, preprocessing to locate the faces was applied. Original face images were manually aligned, cropped, and then resized to 32×32 pixels, with 256 gray levels per pixel. The important statistics of three databases are summarized next.

The Yale face database (http://cvc.yale.edu/projects/yale-faces/yalefaces.html) was constructed at the Yale centre for computational vision and control. There are 15 persons, and each person has 11 different images. The images demonstrate variations in lighting condition and facial expression. Some

FIGURE 1: Face image examples of the Yale database.

FIGURE 2: Face image examples of the ORL database.

FIGURE 3: Face image examples of the CMU PIE database.

sample face images after preprocessing of the database are shown in Figure 1.

In the ORL database (http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html), there are 40 persons, and each person has 10 different images. Some images were captured at different times and have different variations including expression and facial details. Some sample face images after preprocessing of the database are shown in Figure 2.

The CMU PIE database [42] contains 68 individuals with 41368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes under varying poses, illumination, and expressions. In this experiment, we choose the five frontal poses (C05, C07, C09, C27, and C29) and illumination indexed as 10 and 13 which contain 10 front face images for each person. Some sample face images after preprocessing of the database are shown in Figure 3.

The experimental settings are as follows. We randomly selected 10 face images per individual to form the training set for Yale database and 8 face images per individual to form the training set for ORL and CMU PIE databases. The remaining images for each person were used for the testing set. In the training set, we randomly selected l $(= 2, 4)$ face images per individual as labelled data set and the rest as unlabeled data set for each face image database.

To perform face recognition, we first obtain the face subspace by using dimensionality reduction algorithms. Then, the new face images to be identified are projected onto the face subspaces. Finally, the nearest-neighbour classifier is adopted to identify new facial images, where the Euclidean metric is used as the distance measure.

4.2. Compared Algorithms. Five algorithms, which are compared in our experiments, are listed below:

(1) Marginal Fisher analysis (MFA) [19], which provides us with a baseline performance of linear dimensionality reduction algorithms. We can examine the usefulness of kernel approaches by comparing the performance of kernel Marginal Fisher analysis with MFA.

(2) Kernel marginal Fisher analysis (KMFA) is proposed in [19]. This method is the nonlinear extension of

Table 1: Comparisons of recognition accuracy (in percent) on the Yale database.

Algorithm	Two labels	Four labels
MFA	58.4 ± 1.15	68.7 ± 1.27
KMFA	67.8 ± 1.28	76.3 ± 1.33
SDA	69.3 ± 1.49	77.6 ± 1.25
KSDA	70.7 ± 1.32	78.4 ± 1.18
SKMFA	73.6 ± 0.73	81.5 ± 0.69

Table 2: Comparisons of recognition accuracy (in percent) on the ORL database.

Algorithm	Two labels	Four labels
MFA	87.5 ± 1.43	91.3 ± 1.41
KMFA	88.4 ± 1.50	93.8 ± 1.38
SDA	87.9 ± 1.56	93.2 ± 1.42
KSDA	88.6 ± 1.37	94.5 ± 1.29
SKMFA	93.4 ± 1.25	96.1 ± 1.07

Table 3: Comparisons of recognition accuracy (in percent) on the CMU PIE database.

Algorithm	Two labels	Four labels
MFA	70.2 ± 1.90	81.3 ± 1.76
KMFA	72.5 ± 1.85	83.4 ± 1.72
SDA	73.1 ± 1.63	83.7 ± 1.49
KSDA	75.4 ± 1.27	84.6 ± 0.98
SKMFA	79.3 ± 0.98	85.8 ± 0.75

the traditional MFA via kernel trick. The settings of KMFA algorithm are identical to the description in the corresponding paper [19].

(3) Semisupervised discriminant analysis (SDA) is proposed in [30], which is believed to be one of the most representative semisupervised dimensionality reduction algorithms.

(4) Kernel semisupervised discriminant analysis (KSDA) is proposed in [30]. This method is the nonlinear extension of the traditional SDA via kernel trick. The settings of KSDA algorithm are identical to the description in the corresponding paper [30].

(5) Semisupervised kernel marginal Fisher analysis (SKMFA), as described in Section 3, is a new method proposed in this paper.

Note that the settings of the compared algorithms are identical to the description in the corresponding papers. For our proposed SKMFA algorithm, there is a parameter τ which controls the balance contributions of labelled samples against unlabeled samples. We simply set the value of τ as 1, and the effect of parameter selection will be discussed later.

4.3. Face Recognition Results. For each given l (= 2, 4), we average the results over 20 random splits and report the mean as well as the standard deviation. The face recognition accuracies for each algorithm in three face databases are reported on
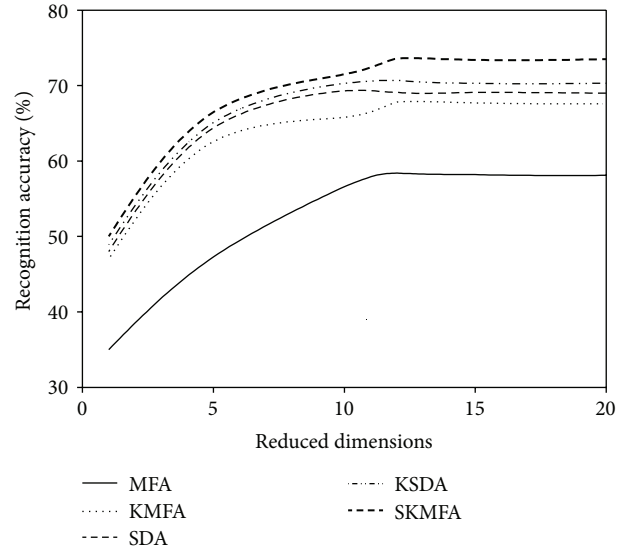

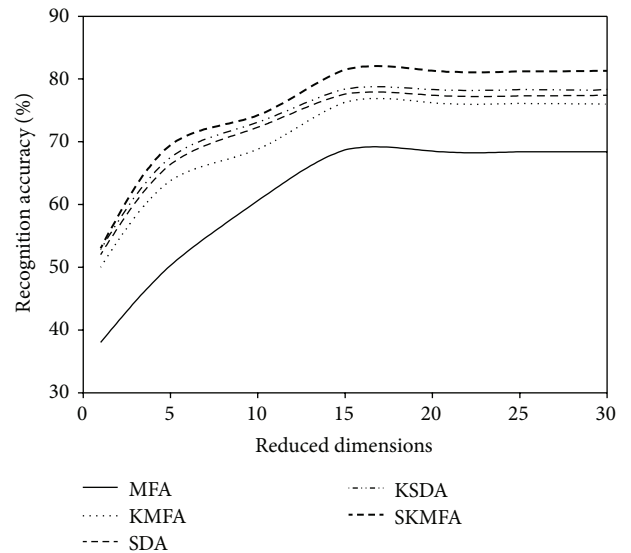Figure 4: Two labelled data for training on the Yale database.

Figure 5: Four labelled data for training on the Yale database.

the Tables 1, 2, and 3, respectively. The recognition accuracies versus the variation of reduced dimensions are shown in Figures 4, 5, 6, 7, 8, and 9. The main observations from the previous performance comparisons include the following.

(1) Our proposed SKMFA algorithm consistently outperforms the MFA, KMFA, SDA, and KSDA algorithms in terms of recognition accuracy, which indicates that SKMFA can effectively use the intrinsic nonlinear manifold structure revealed by unlabeled data to improve the recognition accuracy.

(2) The KMFA and KSDA algorithms achieve higher recognition accuracy than their linear counterparts (i.e., MFA and SDA), respectively, which suggests the effectiveness of kernel approaches.
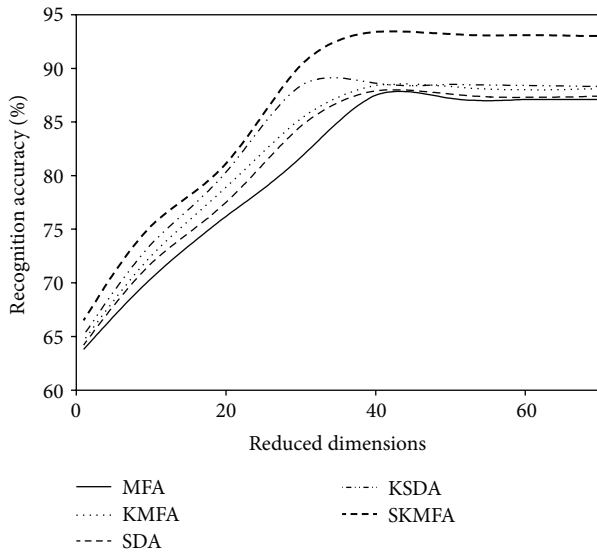
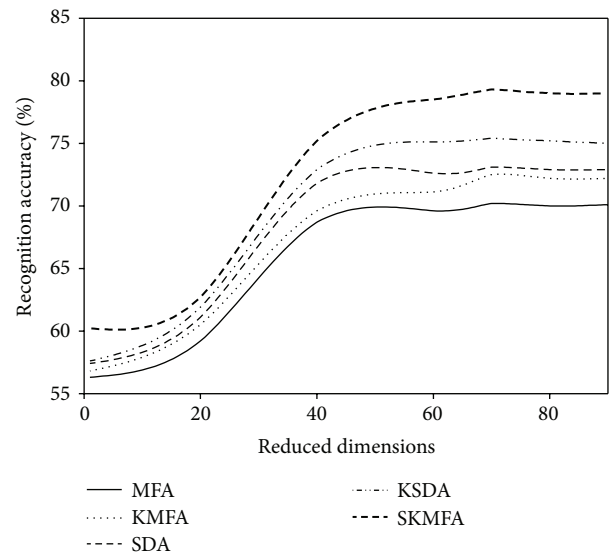FIGURE 6: Two labelled data for training on the ORL database.

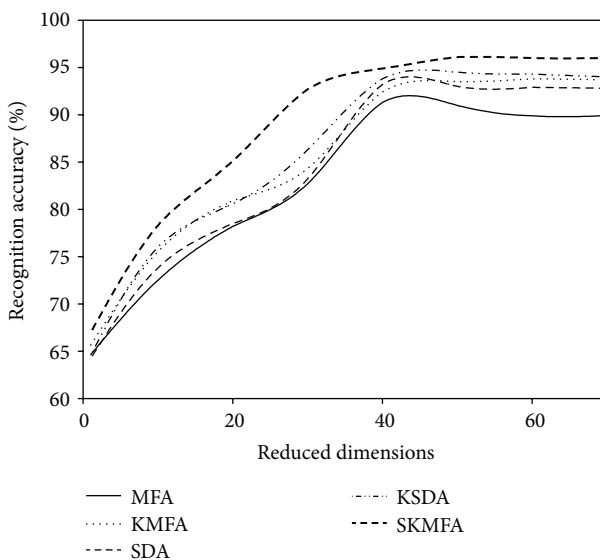FIGURE 8: Two labelled data for training on the CMU PIE database.

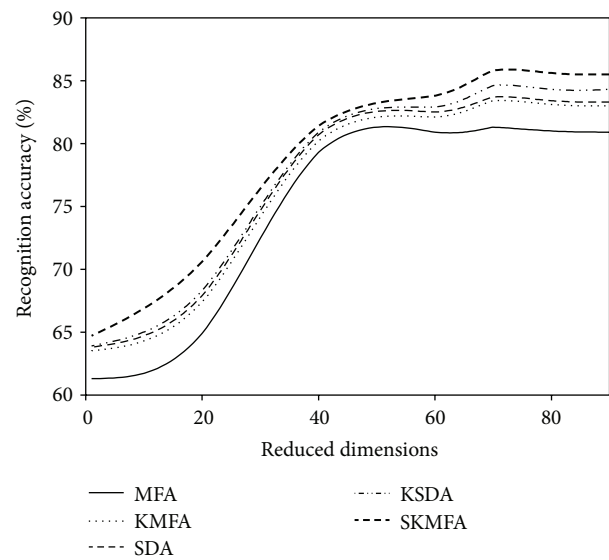FIGURE 7: Four labelled data for training on the ORL database.

FIGURE 9: Four labelled data for training on the CMU PIE database.

(3) The semisupervised algorithms (SKMFA, KSDA, and SDA) achieve higher recognition accuracy than the supervised algorithm (MFA), which demonstrates that these semisupervised algorithms can effectively utilize only a few labelled samples to predict the labels of the unlabeled samples.

(4) The recognition accuracies of SDA and KMFA are almost similar. For some databases, SDA outperforms KMFA, while KMFA is better than SDA for other databases. A possible explanation is as follows: KMFA is only a supervised nonlinear algorithm (not a semi-supervised algorithm), while SDA is only a semisu-pervised linear algorithm (not a nonlinear algo-rithm). Thus, it is hard to evaluate whether nonlinear

extension with kernel trick or semisupervised infor-mation is more important for recognition.

(5) Although KSDA and KMFA belong to kernel-based nonlinear manifold learning algorithm, KSDA per-forms better than KMFA. The possible explanation is that KSDA can utilize a large number of unlabeled data as well as relatively limited labelled data for better discrimination ability.

(6) Although KMFA and SKMFA are the nonlinear extensions of MFA via kernel trick, KMFA performs much worse than SKMFA. This is because KMFA adopts the commonly used data-independent kernel which may not be consistent with the intrinsic mani-fold structure of face images.
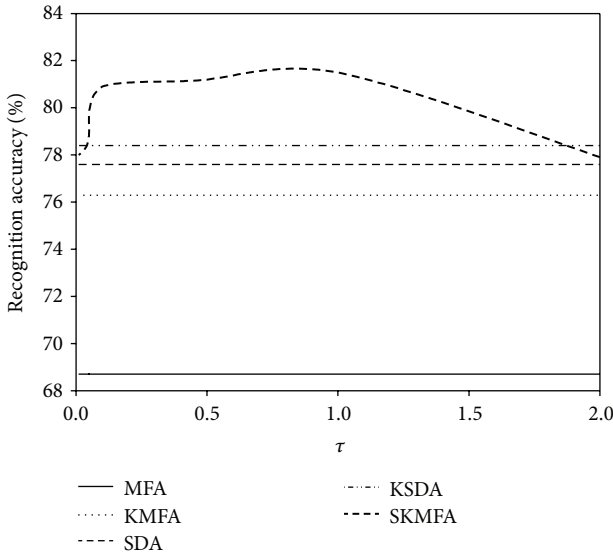
FIGURE 10: The performance of SKMFA varies with the parameter τ on the Yale database.
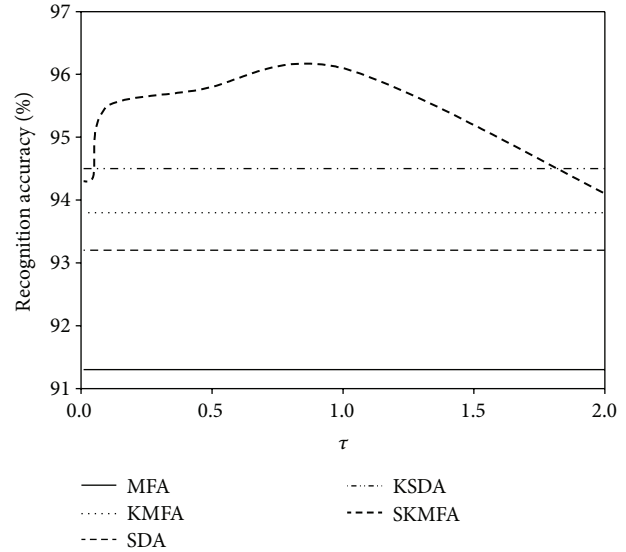

FIGURE 11: The performance of SKMFA varies with the parameter τ on the ORL database.

(7) Our proposed SKMFA algorithm achieves much better performance than the MFA, KMFA, SDA, and KSDA algorithms. The main reason could be attributed to the following fact: first, SKMFA simultaneously considers the intraclass geometry and the interactions of samples from different classes; second, SKMFA successfully avoids the singularity problem without calculating the matrix inverse; third, the manifold adaptive kernel is consistent with the intrinsic manifold structure revealed by unlabeled data points and can effectively capture the nonlinear structure of face images. Therefore, our proposed SKMFA algorithm achieves the best performance among the compared algorithms by simultaneously using the aforementioned optimal strategies.

4.4. Parameter Selection for SKMFA. The parameter $\tau > 0$ is an essential parameter in our SKMFA algorithm which controls the balance contributions of labelled samples against unlabeled samples. We empirically set it to be 1 in the previous experiments. In this subsection, we try to examine the impact of parameter τ on the performance of SKMFA. Figures 10, 11, and 12 show how the average performance of SKMFA varies with τ.

5. Conclusions

In this paper, we have proposed a novel nonlinear algorithm, called semisupervised kernel marginal Fisher analysis (SKMFA), for face recognition. It can make efficient use of both labelled and unlabeled data points for nonlinear dimensionality reduction. The labelled data points are used to maximize the discriminating power, while the unlabeled data points are used to reveal the intrinsic manifold structure. In addition, the manifold adaptive kernel is adopted to further
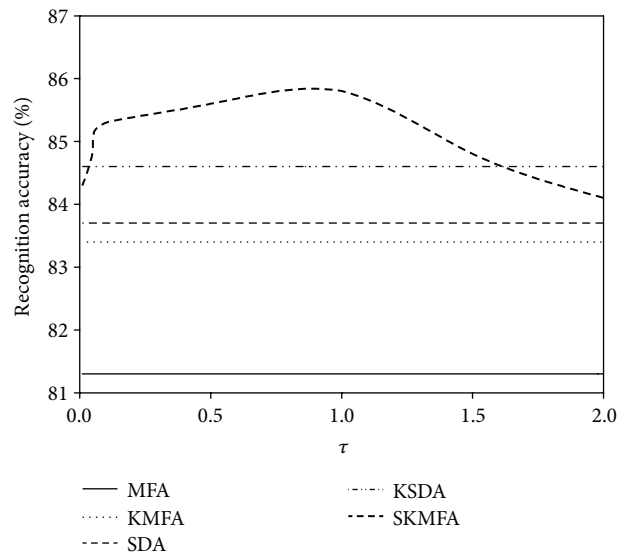

FIGURE 12: The performance of SKMFA varies with the parameter τ on the CMU PIE database.

improve the algorithm performance. Experimental results on three face image databases demonstrate the effectiveness of our proposed algorithm. Since our proposed SKMFA algorithm is a general nonlinear dimensionality reduction algorithm for high-dimensional data, we plan to apply the algorithm to video and audio classification in the future.
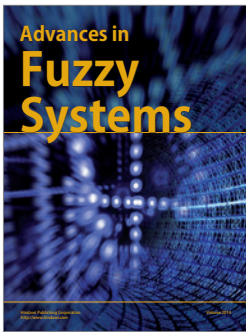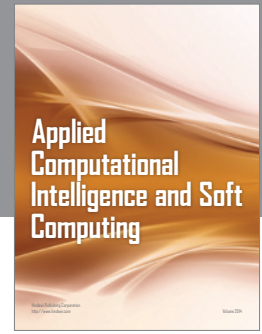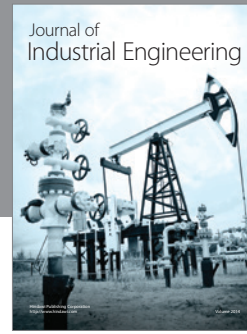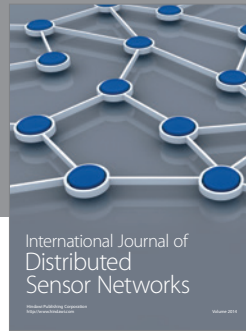
Acknowledgments

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, Hoboken, NJ, USA, 2nd edition, 2000.

[2] Q. Liu, H. Lu, and S. Ma, "Improving kernel fisher discriminant analysis for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42–49, 2004.

[3] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[4] M. Sharkas and M. A. Elenien, "Eigenfaces vs. fisherfaces vs. ica for Face recognition; a comparative study," in *Proceedings of the 9th International Conference on Signal Processing (ICSP '08)*, pp. 914–919, Beijing, China, October 2008.

[5] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," *Advances in Neural Information Processing Systems*, pp. 556–562, 2000.

[6] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[7] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3608–3614, 2006.

[8] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 174–184, 2010.

[9] W. Bian and D. Tao, "Max-min distance analysis by using sequential SDP relaxation for dimension reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 1037–1050, 2011.

[10] D. Xu, Y. Huang, Z. Zeng, and X. Xu, "Human gait recognition using patch distribution feature and locality-constrained group sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 316–326, 2012.

[11] F. Nie, S. Xiang, and C. Zhang, "Neighborhood minmax projections," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 993–998, 2007.

[12] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1208–1213, Beijing, China, October 2005.

[13] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 905–912, June 2006.

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

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

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[17] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1735–1742, June 2006.

[18] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, pp. 585–591, 2003.

[19] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[20] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Science Department; University of Wisconsin, Madison, Wis, USA, 2008.

[21] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the 16th International Conference on Machine Learning*, pp. 200–209, 1999.

[22] M. M. Adankon, M. Cheriet, and A. Biem, "Semisupervised least squares support vector machine," *IEEE Transactions on Neural Networks*, vol. 20, no. 12, pp. 1858–1870, 2009.

[23] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[24] Y. Song, F. Nie, C. Zhang, and S. Xiang, "A unified framework for semi-supervised dimensionality reduction," *Pattern Recognition*, vol. 41, no. 9, pp. 2789–2799, 2008.

[25] M. Zhao, Z. Zhang, and T. W. S. Chow, "Trace ratio criterion based generalized discriminative learning for semi-supervised dimensionality reduction," *Pattern Recognition*, vol. 45, no. 4, pp. 1482–1499, 2012.

[26] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[27] L. Zheng and T. Li, "Semi-supervised hierarchical clustering," in *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM '11)*, pp. 982–991, Vancouver, Canada, December 2011.

[28] S. C. H. Hoi, W. Liu, and S.-F. Chang, "Semi-supervised distance metric learning for collaborative image retrieval," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–7, Anchorage, Alaska, USA, June 2008.

[29] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2010.

[30] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–7, Rio de Janeiro, Brazil, October 2007.

[31] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection," *IEEE Transactions on Neural Networks*, vol. 19, no. 4, pp. 571–585, 2008.

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

[33] W. Bian and D. Tao, "Biased discriminant euclidean embedding for content-based imageretrieval," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 545–554, 2010.

[34] D. Xu and X. Yan, "Semi-supervised bilinear subspace learning," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1671–1676, 2009.

[35] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[36] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[37] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[38] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, pp. 824–831, August 2005.

[39] M. Hein and M. Maier, "Manifold denoising," in *Proceedings of the Neural Information Processing Systems*, pp. 1–8, 2007.

[40] J. Zhuang, I. W. Tsang, and S. C. H. Hoi, "SimpleNPKL: simple non-parametric kernel learning," in *Proceedings of the 26th International Conference On Machine Learning (ICML '09)*, pp. 1273–1280, June 2009.

[41] B. Borchers, "CSDP, a C library for semidefinite programming," *Optimization Methods and Software*, vol. 11, no. 1, pp. 613–623, 1999.

[42] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.