Optimal Kernel Marginal Fisher Analysis for Face Recognition

Ziqiang Wang
Henan University of Technology, Zhengzhou, 450001, China
Email: wzqagent@126.com

Xia Sun
Henan University of Technology, Zhengzhou, 450001, China
Email: sunxiamail@126.com

***Abstract*—Nonlinear dimensionality reduction and face classifier selection are two key issues of face recognition. In this paper, an efficient face recognition algorithm named OKMFA is proposed. The core idea of the algorithm is as follows. First, the high-dimensional face images are mapped into lower-dimensional discriminating feature space by using the feature vector selection-based optimal kernel marginal Fisher analysis(KMFA), then the multiplicative update rule-based optimal SVM classifier is applied to recognize different facial images herein. Extensive experimental results on two benchmark face databases demonstrate the effectiveness and efficiency of the proposed algorithm.**

***Index Terms*—face recognition, kernel marginal Fisher, support vector machine**

I. Introduction

Face recognition (FR) aims to assist a human expert in determining the identity of a test face. FR has attracted the extensive attention of researchers for more than two decades due to its wide range applications in many fields, such as human–computer interfaces, image and video content analysis, multimedia surveillance, and so on. However, the captured face image data often lies in a high-dimensional space, ranging from several hundreds to thousands. Thus, it is necessary and beneficial to transform the face image data from the original high-dimensional space to a low-dimensional one for alleviating the curse of dimensionality. In the low-dimensional feature space, the traditional classification algorithm can be applied to recognize different face images. As a result, numerous face recognition algorithms have been proposed, and surveys in this area can be found in [1]. How to extract discriminating facial features and how to classify a new face image based on the extracted features are two key issues of all these face recognition algorithms. Therefore, this work also focuses on the issues of feature extraction and classifier selection.

Principal component analysis (PCA) and linear discriminant analysis (LDA) are two well-known feature extraction and dimensionality reduction methods for face recognition[2]. PCA is an orthogonal basis transformation where the new basis is found by an eigen-decomposition of the covariance matrix of a normalized data set, it aims to choose a linear transformation for dimensionality reduction that maximizes the scatter of all projected samples. However, PCA is an unsupervised learning method, it does not utilize the class label information. Thus, features extracted by PCA are optimal for face representation and reconstruction, but not optimal for discriminating one face from others. Unlike PCA, LDA is a supervised method, it aims to find the optimal discriminant vectors by maximizing the ratio of the between-class distance to the within-class distance, thus achieving the maximum class discrimination. The discriminant vectors can be readily computed by applying the eigen-decomposition on the scatter matrices. Due to the utilization of label information, LDA is experimentally reported to outperform PCA for face recognition when sufficient labeled face images are provided. Despite the success of LDA in many pattern classification tasks, it often suffers from the small sample size problem when dealing with high-dimensional face data. Moreover, both PCA and LDA are designed for discovering only the global Euclidean structure, whereas the local manifold structure is ignored. Then, they fail to discover the underlying nonlinear structure as traditional linear methods. One way to handle nonlinear face structure can be provided by using kernel theory[3]. Kernel-based dimensionality reduction methods have been extensively investigated in the literature. For example, PCA is generalized to its kernel version, named as KPCA; kernel discriminant analysis (KDA) utilizes the kernel trick to extend the LDA for handling linearly inseparable classification problems. Although both KPCA and KDA have achieved great success in describing the complexity of face images, they fail to discover the intrinsic structure of face images if they are lying on or close to a submanifold of the ambient space. In fact, in many real-world classifications such as face

Manuscript received September 3, 2011; revised October 17, 2011; accepted October 28, 2011.
Project number: 70701013.
Corresponding author: Ziqiang Wang.

doi:10.4304/jcp.7.9.2298-2305

recognition, the local manifold structure is more important than the global Euclidean structure[4].

To discover the intrinsic manifold structure of the face data, nonlinear dimensional reduction algorithms such as ISOMAP[5], locally linear embedding (LLE)[6] and Laplacian eigenmap (LE)[7] were recently developed. ISOMAP, a variant of MDS, aims to preserve globally the geodesic distances between any pair of data points. The goal of LLE is to discover the nonlinear structure via locally linear reconstructions. LE restates the nonlinear mapping problem as an embedding problem for the vertices in a graph and uses the graph Laplacian to derive a smooth mapping. Although all of these algorithms can discover the intrinsic manifold structure, they are defined on the training data points, and the issue of how to map new test data remains difficult. Therefore, they are not suitable for face recognition. To solve the new test data mapping, He et al.[8] proposed the linear manifold learning algorithm named locality preserving projection (LPP), which is obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace-Betrami operator on the manifold. However, these algorithms are designed to best preserve data locality or similarity in the embedding space rather than good discriminating capability. Therefore, these manifold learning algorithms might not be optimal in discriminating face images with different semantic which is the ultimate goal of face recognition. Later, different manifold learning algorithms have been proposed, Analyses and interpretations about these algorithms are given in view of graph embedding framework [9]. The utility of manifold learning has been demonstrated in many pattern recognition applications. Among these various manifold learning algorithms, the graph embedding framework-based marginal Fisher analysis(MFA) has gained significant popularity due to its solid theory foundation and generalization performance[9,10]. Although MFA seems to be more efficient than other manifold learning algorithms for face recognition, it is still a linear technique in nature. So it is inadequate to describe the complexity of real face images due to high variability of the image content and style. In this paper, we discuss how to perform MFA in the Reproducing Kernel Hilbert Space (RKHS), which gives rise to kernel MFA for facial feature extraction.

As for face recognition, classifier selection is another key issue after facial feature extraction. At present, the nearest neighbor (KNN) algorithm is one of the most widely used classifier algorithms. However, for large face image data sets, the computational demand for classifying face image using KNN can be prohibitive. Until now, many classifier algorithm have been proposed for face recognition, such as nearest feature line(NFL)[11], naïve Bayes, neural network and support vector machine(SVM) [12]. Especially, SVM classifier has a very good performance for pattern classification problems by minimizing the Vapnik-Chervonenkis dimensions. The basic idea behind SVM is to find an optimal hyperplane in a high-dimensional feature space that maximizes the margin of separation between the closest training examples from different classes. Although SVM has achieved great success in many pattern classification tasks, its time complexity is cubic in the number of training points, and is thus computationally inefficient on massive face image data sets. In order to overcome the above shortcomings and fully use its advantages such as higher classification accuracy and better generalization ability, we adopted the multiplicative update rule-based optimal training SVM as face classifier.

In this paper, the objective is to improve face recognition performance by simultaneously using kernel MFA and optimal SVM. The rest of this paper is organized as follows. In section II, we give a brief review of MFA. Section III deals with nonlinear dimensional reduction for face recognition by using the optimal kernel MFA. Section IV discusses the optimal training SVM classifier implement. Experiments are reported in Section V. Finally, we give concluding remarks in Section VI.

II. Brief Review Of MFA

Marginal fisher analysis(MFA)[9] is a recently proposed manifold learning method for feature extraction and dimensionality reduction, it is based on the graph embedding framework and explicitly considers the local manifold structure and class label information with margin criterion. MFA aims to preserve the within-class neighborhood relationship while dissociating the submanifolds for different classes from each other, it has achieved good discriminating performance by integrating the information of intraclass geometry and the interclass discrimination.

Given the face image set $X=[x_1,x_2,\ldots,x_n]$, MFA aims to design an intrinsic graph that characterizes the intraclass compactness and another penalty graph which characterizes the interclass separability. For the intrinsic graph, the intraclass compactness is measured as the sum of distances between each sample and its neighbors within the same class. The formal definition of the intraclass compactness is as follows:

$$\begin{aligned}\tilde{S}_c &= \sum_i \sum_{i\in N_{k_1}(j)\,\text{or}\, j\in N_{k_1}(i)} \left\|W^T x_i - W^T x_j\right\|^2 \\ &= 2W^T X\left(D-S\right)X^T W\end{aligned} \quad (1)$$

$$S_{ij}=\begin{cases}1, & \text{if } i\in N_{k_1}(j)\,\text{or}\, j\in N_{k_1}(i)\\ 0, & \text{otherwise}\end{cases} \quad (2)$$

where S is a similarity matrix defined on the data points in the intrinsic graph, $D_{ii}=\sum_j S_{ij}$, and $N_{k_1}(i)$ denotes the index set of the k_1 nearest neighbors of sample x_i that are in the same class.

For the penalty graph, the interclass separability is measured as the sum of distances between margin points and their neighbor points from different classes. The formal definition of the interclass separability is as follows:

$$\begin{aligned}\tilde{S}_p &= \sum_i \sum_{(i,j)\in P_{k_2}(l(x_i))\text{or}(i,j)\in P_{k_2}(l(x_j))} \left\|W^T x_i - W^T x_j\right\|^2 \\ &= 2W^T X\left(D^p - S^p\right)X^T W\end{aligned} \tag{3}$$

$$S_{ij}^p = \begin{cases}1, & \text{if } (i,j)\in P_{k_2}\left(l\left(x_i\right)\right)\text{or}(i,j)\in P_{k_2}\left(l\left(x_j\right)\right) \\ 0, & \text{otherwise}\end{cases} \tag{4}$$

where S_{ij}^p is a similarity matrix defined on the data points in the penalty graph, $D_{ii}^p = \sum_j S_{ij}^p$, $l(x_i)$ is the class label of data point x_i, and $P_{k_2}\left(l\left(x_i\right)\right)$ is a set of data pairs that are the k_2 nearest pairs among the set $\left\{(i,j)\middle| l\left(x_i\right)\neq l\left(x_j\right)\right\}$.

Performing MFA means minimizing the intraclass compactness $\tilde{S}_c$ and maximizing the interclass separability $\tilde{S}_p$. This is equivalent to minimizing the following objective function:

$$\begin{aligned}W_{MFA} &= \arg\min_W \frac{\tilde{S}_c}{\tilde{S}_p} \\ &= \arg\min_W \frac{W^T X\left(D-S\right)X^T W}{W^T X\left(D^p - S^p\right)X^T W}\end{aligned} \tag{5}$$

Finally, the optimal transformation vectors of MFA are the eigenvectors associated with the smallest eigenvalues of the following generalized eigen-problem:

$$X\left(D-S\right)X^T W = \lambda X\left(D^p - S^p\right)X^T W \tag{6}$$

If $X\left(D^p - S^p\right)X^T$ is nonsingular, then the optimal transformation vectors of MFA can be regarded as the eigenvectors of the matrix $\left(X\left(D^p - S^p\right)X^T\right)^{-1} X\left(D-S\right)X^T$ associated with the smallest eigenvalues.

For face recognition, a problem arises that the matrix $X\left(D^p - S^p\right)X^T$ can not be guaranteed to be nonsingular since the number of training face images is usually much smaller than the dimension of each face image. In this case, we can first apply PCA to remove the components corresponding to zero eigenvalues.

III. Optimal Kernel MFA

Although MFA seems to be more efficient than other dimensionality reduction algorithms for facial feature extraction, it often fails to deliver good performance when face images are subject to complex nonlinear changes due to large pose, expression or illumination variations, for it is a linear method in nature. Therefore, a nonlinear version of MFA is required to classify the face images based on their nonlinear structure in their feature space. Employing a nonlinear face image representation algorithm can result in a reduction of the statistical and the perceptual redundancy among representation elements. One way to handle nonlinear structure can be provided by using kernel theory. Inspired by the success of SVM, we introduce the similar scheme to kernelize the linear MFA. The main idea is to nonlinearly map the face image data into a high-dimensional feature space, and then perform to obtain a semantic manifold in that space. Such a generalization is of great importance since the kernelized MFA would generally achieve better recognition accuracy, and relax the restriction of MFA being only a linear manifold learning algorithm.

The idea of kernel MFA(KMFA) is to solve the problem of MFA in an implicit feature space F which is constructed by the kernel trick[12]. The intuition of kernel trick is to map the input data x from the original feature space into a higher dimensional Hilbert space F constructed by the nonlinear mapping

$$\varphi : x \to \varphi\left(x\right) \in F \tag{7}$$

in which the data may be linearly separable. Then building linear MFA algorithms in the feature space implement nonlinear counterparts in the input data space. The map, rather than being given in an explicit form, is presented implicitly by specifying a kernel function $K\left(,\right)$ as the inner product between each pair of points in the feature space.

$$K\left(x_i, x_j\right) = \left(\varphi\left(x_i\right)\cdot\varphi\left(x_j\right)\right) \tag{8}$$

Performing KMFA means minimizing the intraclass compactness $\tilde{S}_c^\varphi$ and maximizing the interclass separability $\tilde{S}_p^\varphi$ in the feature space F simultaneously. According to (5), this is equivalent to minimizing the following objective function:

$$\begin{aligned}W_{MFA} &= \arg\min_W \frac{\tilde{S}_c^\varphi}{\tilde{S}_p^\varphi} \\ &= \arg\min_W \frac{W^T\varphi\left(X\right)\left(D-S\right)\varphi^T\left(X\right)W}{W^T\varphi\left(X\right)\left(D^p - S^p\right)\varphi^T\left(X\right)W}\end{aligned} \tag{9}$$

where $\varphi\left(X\right) = \left[\varphi\left(x_1\right), \varphi\left(x_2\right), \ldots, \varphi\left(x_n\right)\right]$ denotes the face image data matrix in the feature space F.

Then, the eigenvector problem of MFA in the Hilbert space F can be rewritten as follows:

$$\varphi\left(X\right)\left(D-S\right)\varphi^T\left(X\right)W = \lambda\varphi\left(X\right)\left(D^p - S^p\right)\varphi^T\left(X\right)W \tag{10}$$

where the optimal transformation vectors of KMFA are the eigenvectors associated with the smallest eigenvalues of the generalized eigen-problem (10).

Since the eigenvectors of (10) must lie in the span of all the samples in the feature space F, there exist coefficients $\alpha_i, i = 1, 2, \ldots, n$ such that

$$W = \sum_{i=1}^n \alpha_i \varphi\left(x_i\right) = \varphi\left(X\right)\alpha \tag{11}$$

where $\alpha = \left[\alpha_1, \alpha_2, \ldots, \alpha_n\right]^T$.

By using (11) and (8), we can rewrite (10) as follows:

$$K\left(D-S\right)K^{T}\alpha=\lambda K\left(D^{p}-S^{p}\right)K^{T}\alpha \tag{12}$$

Then, the problem of KMFA is converted into finding the eigenvectors of the matrix $\left(K\left(D^{p}-S^{p}\right)K^{T}\right)^{-1}K\left(D-S\right)K^{T}$ associated with the smallest eigenvalues. For a new face image data x , its projection onto W in the feature space F can be calculated as follows:

$$f\left(x\right)=\left(W\cdot\varphi\left(x\right)\right)=\sum_{i=1}^{n}\alpha_{i}K\left(x_{i},x\right) \tag{13}$$

In fact, matrix $K\left(D-S\right)K^{T}$ and $K\left(D^{p}-S^{p}\right)K^{T}$ is usually singular in face recognition, which stems from the fact that the dimension of the kernel feature space is usually much higher than that of the empirical feature space, a deficiency that is generally known as small sample size (SSS) problem. One possible way to address the SSS problem is by performing PCA projection to reduce the dimension of the feature space and make the two matrixes nonsingular.

According to the above derivation of KMFA, we can observe that different kernel function will produce different implicit kernel feature space. However, how to choose a suitable kernel function for a given application is still an open problem so far. In this research, motivated by the fact the inner product between two vectors can be considered as a similarity representation in the implicit feature space[13], we employ the normalized polynomial kernel function.

$$K\left(x_{i},x_{j}\right)=\frac{k\left(x_{i},x_{j}\right)}{\sqrt{k\left(x_{i},x_{i}\right)k\left(x_{j},x_{j}\right)}} \tag{14}$$

where $k\left(,\right)$ is the polynomial kernel. The degree of the polynomial kernel is set to 2 since it has achieved better performance in many pattern recognition tasks[14,15].

In addition, we can observe that the kernel trick-based KMFA algorithm is computationally expensive in the training phase since its computational complexity is proportional to the number of training points needed to represent the transformation vectors from (11). In fact, the dimensionality of the data subspace spanned by $\varphi\left(x_{i}\right)$ is given by the rank of kernel matrix K , and the $rank\left(K\right) \ll n$ for massive training data set. If we replace n with $rank\left(K\right)$ and select a corresponding subset of feature vectors in the feature space F , which will greatly improve the computational efficiency of KMFA. Based on the above consideration, we adopt the feature vector selection methods[16] to accelerate the running speed of KMFA.

The essential idea of the feature vector selection is to find a subset which is sufficient to express all the data as a linear combination of the selected subset in the feature space F . Let the selected feature vector subset $S=\left\{\varphi\left(x_{s1}\right),\varphi\left(x_{s2}\right),\ldots,\varphi\left(x_{sr}\right)\right\}$ in the feature space F is known, where r denotes the number of selected feature vector, then we can estimate the mapping $\widehat{\varphi}\left(x_{i}\right)$ of any input data x_i as a linear combination of φ_S in the feature space F . The formal description is as follows:

$$\widehat{\varphi}\left(x_{i}\right)=\varphi_{S}\cdot\beta_{i} \tag{15}$$

where $\beta_{i}=\left(\beta_{i}^{1},\beta_{i}^{2},\ldots,\beta_{i}^{r}\right)$ is the coefficient vector.

Then, the goal of feature vector selection is to find the coefficients β_i so that the estimated mapping $\widehat{\varphi}\left(x_{i}\right)$ approaches to the real mapping $\varphi\left(x_{i}\right)$ as far as possible, which can be attained by minimizing the following objective function:

$$\delta_{i}=\frac{\left\|\varphi\left(x_{i}\right)-\widehat{\varphi}\left(x_{i}\right)\right\|^{2}}{\left\|\varphi\left(x_{i}\right)\right\|^{2}} \tag{16}$$

The above optimization problem is performed by setting the partial derivative of δ_i with respect to β_i to zero. By using matrix form, the optimal objective function of (16) can be rewritten as follow:

$$\min\delta_{i}=1-\frac{K_{Si}^{T}K_{SS}^{-1}K_{Si}}{K_{ii}} \tag{17}$$

where K_{SS} is a square matrix of dot products of the selected vectors, and K_{Si} is the vector of dot product between x_i and the selected vector set S .

Then, the ultimate goal of feature vector selection method is make (17) apply to all the sample data, which can be summarized as the following form:

$$\max_{S}J_{S}=\frac{1}{n}\sum_{x_{i}\in X}\left(\frac{K_{Si}^{T}K_{SS}^{-1}K_{Si}}{K_{ii}}\right) \tag{18}$$

The above optimal problem of solution can be obtained with an iterative algorithm[16], and the algorithm stops when K_{SS} is no longer invertible or the predefined number of selected vectors is reached.

IV. OPTIMAL TRAINING SVM CLASSIFIER

Once the discriminating facial features are extracted by KMFA, face recognition becomes a pattern recognition task. Pattern recognition systems employing support vector machine (SVM) have drawn much attention due to its good performance in practical applications and their solid theoretical foundations. The essential idea of SVM is to find a linear separating hyperplane which achieves the maximal margin among different classes of data. Furthermore, one can extend SVM to build nonlinear separating decision hyperplanes by exploiting kernel techniques. Although SVM has achieved great success in many pattern classification tasks, its time complexity is cubic in the number of training points, and is thus computationally inefficient on large scale face image data sets. In order to overcome the above shortcomings and fully use its advantages such as higher classification accuracy and better generalization ability, we adopted the multiplicative update rule-based optimal training SVM as face classifier.

Consider n face data points in the low-dimensional feature space extracted by KMFA that belong to two different classes:

$$\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n} \text{ and } y_i \in \{-1, +1\} \tag{19}$$

where x_i is a low-dimension feature vector and y_i is the label of the class that the vector belongs to. SVM aims to separate the two classes of sample data by finding a hyperplane

$$w^T x + b = 0 \tag{20}$$

where w is the normal vector to the hyperplane and b is the corresponding bias term of the hyperplane.

The optimization objective of SVM is to maximize of the margin $2/\|w\|$ and minimize the training error, which can be formally stated as the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{21}$$

subject to

$$y_i\left[w^T\phi\left(x_i\right) + b\right] \geq 1 - \xi_i, \quad \xi_i > 0 \tag{22}$$

where $\phi(\ )$ is the nonlinear mapping function, C is used to balance the tradeoff between maximizing the margin and minimizing the training error, and ξ_i is the slack variable that quantifies SVM training error.

In the primal form, the Lagrangian of the above SVM optimization problem is as follows:

$$\begin{aligned} L = &\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &-\sum_{i=1}^{n}\alpha_i\left[y_i\left(w^T\phi\left(x_i\right)+b\right)-1+\xi_i\right]-\sum_{i=1}^{n}\lambda_i\xi_i \end{aligned} \tag{23}$$

where the Lagrange multipliers $\alpha_i \geq 0$ and $\lambda_i \geq 0$ for all $i = 1, 2, \ldots, n$.

With Lagrange multipliers and Karush-Kuhn-Tucker(KTT) conditions, the solutions of (21) under constraint condition (22) can be obtained by solving its dual problem:

$$Q(\alpha) = \max_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(x_i, x_j\right) \tag{24}$$

subject to

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, n \tag{25}$$

where $K\left(x_i, x_j\right) = \phi^T\left(x_i\right)\phi\left(x_j\right)$ is a kernel function satisfying Mercer's condition.

Once the optimal α is obtained by solving the quadratic programming (QP) problem of (24), the decision function of SVM classifier is given as follows:

$$\begin{aligned} f(x) &= \operatorname{sgn}\left(w^T\phi(x) + b\right) \\ &= \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K\left(x_i, x\right) + b\right) \end{aligned} \tag{26}$$

Note that the above decision function depends only on the training samples with non-zero Lagrange multipliers α_i. Such training samples are known as the support vectors. Meanwhile, the threshold b is computed by averaging $b = y_j - \sum y_i\alpha_i K\left(x_i, x_j\right)$ over all support vectors $x_j\left(\alpha_j > 0\right)$.

In addition, although the quadratic programming (QP) problem of (24) has the important computational advantage of not suffering from of local minima, given n training samples, the naive implementation of QP solver is of $O\left(n^3\right)$ computational complexity, which is computationally infeasible on very large face image data sets. Hence, a replacement of the naive method for solving QP solutions posed by the SVM classifier is highly desirable. To this end, we applied the multiplicative update rule-based method[17] to improve the training speed of SVM classifier.

Since the optimal problem of (24) can be boiled down to the general nonnegative quadratic programming

$$F(v) = \min_{v}\left(\frac{1}{2}v^T A v + b^T v\right) \text{ subject to } v \geq 0. \tag{27}$$

where matrix A is a symmetric and semipositive definite matrix. Hence, the optimization of objective function $F(v)$ is convex.

Due to the nonnegativity constraints in (27), we adopt the multiplicative iterative updates rule to obtain the optimal solution. The iterative update algorithm is implemented according to the positive and negative components of the matrix A in (27).Their definitions are as follows:

$$A_{ij}^{+} = \begin{cases} A_{ij}, if\ A_{ij} \geq 0. \\ 0\ \ , otherwise. \end{cases} \tag{28}$$

$$A_{ij}^{-} = \begin{cases} \left|A_{ij}\right|, if\ A_{ij} < 0. \\ 0\ \ , otherwise. \end{cases} \tag{29}$$

From the above definition, we can easily observe that $A = A^{+} - A^{-}$. Then, the multiplicative iterative updates rule can be defined as follows in terms of nonnegative matrices.

$$v_i \leftarrow v_i\left[\frac{-b_i + \sqrt{b_i^2 + 4\left(A^{+}v\right)_i\left(A^{-}v\right)_i}}{2\left(A^{+}v\right)_i}\right] \tag{30}$$

The remarkable advantage of the multiplicative iterative update rule in (30) is that it can be easily implemented and never violate the nonnegativity condition constraints. Furthermore, it has been proved that the multiplicative update rule has the correct fixed points[17] and can monotonically improve the optimal objective function of (27). Especially, since the optimal objective function of SVM in (24) is a special case of (27), for the training of SVM with the multiplicative iterative update rule, we can make

$$A_{ij} = y_i y_j K\left(x_i, x_j\right), b_i = -1. \tag{31}$$

Then, the multiplicative update rule for solving the objective function of (27) in SVM can be described the following form:

$$\alpha_i \leftarrow \alpha_i \left[\frac{1 + \sqrt{1 + 4\left(A^{+}\alpha\right)_i \left(A^{-}\alpha\right)_i}}{2\left(A^{+}\alpha\right)_i} \right] \tag{32}$$

where A^{+} and A^{-} are defined as in (28) and (29).

In short, the face recognition procedure has three steps. First, we obtain the face subspace with the optimal manifold learning algorithm KMFA; then the new face image to be identified is projected into the face subspace; finally, the optimal training SVM classifier is adopted to identify the new facial image. The outline of the proposed face recognition algorithm is shown in Figure 1.

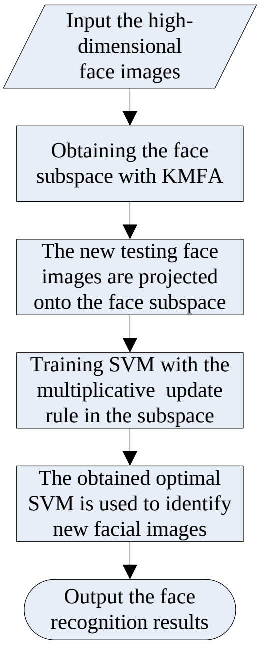

Figure 1. The proposed face recognition algorithm.

V. EXPERIMENTAL RESULTS

In this section, we investigate the performance of our proposed optimal KMFA plus SVM (OKMFA for short) algorithm for face recognition. The algorithm performance is compared with the kernel PCA (KPCA)[18], kernel LDA(KLDA)[19], and kernel LPP(KLPP)[8] algorithms, three of the most popular nonlinear dimensionality reduction algorithms for face recognition. In this experiment, two publicly available benchmark face databases including the FERET standard facial database [20] and Yale database [21] were tested.

Figure 2. The sample image cropped from the face database FERET.

Figure 3. The sample image cropped from the face database Yale.

The FERET face database is a rather larger database. It contains 13,539 face images of 1565 subjects taken during different photo sessions with variations in size, pose, illumination, facial expression and even age. We test the four algorithms on a subset of the FERET database. This subset includes 1,400 images of 200 individuals (each with seven images labeled as ba, bc, bd, be, bf, bg, and bh). Some cropped sample face images from the face database FERET are displayed in Figure 2. All of the gray-level images are aligned by fixing the locations of the two eyes, normalizing in size to a resolution of 32×32 pixels, and preprocessing with histogram equalization. For each individual, *p*(=2,3,4,5) face images are randomly selected for training and the rest are used for testing. To reduce the variation in the recognition results, for each given *p*, we computed the average recognition accuracy of 10 random splits. In general, the performance of the four recognition algorithms varies with the number of dimensions. We only report the best recognition accuracy and the optimal dimensionality obtained by KPCA, KLDA, KLPP, and OKMFA in Table I. As can be seen, our proposed OKMFA algorithm outperforms all the other algorithms with fewer features, and the KPCA algorithm gives relatively poor recognition accuracy.

The Yale face database contains 165 images of 15 individuals (each person has 11 different images) under various facial expressions and lighting conditions. In this experiment, preprocessing to locate the faces was applied. The images were aligned semi-automatically according to the eyes position of each facial image using the eye coordinates. The facial images were cropped, and then re-sized to a resolution of 32×32 pixels. Some cropped sample face images from the face database Yale are displayed in Figure 3. Histogram equalization was used for the normalization of the facial image luminance. A random subset with *p*(=5,6,7,8) face images per individual was taken with labels to form the training set, and the rest of the database was regarded as the testing set. For each given *p*, we average the results over 10 random splits. We only report the best recognition accuracy and the optimal dimensionality obtained by KPCA, KLDA, KLPP, and OKMFA in Table II. We can see that our proposed OKMFA algorithm achieves the best recognition accuracy.

In addition, to verify the efficiency of our proposed OKMFA algorithms, we record the computational time in the experiments. The running times of the four algorithms on the FERET and Yale face database are listed in Table III and Table IV respectively. As can be seen, the time ration for the FERET face database used by the four algorithms are approximately KPCA: KLDA: KLPP: OKMFA = 29:31:26:21.The time ration for the Yale face database used by the four algorithms are approximately KPCA: KLDA: KLPP: OKMFA = 17:19:23:15. These results show that the proposed OKMFA algorithms are much more efficient than the traditional kernel-based dimensionality reduction algorithms. The main reason could be attributed to the fact that the feature vector selection strategy accelerates the running speed of KMFA, and the multiplicative update rule-based method further improve the training speed of SVM classifier. Therefore, our proposed OKMFA algorithm could dramatically reduce the computational time when compared to other three algorithms on large scale face recognition problem.

TABLE I.
RECOGNITION ACCURACY COMPARISONS ON THE FERET DATABASE

Algorithms	2 images	3 images	4 images	5 images
KPCA	74.6%(79)	82.7%(81)	88.6%(56)	91.8%(48)
KLDA	76.5%(42)	84.6%(46)	90.3%(40)	93.4%(69)
KLPP	79.1%(45)	87.9%(50)	91.5%(42)	94.2%(60)
OKMFA	84.7%(40)	92.8%(44)	93.2%(40)	96.7%(46)

TABLE II.
RECOGNITION ACCURACY COMPARISONS ON THE YALE DATABASE

Algorithms	5 images	6 images	7 images	8 images
KPCA	56.3%(72)	64.7%(78)	70.5%(70)	78.2%(74)
KLDA	68.5%(14)	72.6%(14)	81.3%(14)	89.7%(14)
KLPP	79.8%(12)	84.1%(14)	91.6%(14)	95.8%(13)
OKMFA	85.6%(10)	92.5%(12)	95.7%(14)	98.3%(14)

TABLE III.
RUNNING TIME COMPARISONS ON THE FERET DATABASE

Algorithms	Training time(s)	Testing time(s)
KPCA	26.1	3.2
KLDA	28.7	2.5
KLPP	23.8	2.1
OKMFA	19.6	1.4

TABLE IV.
RUNNING TIME COMPARISONS ON THE YALE DATABASE

Algorithms	Training time(s)	Testing time(s)
KPCA	14.9	2.2
KLDA	17.5	1.6
KLPP	21.4	1.8
OKMFA	13.8	1.3

In summary, the main observations from the above performance comparisons include:

(1) Our proposed OKMFA algorithm consistently outperforms KPCA, KLDA, and KLPP algorithms in terms of recognition accuracy and computational time. The superiority of OKMFA stems from two aspects: on the one hand, the kernel MFA explicitly considers the local manifold structure and class label information with margin criterion in the process of nonlinear dimensionality reduction, thus achieving maximum discrimination and improving the computation efficiency of face classifier; one the other hand, the multiplicative updates rule-based optimal SVM can achieve better classification performance and lower computational requirements simultaneously. Hence, the proposed OKMFA algorithm achieved much better performance than other three algorithms.

(2) The manifold learning-based (such as KLPP and OKMFA) algorithms achieve much better performance than KPCA and KLDA, which demonstrates the importance of utilizing local manifold structure. The reason is that KPCA and KLDA are designed for discovering only the global Euclidean structure, whereas the local manifold structure is ignored.

(3) Although OKMFA and KLPP algorithms belong to manifold learning algorithms, our proposed OKMFA algorithm performs better than KLPP. One possible explanation is as follows: although KLPP seeks to preserve local neighbor structure, it does not explicitly exploit the class information for classification. By jointly considering the local manifold structure and the class label information with two graphs, OKMFA achieves much better performance than KLPP in face recognition.

(4) The KPCA algorithm gives the worst recognition accuracy. One possible explanation is as follows: KPCA is an unsupervised algorithm that ignores the valuable label information for classification. Hence, the features extracted by KPCA are optimal for representation, but not optimal for classification.

VI. CONCLUSIONS

In this paper, we have proposed an enhanced face recognition algorithm called OKMFA that combines the advantages of optimal kernel MFA and SVM. The effectiveness and efficiency of this algorithm is evidenced in experimental comparisons on two benchmark face databases with other well-known kernel-based dimensionality reduction algorithms. In the future, we would like to apply our algorithm to many tasks in pattern recognition, data mining, and high-dimensional data processing.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No.70701013, the Natural Science Foundation of Henan Province under Grant No. 102300410020,0611030100, and the Natural Science Foundation of Henan University of Technology under Grant No. 08XJC013 and 09XJC016.

REFERENCES

[1] S.R.Chellappa and C.L.Wilson, "Human and machines recognition of faces: a survey," *Proceedings of IEEE*, vol.83, pp.705–741, May 1995.
[2] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Second edition, Hoboken: Wiley-Interscience,2000.
[3] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge, UK: Cambridge University Press, 2004.
[4] 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,pp.328-340, March 2005.
[5] J.B.Tenenbaum,V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol.290, pp.2319-2323, December 2000.
[6] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*,vol.290, pp.2323-2326, December 2000.
[7] M. Belkin and P.Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol.15, pp.1373-1396, June 2003.
[8] X.He and P.Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, Cambridge, MA:MIT Press,2003,pp.385-392.
[9] 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, pp.40-51, January 2007.
[10] D.Xu, S.Yan, D.Tao, S.Lin, and H.-J.Zhang, "Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Transactions on Image Processing*, vol.16, pp.2811-2821, November 2007.
[11] S.Z.Li and J.W.Lu, "Face recognition using the nearest feature line method," *IEEE Transactions on Neural Networks*, vol.10, pp.439–443, March 1999.
[12] V. N.Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag,1995.
[13] N. Cristianini, J. S. Taylor, and J. Kandola, "Spectral kernel methods for clustering," in *Advances in Neural Information Processing Systems*, Cambridge, MA:MIT Press,2001,pp.649–655.
[14] 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, pp.42-49, January 2004.
[15] H.Li, T.Jiang, and K.Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*,vol.17, pp.157-165, January 2006.
[16] G. Baudat and F. Anouar, "Kernel-based methods and function approximation," in *Proceedings of International Conference on Neural Networks*, Piscataway, NJ:IEEE Press,2001,pp.1244–1249.
[17] F.Sha, Y.Q.Lin, L.K.Saul, and D.D.Lee, "Multiplicative updates for nonnegative quadratic programming," *Neural Computation*, vol.19, pp.2004-2031, August 2007.
[18] M. H. Yang, N.Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces," in *Proceedings of International Conference on Image Processing*, Piscataway, NJ:IEEE Press,2000,pp. 37-40.
[19] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol.12, pp. 2385–2404, October 2000.
[20] P. J. Phillips, H.Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing*, vol. 16, pp. 295–306, April 1998.
[21] Yale Face Database, http://cvc.yale.edu/projects/ yalefaces/ yalefaces.html, 2002.

Ziqiang Wang was born Zhoukou, Henan, China in 1973. He received the PhD degree from China University of Mining &Technology (Beijing) and the Master of Science degree from Xi'an Petroleum University, both in computer application, in 2011 and 2002, respectively. His research interests include data mining and pattern recognition.

He is currently an associate professor in the Henan University of Technology.

Xia Sun was born Xi'an, Shanxi, China in 1978. She received the Master of Science degree in computer application from Huazhong University of Science and Technology in 2008. Her research interests include data mining and pattern recognition.

She is currently a lecturer in the Henan University of Technology.