KERNEL GENERALIZED NONLINEAR DISCRIMINANT ANALYSIS ALGORITHM FOR PATTERN RECOGNITION

Guang Dai [1], *Yuntao Qian* [2]
[1] College of Information Science & Engineering, Wenzhou University, P.R.China
[2] College of Computer Science, Zhejiang University, P.R.China

ABSTRACT

Linear discriminant analysis (LDA) is a very effective tool used for dimensionality reduction and feature extraction in pattern recognition. However, the LDA is inadequate to describe complex and nonlinear patterns. To solve this problem, kernel nonlinear discriminant analysis (K-NDA) has been proposed. Although successful in many cases, classic K-NDA also suffers from the small sample size problem (SSSP) and loses some significant discriminatory information as same as classic LDA. In this paper, a novel K-NDA, i.e., the kernel generalized nonlinear discriminant analysis (KG-NDA) algorithm is introduced to effectively overcome these problems, and it also views the optimal discriminant vectors as a global transform in the feature space to some extent. It not only deals with the nonlinear problem, but also effectively solves the SSSP. The KG-NDA is applied to the experiments on face recognition, and the results tested on two popular databases demonstrate that this method is very effective.

1. INTRODUCTION

Techniques that can introduce low-dimensional feature representation with enhanced discriminatory power are of paramount important in many pattern recognitions. Linear discriminant analysis (LDA) is a classic tool used to reduce the dimensionality and extract the feature in pattern recognition. However, traditional LDA suffers from the small sample size problem (SSSP) that widely exists in high-dimensional pattern recognition tasks, where the number of available samples is smaller than the dimensionality of the samples. The traditional solution to the SSSP [2], i.e., PCA+LDA, will discard some significant discriminatory information. To effectively void the loss, solutions without a separate PCA step, called direct LDA (D-LDA) methods [3,6] have been presented recently. Although successful in many cases, they fail for a nonlinear problem and are inadequate to describe a complex and nonlinear pattern as same as other LDA. Hence, it is reasonable to assume that a better solution to an inherent nonlinear problem can be achieved using nonlinear methods.

In order to overcome the limitation associated with classic LDA, several nonlinear discriminant algorithms [8-11], called kernel nonlinear discriminant analysis (K-NDA) algorithms, have been proposed in recent years, and these algorithms have employed a technique referred to as the kernel trick that was first used in support vectors machine (SVM) to find an optimal separating hyperplane. The basic idea of K-NDA is to first map the input data x into a feature space F via a nonlinear mapping ϕ and then perform a LDA in the feature space F by the kernel trick. Although many K-NDA algorithms have been shown a better performance for a nonlinear problem, they also have two common limitations: 1) in the case of the SSSP, these algorithms discard some significant discriminatory information; 2) these algorithms only extract at most $c-1$ meaningful features, where c is the number of classes invoked.

In this paper, a novel kernel algorithm, i.e., the kernel generalized nonlinear discriminant analysis (KG-NDA) algorithm is introduced to extract the nonlinear feature for pattern recognition. This algorithm not only breaks the inherent limitations of classic K-NDA, but also considers its separability from a global viewpoint in the feature space F to some extent. It has the properties: 1) in the subspace spanned by the first $c-1$ optimal discriminant vectors in the feature space F, the loss of significant discriminatory information in classic K-NDA is completely avoided; 2) in the subspace spanned by the remaining optimal discriminant vectors in the feature space F, some other discriminatory information is obtained by considering its separability from a global view point. The KG-NDA is applied to face recognition, where the SSSP is often encountered and the pattern distribution is generally nonlinear and complex, and the comparative results tested on two popular databases demonstrate that the performance of the KG-NDA is overall superior to those of the existing approaches.

2. THE TRADITIONAL ALGORITHMS FOR KERNEL NONLINEAR DISCRIMINANT ANALYSIS (K-NDA)

For solving nonlinear problems, classic LDA has been generalized to its nonlinear version by the kernel trick, namely K-NDA [8-11]. Let $\phi: x \in R^n \to \phi(x) \in F$ be a nonlinear mapping from the input space to a high-dimensional feature space F with linearly separable properties. The basic idea of K-NDA is to first map the input data x into a feature space F via the mapping ϕ and then performs a LDA in F. However, it is unnecessary to compute explicitly in F but compute the inner product of two vectors in F with an inner product kernel function:

$$k(x,y) = (\phi(x)^T \cdot \phi(y)) . \tag{1}$$

Let $x_i (i = 1,\ldots,N)$ be a vector of the training sample set X with N elements. X_i is subjects of X with N_i elements, and $X = \bigcup_{i=1}^{c} X_i$, $N = \sum_{i=1}^{c} N_i$, where c is the number of classes invoked. The between-class scatter matrix S_b, within-class scatter matrix S_w, and population scatter matrix S_t in F can be expressed as follows, respectively:

$$S_b = \sum_{i=1}^{c} (N_i / N)(m_i - m)(m_i - m)^T \tag{2}$$

$$S_w = (1/N) \cdot \sum_{i=1}^{c} \sum_{x_j \in X_i} (\phi(x_j) - m_i)(\phi(x_j) - m_i)^T \tag{3}$$

$$S_t = S_b + S_w = (1/N) \cdot \sum_{i=1}^{N} (\phi(x_i) - m)(\phi(x_i) - m)^T \tag{4}$$

where $m_i = (1/N_i) \cdot \sum_{x_j \in X_i} \phi(x_j)$ denotes the sample mean of class X_i in F; $m = (1/N) \cdot \sum_{i=1}^{N} \phi(x_i)$ denotes the all sample mean in F. To calculate the optimal discriminant vectors in F, we need to maximize:

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{5}$$

The general algorithms calculating (5) are to utilize the theory of reproducing kernel. By the theory of reproducing kernel, w can be expressed [8,9]:

$$w = \sum_{i=1}^{N} \alpha_i \phi(x_i) \tag{6}$$

Inserting (6) into (5), then (5) becomes a function of $\alpha = (\beta_1, \ldots, \beta_N)$ and can be converted to maximize:

$$J_f(\alpha) = \frac{\alpha^T K_b \alpha}{\alpha^T K_w \alpha} \tag{7}$$

where K_b and K_w can be calculated by the kernel trick, and the detailed procedure can be found in [9].

Then, the maximum criterion $J_f(\alpha)$ can be formed by the leading eigenvectors of $K_w^{-1} K_b$. However, in the case of the SSSP, K_w is singular and is thus not invertible in practice. There are currently two methods used to circumvent this problem. The first method is to replace the inverse matrix K_w^{-1} with a pseudoinverse matrix of K_w, such as generalized discriminant analysis (GDA) [9]. However, this method tends to overfit the training set in some cases. The second method [8,11] will introduce a nonsingular matrix $\tilde{K}_w = K_w + \tau I$ to replace K_w, where $\tau > 0$, and I is the identity matrix. However, these algorithms above have two limitations: 1) they discard the null space of K_w, where may contain the some significant discriminatory information; 2) since the rank of K_b is no more than $c-1$, they only extract at most $c-1$ meaningful features.

Most recently, a novel method, which effectively overcomes the limitation of GDA [9] that the pseudoinverse matrix causes the loss of some significant discriminatory information, has been proposed [10]. This method, i.e., the kernel direct discriminant analysis (KDDA) [10], directly carries out the algorithm of D-LDA of Yu et al [3] in F. As same as the D-LDA [3], the behind idea of the KDDA believes that the significant discriminatory information in F only exists in the intersection space $(S_w(0) \bigcap S_b^{-1}(0))$, and it intends to seek this intersection space, where $S_w(0) = \{x \mid S_w x = 0\}$, and $S_b^{-1}(0) = \{x \mid S_b x \neq 0\}$. To obtain this intersection space, the KDDA first calculates the $c-1$ leading eigenvectors of S_b in F to obtain the space $S_b^{-1}(0)$, and then calculates the eigenvalues and corresponding eigenvectors of $\tilde{S}_W$ to obtain the space $S_w(0)$, where $\tilde{S}_W$ is the projection of S_w in the space $S_b^{-1}(0)$. From the procedure of the KDDA method, it is clear that $R(S_b^{-1}(0)) \leq c-1$, and $R(S_w(0) \bigcap S_b^{-1}(0)) < c-1$, where $R(\bullet)$ denotes the dimensionality of ' $\bullet$ 'obtained in the KDDA. However, according to the analysis in [6], it is obvious that the dimensionality of the intersection space $S_w(0) \bigcap S_b^{-1}(0)$ equals to $c-1$ in essence, since the dimensionality of F is far higher than the number of the training samples in F in the case of the SSSP. As a result, the KDDA has the same limitations described in the algorithms above for the K-NDA.

Hence, many existing KNDA-based algorithms have two common limitations: 1) in the case of the SSSP, they lose some significant discriminatory information; 2) they only extract at most $c-1$ meaningful features.

3. KERNEL GENERALIZED NONLINEAR DISCRIMINANT ANALYSIS (KG-NDA) ALGORITHM

In this section, a novel kernel generalized nonlinear discriminant analysis (KG-NDA) algorithm is introduced to effectively overcome the limitations of the previous algorithms for the K-NDA.

As same as some LDA-based algorithms, the KG-NDA algorithm will replace the conventional Fisher's criterion in (5) with a modified Fisher's criterion, and this modified criterion has been proven to be equivalent to the conventional one [5]. It can be expressed as:

$$J(Y) = \frac{Y^T S_b Y}{Y^T S_t Y} \tag{8}$$

Then, we believe that the optimal discriminant vectors in F can be calculated in $S_t^{-1}(0)$, where $S_t^{-1}(0) = \{x \mid S_t x \neq 0\}$. To see that this conclusion is valid, the feature space F can be split into the null space $S_t(0) = \{x \mid S_t x = 0\}$ and its orthogonal complement space $S_t^{-1}(0)$. Let $Y = Y_1 + Y_2$, and it is clear that $Y_2^T S_t Y_2 = Y_2^T S_b Y_2 = 0$, where $Y_1 \in S_t^{-1}(0)$ and $Y_2 \in S_t(0)$. Hence, it is clear that $J(Y) = J(Y_1)$ in (8). As a result, the optimal discriminant vectors in F can be calculated in $S_t^{-1}(0)$ without any loss of the discriminatory information.

To obtain the optimal discriminant vectors in $S_t^{-1}(0)$, we need to calculate the orthonormal bases of $S_t^{-1}(0)$. Then, S_t in (4) can be rewritten here as follows:

$$S_t = \sum_{i=1}^{N} \overline{\phi}(x_i) \overline{\phi}(x_i)^T = \Phi_t \Phi_t^T \tag{9}$$

where $\overline{\phi}(x_i) = \sqrt{1/N} \cdot (\phi(x_i) - m)$, $\Phi_t = [\overline{\phi}(x_1), \ldots, \overline{\phi}(x_N)]$

As similar as the KDDA [10], the orthonormal bases of $S_t^{-1}(0)$ can be obtained by calculating the corresponding orthonormal eigenvectors of all positive eigenvalues of S_t. Since the dimensionality of F, denoted as N', could be arbitrarily large or possible infinite, it is intractable to directly compute those orthonormal eigenvectors of the $N' \times N'$ matrix S_t. Fortunately, as described in [1,3,10], those orthonormal eigenvectors can be indirectly derived from the eigenvectors of $\Phi_t^T \Phi_t$ (with size $N \times N$).

For all training samples $\{\phi(x_i)\}_{i=1}^{N}$ in F, we can define a $N \times N$ kernel matrix K as follows:

$$K = (k_{i,j})_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}} \tag{10}$$

where $k_{i,j} = \phi(x_i)^T \phi(x_j)$. Hence, by the kernel trick, $\Phi_t^T \Phi_t$ can be expressed as follows:

$$\Phi_t^T \Phi_t = \frac{1}{N}(K - \frac{1}{N}(K \cdot I_{N\times N} + I_{N\times N} \cdot K) + \frac{1}{N^2} I_{N\times N} \cdot K \cdot I_{N\times N}) \tag{11}$$

where $I_{N\times N}$ is the $N \times N$ matrix with all terms being one. Let λ_i and e_i ($i = 1, \ldots, m$) be the i -th positive eigenvalue and corresponding eigenvector of $\Phi_t^T \Phi_t$. According to [1,3,10], it is clear that $v_i = \Phi_t e_i \lambda_i^{-1/2}$ ($i = 1, \ldots, m$) constitute the orthonormal bases of $S_t^{-1}(0)$. Then, the optimal discriminant vectors in F are transformed and calculated in the projective space R^m of $S_t^{-1}(0)$. Then, $J(Y)$ in (8) can be rewritten in the projective space R^m:

$$J(U) = \frac{U^T \tilde{S}_b U}{U^T \tilde{S}_t U} \tag{12}$$

In addition, $\tilde{S}_b = V^T S_b V$, $\tilde{S}_w = V^T S_w V$, and $\tilde{S}_t = V^T S_t V$ are the projection of S_b S_w and S_t in $S_t^{-1}(0)$, respectively. It is obvious that $\tilde{S}_b$, $\tilde{S}_w$ is semi-positive definite and $\tilde{S}_t$ is positive definite. There are two approaches to calculate these scatter matrices in the projective space R^m.

A. By mapping all the training samples $\{\phi(x_i)\}_{i=1}^{N}$ in F into $S_t^{-1}(0)$, the corresponding training samples $\{y_i\}_{i=1}^{N}$ in the projective space R^m can be obtained by the kernel trick as follows:

$$y_i = V^T \phi(x_i) = \sqrt{\frac{1}{N}} E^T (k_{i,1} - \frac{1}{N}\sum_{j=1}^{N} k_{i,j}, \ldots, k_{i,N} - \frac{1}{N}\sum_{j=1}^{N} k_{i,j})^T \tag{13}$$

where $V = [v_1, \ldots, v_m]$, $E = (e_1 \lambda_1^{-1/2}, \ldots, e_m \lambda_m^{-1/2})$, $y_i \in R^m$. According to the definition of scatter matrices, $\tilde{S}_b, \tilde{S}_w, \tilde{S}_t$ can be

calculated using the $\{y_i\}_{i=1}^{N}$. It should be noted at this point, by a similar approach, a very complex procedure of eigen-analysis of scatter matrices in the KDDA algorithm would be very simply.

B. Using the kernel trick and the kernel matrix K in (10), $\tilde{S}_b$ $\tilde{S}_w$ and $\tilde{S}_t$ are directly calculated as similar as the KDDA algorithm: $\tilde{S}_{b_r} = V^T S_b V = E^T \Pi^T \Pi E$, $\tilde{S}_w = V^T S_w V = E^T \Xi^T \Xi E$, and $\tilde{S}_t = V^T S_t V = diag(\lambda_1, \ldots \lambda_m)$,

where

$$\Pi^T = K \cdot \mathrm{T}_{N\times m} / N - K \cdot \Gamma_{N\times m} / N^2 - \mathrm{I}_{N\times N} \cdot K \cdot \mathrm{T}_{N\times m} / N^2 + \mathrm{I}_{N\times N} \cdot K \cdot \Gamma_{N\times m} / N^3$$

$$\Xi^T = K / N - \mathrm{I}_{N\times N} \cdot K / N^2 - K \cdot \mathrm{A}_{N\times N} / N + \mathrm{I}_{N\times N} \cdot K \cdot \mathrm{A}_{N\times N} / N^2$$

$E = (e_1 \lambda_1^{-1/2}, \ldots, e_m \lambda_m^{-1/2})$,

where $\mathrm{A}_{N\times N} = diag(A_1, \ldots, A_m)$, A_i is the $N_i \times N_i$ matrix with all terms being $1/N_i$, $\mathrm{T}_{N\times m} = diag(\mathrm{T}_1, \ldots, \mathrm{T}_m)$, $\Gamma_{N\times m} = (\Gamma_1, \ldots, \Gamma_m)$, T_i is a $1 \times N_i$ vector with all terms being $1/\sqrt{N_i}$, Γ_i is a $1 \times N$ vector with all terms being $\sqrt{N_i}$.

Then, we will carry out the generalized optimal discriminant vectors (GODV) algorithm of Guo et al [5] to calculate the optimal discriminant vectors with respect to Fisher criterion (12). The GODV method of Guo et al [5] is the extension of the GODV method of Liu et al [4], and it can effectively prevent some GODV, which are meaningless for classification, from being calculated in the GODV method of Liu et al [4]. It can view the optimal set of discriminant vectors as a global transform and consider its separability from a global viewpoint. According to the GODV method [5], $\tilde{S}_w$ can be split into its null space $\tilde{S}_w(0) = span\{\gamma_1, \ldots, \gamma_l\}$ and its orthogonal complement space $\tilde{S}_w^{-1}(0) = span\{\gamma_{l+1}, \ldots, \gamma_m\}$, where $\gamma_1, \ldots, \gamma_m$ are the orthonormal bases of R^m. In fact, it can be verified that all discriminatory information with respect to Fisher criterion (12) is contained in these two subspaces [13]. It is clear that the within-class distance equals to zero in $\tilde{S}_w(0)$, and the between-class distance equals to nonzero in $\tilde{S}_w^{-1}(0)$. Hence, in $\tilde{S}_w(0)$, the Fisher criterion (12) can be replaced by $\hat{J}(U) = U^T \tilde{S}_b U$. To calculate the optimal discriminant vectors in $\tilde{S}_w(0)$, let $P_1 = [\gamma_1, \ldots, \gamma_l]$ and $\overline{S}_b = P_1^T \tilde{S}_b P_1$, calculate $\overline{S}_b$'s orthonormal eigenvectors $z_1, \ldots, z_l$. It is easy to be seen that $P_1 z_i (i = 1, \ldots, l)$ constitute all optimal discriminant vectors in $\tilde{S}_w(0)$ and $VP_1 z_i (i = 1, \ldots, l)$ constitute all optimal discriminant vectors in the intersection space $(S_w(0) \bigcap S_b^{-1}(0))$ (in fact, the space $(S_w(0) \bigcap S_b^{-1}(0))$ is equivalent to the space $(S_w(0) \bigcap S_t^{-1}(0))$). It is obvious that l must be $c-1$, since the dimensionality of the feature space F is far higher than the number of the training samples in the feature space F [6]. It is clear that the loss of the discriminatory information of those traditional KNDA-based algorithms can be effectively avoided here. In addition, according to the GODV method [5], the remaining optimal discriminant vectors with respect to Fisher criterion (12) can be calculated in $\tilde{S}_w^{-1}(0)$. Let $P_2 = (\gamma_{l+1}, \ldots, \gamma_m)$ and $\hat{S}_b = P_2^T \tilde{S}_b P_2$, $\hat{S}_t = P_2^T \tilde{S}_t P_2$, those optimal discriminant vectors can be calculated by the GODV method of Liu et al [4] as follows:

(1) z_{l+1} is the unit vector, which maximizes:

$$J_{l+1}(z) = \frac{\sum_{i=1}^{l} z_i^T P_1^T \tilde{S}_b P_1 z_i + z^T \hat{S}_b z / \| z \|}{\sum_{i=1}^{l} z_i^T P_1^T \tilde{S}_t P_1 z_i + z^T \hat{S}_t z / \| z \|} \tag{14}$$

(2) $(z_j)_{j=l+2}^{l'}$ are the unit vectors, which can be calculated by optimizing problem:

$$\max_{Z_i^T Z = 0, \|Z\| = 1} \{J_j(z)\}, \quad i = l+1, \ldots j-1$$

$$J_j(z) = \frac{\sum_{i=1}^{l} z_i^T P_1^T \tilde{S}_b P_1 z_i + \sum_{i=l+1}^{j-1} z_i^T \hat{S}_b z_i + z^T \hat{S}_b z / \| z \|^2}{\sum_{i=1}^{l} z_i^T P_1^T \tilde{S}_t P_1 z_i + \sum_{i=l+1}^{j-1} z_i^T \hat{S}_t z_i + z^T \hat{S}_t z / \| z \|^2} \tag{15}$$

Then, $P_2 z_i (i = l+1, \ldots, l')$ constitute the optimal discriminant vectors in $\tilde{S}_w^{-1}(0)$ and $VP_2 z_i (i = l+1, \ldots, l')$ constitute the remaining optimal discriminant vectors in F. It is clear that the $VP_2 z_i (i = l+1, \ldots, l')$ constitute the optimal discriminant vectors of $\tilde{S}_b^{-1}(0) \bigcap \tilde{S}_w^{-1}(0)$. From the procedure above, we can see that $Y_i = VP_1 z_i (i = 1, \ldots, l)$ and $Y_i = VP_2 z_i (i = l+1, \ldots, l')$ constitute all optimal discriminant vectors in F.

From the procedure of calculating the discriminant vectors in F and the GODV method [5], it is obvious that the optimal discriminant vectors have been considered from a global viewpoint in F, since the discriminant vectors are calculated in $S_t^{-1}(0)$ without any loss of discriminatory information. In addition, the loss of significant discriminatory information in the traditional K-NDA methods can be completely avoided in this method. We call this method acquired the kernel generalized nonlinear discriminant vectors (KG-NDA) method.

For an input pattern x, its projection into the subspace spanned by $\Theta = [Y_1, \ldots, Y_{l'}]$, can be calculated by $z = \Theta^T \phi(x)$, and this expression can be rewritten by the kernel trick as follows:

$$z = (P_1 z_1, \ldots, P_1 z_l, P_2 z_{l+1}, \ldots, P_2 z_{l'})^T$$
$$\bullet \sqrt{\frac{1}{N}} E^T (k(x, x_1) - \frac{1}{N} \sum_{i=1}^{N} k(x, x_i), \ldots, k(x, x_N) - \frac{1}{N} \sum_{i=1}^{N} k(x, x_N))^T \tag{16}$$

where $E = (e_1 \lambda_1^{-1/2}, \ldots, e_m \lambda_m^{-1/2})$.

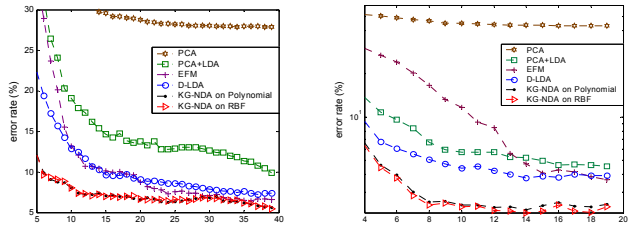
Thus, a low-dimensional representation with enhanced discriminant power by the KG-NDA method has been introduced.

4. EXPERIMENTAL RESULTS

We will assess the feasibility and performance of the KG-NDA on the face recognition task, using the ORL and the UMIST databases. The ORL database is composed of 400 images with ten different images for each of the 40 distinct subjects. The variations of the images are across pose, size, time, and facial expression. The UMIST database is a multiview database, consisting of 575 gray-scale images of 20 subjects, each coving a wide range of poses from profile to frontal view as well as race, gender and appearance. The spatial and grey-level resolutions of all images are scaled into 92x112 and 256, respectively.

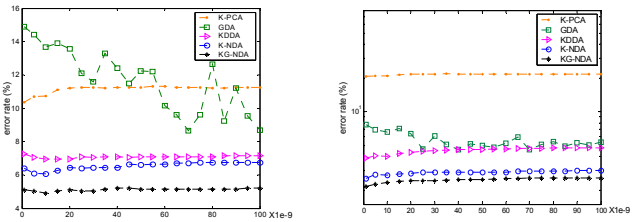
In the following experiments, each one of the two databases is randomly divided into a training set and a test set, and there is no overlapping between the two. For the ORL database, five images are randomly chosen from the ten images available for each subject for training, while the remaining five images are used for testing. For the UMIST database, five images are randomly chosen from the images available for each subject for training, while the remaining images are used for testing. The nearest neighbor classifier is used for classification. In addition, each experiment is repeated ten times and the results reported in this paper are an average of them.

The first experiment will compare the KG-NDA with the linear feature extraction approaches, including: PCA [1], PCA+LDA [2], EFM [12], D-LDA [3]. For the sake of simplicity, we only implement the KG-NDA with the polynomial kernel $(k(z_1, z_2) = ((z_1^T \cdot z_2) \cdot 1e-9 + 1)^2)$ and the RBF kernel $(k(y_1, y_2) = \exp((-\| y_1 - y_2 \|^2)/1e9)$. Fig.1 (a) and Fig.1 (b) depict the error rates as functions of the number of feature vectors within the range from 5 to 39 on the ORL database and within the range from 5 to 19 on the UMIST database, respectively. According to Fig.1, it is clear that the performance of the KG-NDA is overall superior to those of the four linear methods. In addition, by comparing Fig.1(a) with Fig.1(b), it is clear that the KG-NDA is more effective on the UMIST database than on the ORL database, since the face patterns in the UMIST database are subject to larger variations.

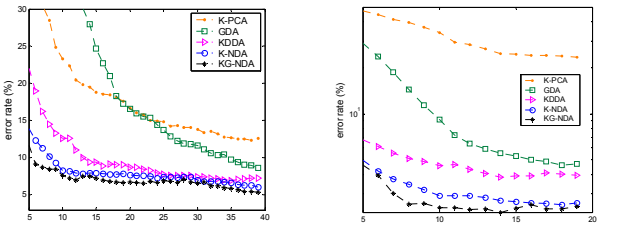


(a):ORL (b):UMIST

Fig.1. Comparison of error rates of the KG-NDA method and the various linear methods.



(a):ORL (b):UMIST



(c):ORL (d):UMIST

Fig.2. Comparison of error rates based on Polynomial kernel function. (a), (b) error rates as functions of a . (c), (d) error rates on the optimal parameter a .

The second experiment will compare the KG-NDA with the nonlinear approaches by the kernel trick, including: K-PCA [7], GDA [9], KDDA [10], and K-NDA [11]. In addition, we only carry out these five methods with the polynomial kernel $(k(z_1,z_2)=((z_1^T\cdot z_2)\cdot a+b)^d)$. As similar as the KDDA, for the sake of simplicity, we only discuss the influence of a , while $b=1$, and $d=2$ are fixed. Fig.2(a) and Fig.2(b) depict the error rates as functions of a within the range from $1e-9$ to $1e-7$ on the optimal number of feature vectors, which can be found by searching the number of used feature vectors that leads to the minimum summation of the error rate over the variation range of a [10]. Fig.2(c) and Fig.2(d) depict the error rates as functions of the number of feature vectors within the range from 5 to 39 on the ORL database and within the range from 5 to 19 on the UMIST database respectively, where the various methods are carried out on the polynomial kernels $(k(z_1,z_2)=((z_1^T\cdot z_2)\cdot a+1)^2)$ with the optimal parameter a . In Fig.2(a) and Fig.2(b), the optimal numbers of feature vectors in the GDA, the KDDA and the K-NDA are about 39 on the ORL database and 19 on the UMIST database, while the optimal numbers of feature vectors in both the KG-NDA and the K-PCA are often beyond 39 on the ORL database and 19 on the UMIST database. As a result, it can conclude that the KG-NDA can obtain some discriminatory information contained in $S_b^{-1}(0)\bigcap S_w^{-1}(0)$, and it is often omitted in the previous methods. In addition, the K-PCA only achieves simply object reconstruction in F , so it always has highest error rates shown in Fig.2 on both databases. From Fig.2, we can see that the performance of the KG-NDA is overall superior to those of the other four nonlinear methods, and it can effectively compensate the limitations and shortcomings of these methods. In addition, it is worthy to mention here that the computational requirements of the novel algorithm are tolerable to those of other nonlinear methods.

5. CONCLUSIONS

A novel K-NDA method, i.e., the KG-NDA has been introduced in this paper. It not only deals with a nonlinear problem, but also effectively solves the SSSP. In addition, it also views the optimal discriminant vectors as a global transform in the feature space to some extent. We apply this method to extract the nonlinear feature for face recognition, where the SSSP widely exists and the pattern distribution is generally nonlinear and complex, and experimental results indicate that the performance of the KG-NDA is overall superior to those obtained by the existing approaches. We also expect that in addition to face recognition, the KG-NDA algorithm will provide excellent performance in many pattern recognition tasks.

6.ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of P.R.China(No.60103018). The authors would like to thank Mr. D.Graham *et al.* for providing the UMIST database, and thank AT&T Lab Cambridge for providing the ORL database. In addition, the authors would like to thank Mr. B.Scholkopf *et al.* for providing the code of the K-PCA algorithm, thank Mr. G.Baudat *et al.* for providing the code of the GDA algorithm, and thank Mr. J.W.Lu *et al.* for providing the code of the KDDA algorithm.

7. REFERENCES

[1] M.Turk, A.P.Pentland, "Eigenfaces for recognition," *J.Cogni. Neurosci.*, vol.3, no.1, pp.71–86, 1991.

[2] P.N.Belhumeur, J.P.Hespanha, D.J.Kriegman, "Eigenfaces vs.Fish-erfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol.19, pp.711–720, July 1997.

[3] H.Yu, J.Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recogniti.*, vol.34, pp.2067–2070, 2001.

[4] K.Liu, Y.Q.Cheng, J.Y.Yang, "A generalized optimal set of discriminant vector," *Pattern Recogniti.*, vol.25, pp.731–739, 1992.

[5] Y.F.Guo, T.T.Shu, J.Y.Yang, S.J.Li, "Feature extraction method based on the generalized Fisher discriminant criterion and facial recognition," *Pattern Analysis & Applications*, vol.4, pp.61-66, 2001.

[6] Y.F.Guo, L.D.Wu, "A novel optimal discriminant principle in high dimensional spaces," *in Proc. IEEE ICDL,* 2002.

[7] B.Scholkopf, A.Smola, K.R.Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.,* vol.10, pp.1299-1319, 1999.

[8] S.Mika, G.Ratsch, J.Weston, B.Scholkopf, K.-R.Muller, "Fisher discriminant analysis with kernels," *in Proc. Neural Networks for Signal Processing IX*, Y.-H.Hu, J.Larsen, E.Wilson, and S.Douglas, Eds:IEEE, 1999, pp.41–48.

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

[10] J.W.Lu, K.N.Plataniotis, A.N.Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE trans. Neural Networks*, vol.14, pp117-126. 2003.

[11] Q.S.Liu, R.Huang, H.Q.Lu, S.D.Ma, "Face recognition using kernel based Fisher discriminant analysis," *in Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2002.

[12] C.J.Liu, H.Wechsler, "Robust coding schemes for indexing and retrieval from large face databases," *IEEE Trans. Image Processing,* vol.9, pp.132-137, Jan. 2000.

[13] J.Yang, J.Y.Yang, "Optimal FLD algorithm for facial feature extraction," *SPIE Processing of the Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision,* vol.4572, pp.438-444, 2001.