MODIFIED KERNEL-BASED NONLINEAR FEATURE EXTRACTION

Guang Dai *Yuntao Qian* *Sen jia*

College of Computer Science, Zhejiang University, P.R.China
Email: {dai_guang@yahoo.com.cn, ytqian@zju.edu.cn, zjujiasen@hotmail.com}

ABSTRACT

Feature extraction techniques are widely used in many applications to pro-process data in order to reduce the complexity of subsequent processes. A group of kernel-based Fisher discrminant analysis (KFDA) algorithms has attracted much attention due to their high performance. In this paper, the inherent limitations of those KFDA algorithms have been discussed and the novel algorithm will be proposed to effectively overcome those limitations. Experimental results on face recognition suggest that this proposed algorithm is superior to the existing methods in terms of correct classification rate.

Keywords: Feature extraction, kernel Fisher discriminant analysis, face recognition.

1. INTRODUCTION

Feature extraction is one of the most significant and fundamental problems in many applications such as pattern recognition and data mining, and extracting efficient feature is always the key to solving a problem in those applications. Principal component analysis (PCA) and linear discriminant analysis (LDA) are two classic tools widely used for data reduction and feature extraction [1,2,3,4,7]. For solving problems of classification, it is generally believed that LDA-based algorithms outperform PCA-based ones, since the former optimizes the low-dimensional representation of the objects with focusing on the most discrminant feature extraction while the latter only achieves simply object reconstruction [2]. Although LDA-based algorithms have been proven successful on classification problems, those algorithms fail for a nonlinear problem and are inadequate to describe the complex and nonlinear patterns.

Recently, the extension of linear methods to nonlinear ones, using the so-called kernel trick that was first used in Support Vector Machine (SVM) can efficiently construct nonlinear relations of the input data in a very high-dimensional feature space obtained by a nonlinear mapping $\phi : x \in R^N \to f \in F$, and then they only depend on inner products in the feature space F, but need not to compute in the feature space explicitly [5,6,8,9]. As nonlinear extensions of PCA and LDA, kernel PCA (KPCA) and kernel Fisher discrminant analysis (KFDA) have already been shown to provide a better performance than a linear PCA and LDA in several applications, respectively [5,6,8,9,10]. The basic ideas of KPCA and KFDA are to first map the input date x into a feature space F via a nonlinear mapping ϕ and then perform PCA and LDA in F, respectively. Since KPCA is inadequate for discriminating purposes as same as PCA, KFDA-based algorithms can outperform KPCA-based ones for solving a problem of classification [8,9]. Mika *et al.* [6] first proposed the two-class nonlinear discriminant algorithm by the kernel trick and G.Baudat *et al.* [5] extended this algorithm to multiclass problems. Subsequently, another form of multiclass nonlinear discriminant algorithm was also developed [8]. However, the general KFDA-based algorithms have two limitations: 1) in the case of small sample size problem (SSSP), these algorithms discard some significant discriminatory information; 2) these algorithms only extracts at most $c-1$ meaningful features, where c is the number of classes involved. In this paper, we propose a modified kernel-based nonlinear feature extraction algorithm, which can break the limitations above and is very useful for the SSSP. This proposed algorithm has the properties: in the space spanned by the first $c-1$ optimal discrminant vectors in the feature space F, the within-class distance of the training samples in the feature space F equals to zero, while the between-class distance of the training samples in the feature space F does not equal to zero; in the space spanned by the remaining optimal discriminant vectors in the feature space F, some other discriminatory information is also obtained. We apply this modified algorithm to face recognition, where the SSSP widely exists and the pattern distribution is generally nonlinear, and the experimental results reveal that this algorithm has the better performance for feature extraction.

2. KERNEL FISHER DISCRIMINANT VECTORS

For solving nonlinear problems, classic LDA has been generalized to its nonlinear version by the kernel trick, namely KFDA [6,8,9]. 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, where different classes of objects are supposed to be linearly separable. The basic idea of KFDA will seek to find a linear transformation in F which can maximize the between-class scatter and minimize the within-class scatter 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:

$$k(x, y) = (\phi(x)^T \cdot \phi(y)) . \quad (1)$$

Let $X = \{x_1, \ldots, x_N\}$ be the set of n -dimensional known-class training samples; there are c classes of samples: $X_1, \ldots, X_c$. It means that each x_i belongs to a class X_j, i.e., $x_i \in X_j$, $i = 1, \ldots, N$, $j = 1, \ldots, c$. $N_i, i = 1, \ldots, c$ is the number of samples which belong to class $X_i, i = 1, \ldots, c$, $N_1 + \ldots + N_c = N$. Then, let the training samples $X = \{x_1, \ldots, x_N\}$ project into the feature space F via the nonlinear mapping ϕ, and obtain the corresponding training samples $\{\phi(x_i)\}_{i=1}^N$ in F. 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:

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

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

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

where $m_i = (1/N_i) \sum_{x_j \in X_i} \phi(x_j)$ denotes the sample mean of class X_i in F; $m = \sum_{i=1}^{N} \phi(x_i)$ is the all sample mean in F. To

perform KFDA in F and 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 [5,6]; in addition, a new algorithm for calculating (5), which will directly carry out a direct LDA algorithm [8] in F, has been proposed, recently. The follows will discuss the two approaches.

2.1 The General Algorithms for the KFDA

By the theory of reproducing kernel, the general algorithms for the KFDA believe that w is in a space spanned by the training samples $\{\phi(x_i)\}_{i=1}^N$ in F and can be expressed:

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

Inserting (6) into (5), then (5) becomes a function of $\alpha = (\alpha_1, \ldots, \alpha_N)$ and can be converted to maximize:

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

where $K_b = \sum_{i=1}^c (N_i / N) \cdot (M_i - M)(M_i - M)^T$,

$K_w = (1/N) \cdot \sum_{i=1}^c \sum_{x_j \in X_i} (\xi_j - M_i)(\zeta_j - M_i)^T$,

with $M_i = ((1/N_i) \cdot \sum_{j=1}^{N_i} k(x_1, x_j), \ldots, (1/N_i) \cdot \sum_{j=1}^{N_i} k(x_N, x_j))^T$,

$M = ((1/N) \cdot \sum_{j=1}^{N} k(x_1, x_j), \ldots, (1/N) \cdot \sum_{j=1}^{N} k(x_N, x_j))^T$,

$\xi_j = (k(x_1, x_j), \ldots, k(x_N, x_j))^T$.

Therefore, the optimization solution to (7) is exactly as same as that to the algorithms of traditional LDA, and the maximum criterion $J_f(\alpha)$ can be formed by m eigenvectors corresponding to the first m eigenvalues of $K_w^{-1} K_b$. However, in the case of the smaller training set, K_w is singular and is thus not invertible in practice. There are currently two methods used to circumvent this problem of the noninvertibility of the matrix K_w. 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) [5]. However, this method tends to overfit the training set in some cases. The second method will introduce a nonsingular matrix $\tilde{K}_w = K_w + \tau I$ to replace the matrix K_w, where $\tau > 0$ and is called the conditioning coefficient, and I is the identity matrix. However, these algorithms described above have two same limitations: 1) they discard the null space of K_w, where may contain some significant discriminatory information; 2) since the rank of K_b is no more than $c-1$, these algorithms mentioned previously only extracts at most $c-1$ meaningful features.

2.2 Kernel Direct Discriminant Analysis Algorithm

Recently, a novel so-called kernel direct discriminant analysis (KDDA) algorithm has been proposed, and it is more useful for the smaller training set [8]. This algorithm generalizes the strengths of the recently proposed direct-LDA (D-LDA) of Yu *et al.* [3] and the kernel trick, and it can effectively overcome the limitations of the GDA [5] that the pseudoinverse matrix can cause some loss of the significant discriminatory information. The basic idea of the KDDA directly carries out the D-LDA algorithm in F. As same as the D-LDA [3], the KDDA believes that the null space of S_w in F may contain significant discriminatory information if the projection of S_b in F is not zero in that direction, and that no significant discriminatory information will be lost if the null space of S_b is discarded. As same as the D-LDA[3], the KDDA intends to seek the intersection space $(S_w(0) \cap \overline{S_b(0)})$, where $S_w(0) = \{x \mid S_w x = 0\}$, and $\overline{S_b(0)} = \{x \mid S_b x \neq 0\}$. In order to obtain this intersection space, the KDDA will first calculate the $c-1$ corresponding eigenvectors of all positive eigenvalues of S_b to obtain the space $\overline{S_b(0)}$, and then calculate the eigenvalues and corresponding eigenvectors of $\ddot{S}_w$, which is the projection of S_w in $\overline{S_b(0)}$, to obtain the space $S_w(0)$. From the procedure of KDDA, it is clear that $d(S_b(0)) \leq c-1$, and $d(S_w(0) \cap \overline{S_b(0)}) < c-1$, where $d(\bullet)$ denotes the dimensionality of the space' • '. However, according to the analysis in [4], it can be seen that $d(S_w(0) \cap \overline{S_b(0)}) = c-1$, since the dimensionality of the feature space F is far higher than the number of the training samples. As a result, the KDDA algorithm has the same limitations and shortcomings described in the general algorithms for the KFDA above. It not only loses some discriminatory information, but also only extracts at most $c-1$ meaningful features.

3. THE NOVEL ALGORITHM FOR THE KERNEL FISHER DISCRIMINANT VECTORS

In this section, a novel algorithm for the kernel Fisher discrminant vectors will be introduced, and it can effectively overcome the shortcomings and limitations of the previous algorithms for the kernel Fisher discriminant vectors. This proposed algorithm has two novel properties: 1) in the space spanned by the first $c-1$ optimal discrminant vectors in F, the within-class distance of the training sample in F equals to zero, while the between-class distance of the training sample in F does not equal to zero (in fact, the intersection space $(S_w(0) \cap \overline{S_b(0)})$ is completely obtained in this step); 2) in the space spanned by the remaining optimal discriminant vectors in F, some other discriminatory information can also be obtained in F. As same as some LDA-based algorithms, we believe that the optimal discriminant vectors in F can be calculated in the space $\overline{S_t(0)}$, where $\overline{S_t(0)} = \{x \mid S_t x \neq 0\}$. Otherwise, if $x \in S_t(0) = \{x \mid S_t x = 0\}$, it is clear that $x^T S_t x = x^T S_b x = 0$, i.e., the between-class distance in F equals 0, which is obviously meaningless for classification.

It is clear that S_t in (4) can be rewritten here as follows:

$$S_t = \sum_{i=1}^N \overline{\phi}(x_i) \overline{\phi}(x_i)^T = \Phi_t \Phi_t^T \tag{8}$$

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

As similar as the KDDA, the orthonormal bases of $\overline{S_t(0)}$ can be obtained by calculating the corresponding orthonormal eigenvectors of all positive eigenvalues of S_t. Since the dimensionality of the feature space 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 [1,3,8], those orthonormal eigenvectors can be indirectly derived from the eigenvectors of the matrix $\Phi_t^T \Phi_t$ (with $N \times N$).

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

$$K = (k_{i,j})_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}} \tag{9}$$

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:

$$\Phi_t^T \Phi_t = \frac{1}{N}(K - \frac{1}{N}(K \cdot I_{N\times N} + I_{N\times N} \cdot K) + \frac{1}{N^2} I_{N\times N} \cdot K \cdot I_{N\times N}) \tag{10}$$

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 [3,8], it is clear that the $v_i = \Phi_t e_i \lambda_i^{-1/2}$ ($i = 1,\ldots,m$) can constitute the orthonormal bases of $S_t(0)$ in F. Hence any optimal discriminant vector x in F can be expressed:

$$x = V \cdot y \tag{11}$$

where $y \in R^m$, $V = [v_1,\ldots,v_m]$.

Hence, the Fisher discriminant criterion $J(w)$ in F can be transformed and carried out in the projective space R^m of $S_t(0)$. By mapping the training samples $\{\phi(x_i)\}_{i=1}^{N}$ into the space $S_t(0)$, the corresponding training samples $\{y_i\}_{i=1}^{N}$ in the projective space R^m of $S_t(0)$ can be obtained as follows:

$$y_i = V^T \phi(x_i) = \sqrt{\frac{1}{N}} E^T (k_{i,1} - \frac{1}{N}\sum_{j=1}^{N} k_{i,j}, \ldots, k_{i,N} - \frac{1}{N}\sum_{j=1}^{N} k_{i,j})^T \tag{12}$$

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

Hence, the corresponding $\tilde{S}_b$, $\tilde{S}_w$, $\tilde{S}_t$ in R^m can be directly calculated in the projective space R^m of $S_t(0)$, since m is far less than N. As a result, the criterion function $J(w)$ in (5) can be rewritten in the projective space R^m:

$$J(Y) = \frac{Y^T \tilde{S}_b Y}{Y^T \tilde{S}_w Y} \quad \text{or} \quad J(Y) = \frac{Y^T \tilde{S}_b Y}{Y^T \tilde{S}_t Y} \tag{13}$$

It is easy to be seen that $\tilde{S}_b$, $\tilde{S}_w$ is semi-positive definite and $\tilde{S}_t$ is positive definite. In fact, the former one is equivalent to the latter one in (13)[8]. Then, we will carry out OFLD [11] algorithm to calculate the optimal discriminant vectors with respect to Fisher criterion (13). The within-class matrix $\tilde{S}_w$ can be split into its null space $\tilde{S}_w(0) = span\{\gamma_1,\ldots,\gamma_l\}$ and its orthogonal complement space $\tilde{S}_w(0) = span\{\gamma_{l+1},\ldots,\gamma_m\}$, where $\gamma_1,\ldots,\gamma_m$ are the orthonormal basis of R^m. In fact, it can be verified that all discriminatory information with respect to Fisher criterion (13) is contained in these two subspaces [11]. 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 $\tilde{S}_w(0)$. Hence, in $\tilde{S}_w(0)$, the Fisher criterion (13) can be replaced by $\hat{J}(Y) = Y^T \tilde{S}_b Y$. In order 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 $P_1 z_i (i = 1,\ldots,l)$ can constitute all optimal discriminant vectors in $\tilde{S}_w(0)$ and $VP_1 z_i (i = 1,\ldots,l)$ can constitute all optimal discriminant vectors in the intersection space $(S_w(0) \cap S_b(0))$. It is obviously that l must be $c-1$, since the dimensionality of the feature space F is far higher than the number of the training samples [4]. From [4], in the space spanned by those $c-1$ optimal discrminant vectors $VP_1 z_i (i = 1,\ldots,l)$ in F, the within-class distance of the training sample in F equals to zero, while the between-class distance of the training sample in F does not equal to zero. In addition, for optimal discriminant vectors in the space $\tilde{S}_w(0)$, let $P_2 = (\gamma_{l+1},\ldots,\gamma_m)$ and $\hat{S}_b = P_2^T \tilde{S}_b P_2$, $\hat{S}_t = P_2^T \tilde{S}_t P_2$, calculate $l' - l$ eigenvectors $z_{l+1},\ldots,z_{l'}$ corresponding to the first $l' - l$ leading eigenvalues of $\hat{S}_b^{-1} \cdot \hat{S}_t$. Then, $P_2 z_i (i = l+1,\ldots,l')$ constitute the optimal discriminant vectors in $\tilde{S}_w(0)$ and $VP_2 z_i (i = l+1,\ldots,l')$ constitute the remaining optimal discrminant vectors in F. It is clear that the $P_2 z_i (i = l+1,\ldots,l')$ constitute the optimal discriminant vectors of $S_b(0) \cap S_w(0)$. From the procedure above, we can see that $Y_i = VP_1 z_i (i = 1,\ldots,l)$, $Y_i = VP_2 z_i (i = l+1,\ldots,l')$ constitute all optimal discrminant vectors in F.

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

$$z = (P_1 z_1,\ldots,P_1 z_l, P_2 z_{l+1},\ldots,P_2 z_{l'})^T$$
$$\bullet \sqrt{\frac{1}{N}} E^T (k(x,x_1) - \frac{1}{N}\sum_{i=1}^{N} k(x,x_i),\ldots,k(x,x_N) - \frac{1}{N}\sum_{i=1}^{N} k(x,x_N))^T \tag{14}$$

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

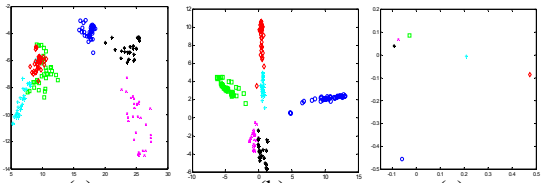
Thus, a low-dimensional representation $z = \Theta^T \phi(x)$ with enhanced discrminant power in F has been introduced. In addition, when $\phi(x) = x$, some D-LDA-based algorithms [4,7] are also viewed as a special case of the proposed kernel algorithms.

4. EXPERIMENTS

In order to test the effectiveness of the proposed algorithm, we will apply it to face recognition, where the SSSP widely exists and the pattern distribution is generally nonlinear and complex. In the following experiments, we have used the UMIST database (images.ee.umist.ac.uk) [12], which is a multiview database and consists of 575 gray-scale images of 20 subjects, each covering a wide range of poses from profile to frontal views as well as face gender and appearance. All original images are resized into 112X92 with 256-level gray scale.

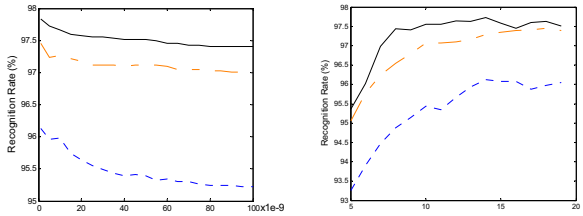
The first experiment will compare the feature distribution of this article algorithm with those of the D-LDA [3] and the KDDA [8]. And we only use a subset of the database, which contains 204 images of six randomly selected subjects (classes). Fig.1 depicts the first two most discriminant features extracted by the D-LDA, the KDDA and the novel algorithm, where a RBF kernel $k(z_1, z_2) = \exp(-\| z_1 - z_2 \|^2 / \sigma^2)$ with $\sigma^2 = 1e7$ is used. It is clear that some classes are still non-separable in the D-LDA-based subspace while they can more linearly separable in the KDDA-based subspace. In addition, those features that belong to the same class will cluster into the same point in the novel algorithm-based subspace, since the within-class distance of those features equals to zero. Hence, it is clear that the feature distribution in the novel algorithm-based subspace is more linearly separable than that in the KDDA-based subspace.

The following experiment will compare the novel algorithm with the KDDA [8] and the KFDA [9] in terms of the correct classification rates. The KDDA algorithm [8] can effectively compensate the limitation of the GDA [5] that the pseudoinverse matrix used in the GDA loses some significant discriminatory information, and the KFDA algorithm [9] that used the second method in Section 2 to solve the SSSP is more fit to the training set than the GDA [5] in many cases. In this experiment, we



(a) (b) (c)

Fig.1 Distribution of 204 samples of six classes in D-LDA-, KDDA- and the novel algorithm-based subspace. (a) for D-LDA; (b) for KDDA; (c) for the novel algorithm.



(a) (b)

Fig.2. Comparison rates, where ':': KDDA, '-.': KFDA, '-': the novel algorithm. (a): the recognition rates as functions of a ; (b): the recognition rates as functions of the number of feature vectors.

randomly chose the 5 training images per person from the database, and a training set of 100 images and a test set of remaining 465 images are created for the following experiment. The nearest neighbor classifier is used for classification. We do each experiment on 10 times and the results reported in this paper are an average of them.

Fig.2 depicts the recognition rates of those three methods when the polynomial kernel $(k(z_1, z_2) = (a(z_1 \cdot z_2) + b)^d)$ is used. 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) describes the recognition 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 highest summation of the recognition rate over the variation range of a . In addition, the optimal number of feature vectors in the novel algorithm is always more than 19, whereas the optimal number of feature vectors in the KDDA [8] and the KFDA [9] is often about 19. As a result, it can conclude that some discriminatory information can be contained in the space $S_b(0) \bigcap S_w(0)$ and it is often omitted in many traditional KFDA algorithms such as the KDDA [8] and the KFDA [9]. Fig.2 (b) describes the recognition rate curves as functions of the number of feature vectors within the range from 3 to 19, where the polynomial kernel with $a = 1e-19$ is used. According to Fig.2, we can see that the performance of the novel algorithm is overall superior to those of the other two algorithms, and it can effectively compensate the limitations and shortcomings of those algorithms. In addition, it is worthy to mention here that the computational requirements of the novel algorithm are tolerable to those of the KDDA algorithm and the KFDA algorithm.

Other comparative experiments on the different kernel functions with the different parameters have also been carried out, and the comparative experiments of applying those algorithms to other popular databases (the ORL database. available: www.uk.research.att.com) have been carried out too. In addition, we compare this novel algorithm with the GDA method [5] and the KPCA method [10] too. All results show that this novel algorithm is very effective. However, we refer reader to those results duo to space limitations.

5. CONCLUSIONS

In this paper, we have developed a novel algorithm for the kernel nonlinear Fisher discriminant analysis. The proposed method combines kernel-based methodologies with optimal discriminant analysis techniques. This algorithm can effectively break the inherent limitations in the general earlier kernel discriminant analysis algorithms. This algorithm has so properties: in the space spanned by the first $c-1$ optimal discriminant vectors in F , the within-class distance of the training samples in F equals to zero, while the between-class distance of the training samples in F does not equals to zero; in the space spanned by the remaining optimal discriminant vectors in F , some other discriminatory information in F can be obtained. We have applied this algorithm to the face recognition, and the experimental results tested on the different kernel function with the range parameter show that this algorithm is very effective.

6.ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of P.R.China (No.60103018). In addition, the author would like to thank Dr.D.Graham and Dr.N.Allinson for providing the UMIST database, and thank Dr.J.W.Lu for providing the code of the KDDA algorithm.

7. REFERENCES

[1] M.Turk, A.P.Pentland, "Eigenfaces for recognition," *J.Cogn. Neurosci.*, vol.3, no.1, pp.71–86, 1991.
[2] P.N.Belhumeur, J.P.Hespanha, D.J.Kriegman, "Eigenfaces vs.Fisher-faces: 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] Y.F.Guo, L.D.Wu, "A novel optimal discriminant principle in high dimensional spaces," *in Proc. IEEE ICDL,* 2002.
[5] G.Baudat, F.Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput*., vol.12, pp.2385-2404, 2000.
[6] 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.
[7] J.Yang, J.Y.Yang, "Why can LDA be performed in PCA transformed space," *Pattern recogniti.*, vol.36, pp.563-566, 2003.
[8] 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.
[9] 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.
[10] B.Scholkopf, C.J.C.Burges, A.J.Smola, *Advances in kernel Methods-Support Vector Learning*. Cambridge, MA: MIT Press, 1998.
[11] 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.
[12] D.B.Graham, N.M.Allinson, "Characterizing Virtual Eigensignatures for General Purpose Face Recognition," *in Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences,* H.Wechsler, P.J.Phillips, V.Bruce, F.Fogelman-Soulie, T.S.Huang, Eds., 1998, vol.163, pp.446-456.