Fisher+Kernel Criterion for Discriminant Analysis*

[1]Shu Yang, [2]Shuicheng Yan, [3]Dong Xu, [2,4]Xiaoou Tang, [1]Chao Zhang
[1]National Laboratory on Machine Perception, Peking University, Beijing, P.R. China
[2]Department of Information Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong
[3] MOE-Microsoft Key Laboratory of Multimedia Computing and Communication & Department of EEIS,
University of Science and Technology of China, Hefei, Anhui, P. R. China
[4]Microsoft Research Asia, Beijing, P.R. China
Contact: yangshu@cis.pku.edu.cn

Abstract

We simultaneously approach two tasks of nonlinear discriminant analysis and kernel selection problem by proposing a unified criterion, Fisher+Kernel Criterion. In addition, an efficient procedure is derived to optimize this new criterion in an iterative manner. More specifically, original input vector is first transformed into a higher dimensional feature matrix through a battery of nonlinear mappings involved in different kernels. Then, based on the feature matrices, FKC is presented within two coupled projection spaces: one projection space is used to search for the optimal combinations of kernels; while the other encodes the optimal nonlinear discriminating projection directions. Our proposed method is a unified framework for both kernel selection and nonlinear discriminant analysis. Besides, the algorithm potentially alleviates overfitting problem existing in traditional KDA and has no singularity problems in most cases. The effectiveness of our proposed algorithm is validated by extensive face recognition experiments on several datasets.

1. Introduction

Kernel Discriminant Analysis (KDA) has been widely used in computer vision, especially for face recognition [9] [16]. KDA first maps the original data to a higher dimensional feature space via a nonlinear mapping, and then applies Linear Discriminant Analysis (LDA) [6] in the transformed feature space. In KDA, we need not explicitly know the nonlinear mapping; and the kernel function, *i.e.* inner product of the data pair in feature space, is enough to derive final solution [13][15]. A kernel determines the induce bias of a learning algorithm on a specific data set; thus a proper way to select optimal kernel is crucial for such learning algorithms as KDA.

There have been some attempts to pursue more effective kernels [2] [8]. Generally speaking, these methods can be classified into two types: one is independent to the subsequent learning algorithm and the other is dependent. For the first type, new kernels are the refined ones of the traditional kernels with special motivations, such as [7] and the cosine kernel [10]. These methods are not always effective for specific algorithms in practice. For the second type, such as boosting kernel for SVM [1] and RBF kernel parameter selection for KDA [5], the procedure for kernel selection is time consuming and is usually restricted to a certain kind of kernels.

As discussed in this work, it is infeasible to conduct kernel selection in KDA due to the overfitting problem existing in traditional KDA when the kernel Gram matrix is nonsingular. In such cases, the data of the same class are mapped onto the same point; therefore traditional KDA cannot distinguish which kernel is better. On the other hand, searching for an optimal combination of kernels [3] [4] is regarded as a proper way for kernel selection. But simple linear combination of the kernel matrix is far from satisfactory, and the overfitting problem still exists if the combined kernel Gram matrix is nonsingular.

Motivated by the above observations, we propose a novel criterion, called Fisher+Kernel Criterion (FKC) for discriminant analysis, in which kernel selection and nonlinear discriminant analysis problems are simultaneously solved by optimizing a united objective function. More specifically, we first construct a compound transformation by combining several different nonlinear mappings involved in different kernels. The original input vector is mapped to a feature matrix through this compound map. Then these feature matrices are projected into a lower dimensional feature space by optimizing the FKC. Here in FKC, two projection spaces corresponding to the two dimensions of a matrix are used in the optimization. The right-projection space is used to search for a better combination of the simple nonlinear mapping functions; while the left-projection is devoted to finding the optimal projection directions for nonlinear discriminant analysis. Finally, an iterative optimization procedure is designed to optimize this criterion based on matrix analysis techniques. The integration of FKC and the solution procedure is called Fisher+Kernel Analysis (FKA) in this work.

The rest of this paper is organized as follows. First, we provide an overview of the KDA algorithm in Section 2. In Section 3, we present the Fisher+Kernel Criterion and its optimization procedure. The detailed analysis of the FKA is presented in Section 4. Then, the face recognition experiments on different benchmark databases are given in Section5. Finally, we conclude this paper in Section 6.

*This work was performed at Microsoft Research Asia.

1063-6919/05 $20.00 © 2005 **IEEE**

2. Kernel Discriminant Analysis

Given a collection of training image samples denoted as $\chi = \{x_1, x_2 \cdots x_n\}$, where $x_i \in \mathbb{R}^m$ and sample x_i belong to the l_i *-th* class, $l_i \in \{1, 2, ..., p\}$ and n_l is the number of samples belonging to the *l-th* class.

Let ϕ: $x \in \chi \rightarrow \phi(x) \in \mathcal{F}$ be a nonlinear mapping from the original input space to a higher dimensional feature space $\mathcal{F}$; and the inner product in the feature space is defined as a kernel function $k(x, y) = \phi(x) \bullet \phi(y)$.

The main idea of KDA is to apply the Fisher Criterion in the higher dimensional feature space. So the criterion for KDA algorithm can be given as follows

Criterion 1 $\psi^* = \arg\max\limits_{\psi} \dfrac{\left|\left(\psi^T S_B \psi\right)\right|}{\left|\left(\psi^T S_W \psi\right)\right|}$

where $\psi \in span\{\phi(x_i), i = 1, ..., n\} \subset \mathcal{F}$, $\psi = \sum_{i=1}^{n} \alpha_i \phi(x_i)$

And the intra-class scatter matrix S_w and the inter-class scatter matrix S_B are

$$\begin{aligned} S_W &= \sum\nolimits_{l=1}^{n} \left(\phi(x_i) - \overline{\phi}^{l_i}\right)\left(\phi(x_i) - \overline{\phi}^{l_i}\right)^T \\ S_B &= \sum\nolimits_{l=1}^{p} n_l \left(\overline{\phi}^{l} - \overline{\phi}\right)\left(\overline{\phi}^{l} - \overline{\phi}\right)^T \end{aligned} \qquad (1)$$

where $\overline{\phi}^l$ is the average of the mapped samples belonging to the class-*l*, and $\overline{\phi}$ is the average of all mapped samples. Denote $\Phi = \left(\phi(x_1), \cdots, \phi(x_n)\right)$, then we have

$$\begin{aligned} S_W &= \Phi \sum_{i=1}^{n} \left(e_i - \frac{1}{n_{l_i}} \sum_{j=1}^{n} \delta(l_i, l_j) e_j \right)\left(e_i - \frac{1}{n_{l_i}} \sum_{j=1}^{n} \delta(l_i, l_j) e_j \right)^T \Phi^T \\ &\triangleq \Phi M_w \Phi^T \end{aligned}$$

where $\delta(i, j) = \begin{cases} 1 & if \;\; i = j \\ 0 & else \end{cases}$, and e_i is a *n*-dimensional vector with $e_i(j) = \delta_{ij}$. Let S_T be the total scatter matrix as

$$S_T = \Phi \sum_{i=1}^{n} \left(e_i - \frac{1}{n} \sum_{j=1}^{n} e_j \right)\left(e_i - \frac{1}{n} \sum_{j=1}^{n} e_j \right)^T \Phi^T \triangleq \Phi M_T \Phi^T$$

Then S_B can be written as

$$S_B = S_T - S_W = \Phi\left(M_T - M_W\right)\Phi^T \triangleq \Phi M_B \Phi^T \qquad (2)$$

Thus, in the higher or even infinite dimensional feature space $\mathcal{F}$, we can define the scatter matrices directly as

$$S_W = \Phi^T M_W \Phi \text{ and } S_B = \Phi^T M_B \Phi \qquad (3)$$

Denote the kernel Gram matrix $K = \Phi^T \Phi$, then the optimal solution of criterion 1 can be obtained as

$$K^T M_B K \alpha = \lambda K^T M_W K \alpha \qquad (4)$$

From Eq. (4), we have the following observations: class information is contained in matrices M_B and M_W while distribution information of training data is included in the matrix *K*. For classification problem with fixed number of samples, learning algorithm is fully decided by its kernel Gram matrix. Moreover, we have $rank(K^T M_W K) \leq rank(M_W) = n - p$; hence the dimension of the null space of $K^T M_W K$ is no less than *p*; and the number of α satisfying $\max\limits_{\Psi} \dfrac{\left|\left(\psi^T S_B \psi\right)\right|}{\left|\left(\psi^T S_W \psi\right)\right|} = +\infty$ is at least p. As proved in appendix, when kernel gram matrix is nonsingular, samples of the same class are mapped on the same point in the learned lower dimensional feature space, which leads to overfitting. And this overfitting makes KDA fail to determine which kernel is better since all the criterion values will be infinite if nonsingular kernels are used.

These considerations motivate us to propose a criterion which has the ability to distinguish different kernels and meanwhile can be solved efficiently. To this end, we propose Fisher+Kernel Criterion as follows.

3. Fisher+ Kernel Criterion and Optimization

Define $\| A \|_F = \sqrt{\sum_{i=1, j=1}^{m} A_{ij}^2}$ as Frobenius norm of matrix A. Denote a collection of nonlinear mapping functions as: $\phi_j : \chi \rightarrow \mathcal{H}_j$, where $j \in \{1, \cdots, f\}$ and $\mathcal{H}_j$ is a Hilbert space. $\mathcal{H}$ is a Hilbert space as the direct sum of $\mathcal{H}_j$. So, $\hat{\phi}_j(x) \in \mathcal{H}$, $j \in \{1, \cdots, f\}$, is a vector expanded by $\phi_j(x)$.

Traditional kernel methods are processed in the single Hilbert space $\mathcal{H}_j$; but here we intend to process the kernel selection in the larger Hilbert space $\mathcal{H}$. Thus the *i-th* original sample x_i is first mapped onto a so called feature matrix $\Phi(x_i) = [\hat{\phi}_1(x_i), \cdots \hat{\phi}_f(x_i)]$ in that larger space $\mathcal{H}$.

3.1 Fisher + Kernel Criterion

Since kernel function is the inner product of $\hat{\phi}_j$ in Hilbert space $\mathcal{H}$, selection for kernels is actually the selection of nonlinear map $\hat{\phi}_j$. Besides, we intend to learn the most discriminating features with Fisher Criterion, which has achieved good performance in many real world problems. To fuse kernel selection problem into Fisher Criterion, we propose Fisher +Kernel Criterion below:

Critrion2 $(U^*, V^*) = arg \max\limits_{U,V} \dfrac{\sum_c n_c \| U^T \overline{\Phi}_c V - U^T \overline{\Phi} V \|_F^2}{\sum_i \| U^T \Phi(x_i) V - U^T \overline{\Phi}_{l_i} V \|_F^2}$

where $U \in span(\hat{\phi}_j(x_i)) \subset \mathcal{H}$, $V \in \mathbb{R}^{f \times f'}$, $\overline{\Phi}$ is total average matrix of all the feature matrix $\Phi(x_i)$, $\overline{\Phi}_c$ is average matrix of the $\Phi(x_i)$ belonging to class c, and so for $\overline{\Phi}_{l_i}$.

In criterion 2, right matrix V is to search a better combination of $\hat{\phi}_j, (j = 1, \cdots f)$, so right-projection space is a finite dimensional Euclidean space. While left matrix is used to find the most discriminative subspace in $\mathcal{H}$; therefore left-projection space is a subspace of $\mathcal{H}$.

Our criterion is based on the whole Hilbert space $\mathcal{H}$, so inner product should be defined between any two vectors in $\mathcal{H}$ for similarity measure; to simplify the algorithms, we assume that the feature vectors mapped by different ϕ_j are independent, i.e.

$$\hat{\phi}_j(x) \bullet \hat{\phi}_j(y) = \hat{k}^j(x, y) \text{ and } \hat{\phi}_i(x) \bullet \hat{\phi}_j(y) = 0 \quad (i \neq j) \quad (5).$$

Specifically, we can define $\hat{\phi}_j = (\overbrace{0 \cdots 0}^{1 \cdots j-1}, \ \phi_j^{'} \ , \overbrace{0 \cdots 0}^{j+1 \cdots f})'$.

Conventionally, we can pursue left projection subspace in $span(\hat{\phi}_j(x_i))$, yet it will result in a very large scale optimization problem. To cut the overhead of calculation, we propose two methods to constrain the left-projection space, which leads to two different algorithms, called FKA01 and FKA02, respectively.

1). FKA01: Assume left-projection space is constrained in $span\left(\tilde{\phi}(x)\right)$, where $\tilde{\phi}(x) = \sum_{j=1}^{f} \hat{\phi}_j(x)$ (6).

Denote $U = \left[\tilde{\phi}(x_1), \tilde{\phi}(x_2), \cdots \tilde{\phi}(x_n)\right] \bullet L \triangleq \tilde{\Phi} \bullet L$, $L \in \mathbb{R}^{n \times m}$. Suppose the projection of $\Phi(x_i)$ in left projection space is K_i, then the element of K_i, *i.e.* $K_i(a,b)$, has the form:

$$K_i(a,b) = \tilde{\phi}(x_a) \bullet \hat{\phi}_b(x_i) = \phi_b(x_a) \bullet \phi_b(x_i) = k^b(x_a, x_i) .$$

Different kinds of kernel functions k^b are chosen from the kernel bank $\{k^j\}$, which can be obtained as a kind of prior knowledge given by users.

2).FKA02: Another way to alleviate computational cost is to use the mapping of class centroids approximate the $span(\hat{\phi}_j(x_i))$, namely, $\hat{\phi}_j(\overline{x}_s)$ is used in left projection space instead of all $\hat{\phi}_j\left(x_i\right)$, where $\overline{x}_s$, $s \in \{1, \cdots p\}$ is the centroid of *s-th* class. Then we can define

$$U = [\Phi(\overline{x}_1), \Phi(\overline{x}_2), \cdots \Phi(\overline{x}_p),] \bullet L \triangleq \tilde{\Phi} \bullet L , L \in \mathbb{R}^{fp \times m} .$$

Denote the projection of $\Phi(x_i)$ in the projection space as K_i, then the element of K_i, *i.e.* $K_i(a,b)$, has the form:

$$\begin{aligned} K_i(a,b) &= \hat{\phi}_{a_1}(\overline{x}_{a_0}) \bullet \hat{\phi}_b(x_i) = \delta_{a_1,b} \cdot \phi_b(\overline{x}_{a_0}) \bullet \phi_b(x_i) \\ &= \delta_{a_1,b} \cdot k^b(\overline{x}_{a_0}, x_i), \qquad a = (a_0 - 1) f + a_1 \end{aligned}$$

Though the two mentioned methods have different projection spaces, both of them satisfy the same criterion 2 and have the same form. So we can solve these two algorithms in the same way as follows:

$$\text{Denote that} \quad \overline{K}_c = \tilde{\Phi}^T \bullet \overline{\Phi}_c \qquad \overline{K} = \tilde{\Phi}^T \bullet \overline{\Phi}$$
$$K_i = \tilde{\Phi}^T \bullet \Phi_i \qquad \overline{K}_{l_i} = \tilde{\Phi}^T \bullet \Phi_{l_i} \qquad (7)$$

Then the criterion 2 can be simplified to

```
**Fisher +Kernel Analysis:**
Given input sample set $\{x_i \in \mathbb{R}^m\}$, i=1,...,n , class labels $l_i \in \{1, 2, ..., p\}$ and the desired final dimensions $m$ and $f'$.
1. Construct the kernel matrix defined in the Eq. (7)
2. Initiate $V_0$.
3. for $t$ =1, 2, … , $f$, do
    a) For given $V_{t-1}$, calculate the optimal $L_t$ from the Eq. (8) by using a generalized eigenvector decomposition method.
    b) For given $L_t$, calculate the optimal $V_t$ from Eq. (11) using a general eigenvector decomposition method.
    c) If $\| L^t - L^{t-1} \|_F < \varepsilon$ , $\| V^t - V^{t-1} \|_F < \varepsilon$ and $t > 2$, go to step 4; else, continue;
4. Output the projections $L = L_t \in \mathbb{R}^{fp \times m} or \mathbb{R}^{n \times m}$ and $V = V_t \in \mathbb{R}^{f \times f'}$.
```

Figure1. The procedure for Fisher+Kernel Analysis

Criterion3 $(L^*, V^*) = arg \max\limits_{L,V} \dfrac{\sum_c n_c \| L^T \overline{K}_c V - L^T \overline{K} V \|_F^2}{\sum_i \| L^T K_i V - L^T \overline{K}_{l_i} V \|_F^2}$

The kernel matrices can be directly computed from the kernel function bank $\{k^i\}$. Also, K matrices in criterion 3 has finite dimensions, thus criterion 3 can be solved using general matrix analysis techniques. To the best of our knowledge, criterion 3 has no closed-form solution; here we present an iterative procedure to solve this problem.

3.2 Iterative optimization procedure

For a given $V \in \mathbb{R}^{f \times f'}$, the objective function of criterion 3 is rewritten as:

$$\begin{aligned} \frac{\sum_c n_c \| L^T \overline{K}_c V - L^T \overline{K} V \|^2}{\sum_i \| L^T K_i V - L^T \overline{K}_{l_i} V \|_F^2} &= \frac{\sum_c n_c \| L^T \overline{K}_c^v - L^T \overline{K}^v \|_F^2}{\sum_i \| L^T K_i^v - L^T \overline{K}_{l_i}^v \|_F^2} \\ &= \frac{Trace(L^T S_B^v L)}{Trace(L^T S_W^v L)} \end{aligned} \quad (8)$$

where the symbol A^v with superscript v means $A^v \doteq AV$,

and
$$\begin{aligned} S_B^v &= \sum_c n_c (\overline{K}_c^v - \overline{K}^v)(\overline{K}_c^v - \overline{K}^v)^T \\ S_W^v &= \sum_i (K_i^v - \overline{K}_{l_i}^v)(K_i^v - \overline{K}_{l_i}^v)^T \end{aligned} \quad (9)$$

and the optimal L for the given V can be obtained from

$$S_B^V x = \lambda S_w^V x \quad (10)$$

Similarly, for a given $L \in \mathbb{R}^{n \times m} or \mathbb{R}^{fp \times m}$, the objective function of criterion 3 can be reorganized as follows

$$\frac{\sum_c n_c \| L^T \overline{K}_c V - L^T \overline{K} V \|_F^2}{\sum_i \| L^T K_i V - L^T \overline{K}_{l_i} V \|_F^2} = \frac{Trace(V^T S_B^L V)}{Trace(V^T S_W^L V)} \quad (11)$$

where the symbol A^L with superscript L means $A^L \doteq L^T A$ and

$$S_B^L = \sum_c n_c (\bar{K}_c^L - \bar{K}^L)^T (\bar{K}_c^L - \bar{K}^L)$$
$$S_W^L = \sum_i (K_i^L - \bar{K}_{l_i}^L)^T (K_i^L - \bar{K}_{l_i}^L) \quad (12)$$

The optimal solution V for the given L can also be calculated similar to Eq. (10).

Then, we develop the so called Fisher+Kernel Analysis (FKA) procedure to iteratively optimize the matrix L and V for the local optimum as in Figure 1.

4. Algorithmic Analysis

Here, we discuss the characteristics of FKA as follows:

Merge kernel selection into discriminant analysis. Kernel selection and discriminant analysis are carried out simultaneously using two projection spaces. Also we only need the final projection of Gram matrices rather than the exact form of final selected kernel function. Thus FKA avoids the complex process in kernel selection and can also be generalized into other kernel algorithm such as KPCA. Besides, the kernel selection result is optimal for the whole projection matrix. That is quite different from most other algorithms [5], which can only guarantee that the selected kernel is optimal for the first dimension of the projection matrix.

Process discriminant analysis in a larger Hilbert space. In traditional KDA, discriminant analysis is processed in each $\mathcal{H}_i$, and our algorithm is carried out in the whole feature space $\mathcal{H}$. If traditional KDA classification is taken as a nonlinear classifier, the new algorithm can be regarded as a synthesis of several different classifiers. This brings possibility of capturing more discriminating features of the data. Moreover, this larger space enables our algorithm capability to select kernel among various different kinds of kernel function.

Could avoid overfitting in traditional KDA. As proved in appendix, when K is a nonsingular matrix, there exists overfitting problem in KDA algorithm. Most of the kernel matrix is nonsingular, e.g. RBF kernel cases [15]. So in this condition, it is infeasible to pursue the selection of kernel under traditional Fisher Criterion, for all the kernels are equal under this criterion.

In our algorithms, original image vectors are mapped to feature matrices. Take FKA01 for example and suppose V^* is the optimal V matrix. Note that $K_i^v \in \mathbb{R}^{n \times f}$ and $S_W^v = \sum_i (K_i^v - \bar{K}_{l_i}^v)(K_i^v - \bar{K}_{l_i}^v)^T \in \mathbb{R}^{n \times n}$. So, $rank(S_w)$ is around $\min\{(n-p)*f, n\}$. When we use multiple kernels, there is $(n-p)*f >> n$, therefore the matrix S_w can be guaranteed to be in full rank in most cases. The case of FKA02 is similar to that; hence both of our algorithms can alleviate overfitting to some extent. Therefore, it is different with those combined kernel methods which simply add weight to each kernel matrix, whereas the overfitting and singular problem still exists for KDA.

5. Experiments

In this section, we conduct a series of experiments to compare FKA algorithm with traditional KDA. We use Gaussian kernel: $k^i(x,y) = \exp\left(-\|x-y\|^2 / 2\sigma_i^2\right)$ and set the parameter $\sigma_i = i \times \sigma, i \in 1, \cdots 10$, where σ is the standard variation of training data. $\{k^i, i = 1, \ldots, 10\}$ is used as kernel bank in FKA. Algorithms are tested on the benchmark face databases, ORL [11], FERET [12] and CMU PIE [14]. For ease of presentation, each experiment is named as Gm/Pn, which means m images per person are *randomly* selected as gallery set and other n for probe set. Histogram equilibrium has been applied as preprocessing step and nearest neighbor is used as final classifier.

5.1. ORL database

The ORL database contains 400 images of 40 individuals. Some images were captured at different times and had different variations including expression and facial details. The images were taken with a tolerance for some tilting and rotation of the face up to 20 degrees. All images are grayscale and normalized to a resolution of 46*56 pixels. Five sample images of one person in the ORL database are shown in Figure 2.

Figure 2. Five sample images in ORL face database

Figure 3 and 4 display the error rates of three sets tests compared between KDA and FKA. Vertical axis represents error rate of different algorithm. Horizontal axis represent different algorithms, where the numbers from 1 to 10 denote KDA algorithms using ten different parameters while 11 and 12 represent FKA02 and FKA01 taking ten kernels in KDA as the kernel bank, respectively. The whole databases are divided into gallery and probe set as: Test A G3/P7, Test B G2/P8 and Test C G5/P5. From Figure 3, we can see that FKA02 has much lower error rate than the KDAs with different kernels in both cases.

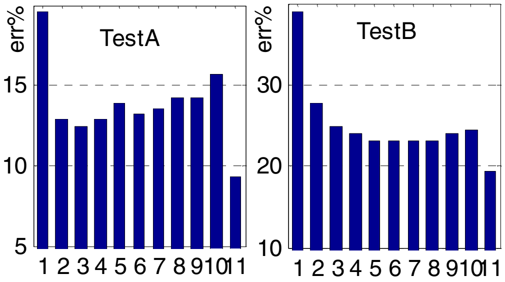

Figure 3. Error Rate of FKA02 vs. KDAs on ORL database.

In Figure 4, we show the results of both FKA 01 and 02 in Test C to compare with ten KDAs with different kernels. It also confirms that both FKA algorithms outperform all the KDAs.

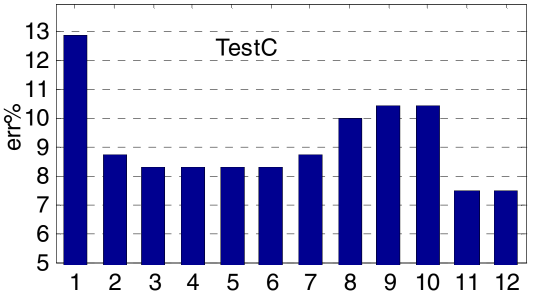

Figure 4. Error Rate of FKA02 &FKA01 vs. KDAs on the ORL database. Note that horizontal axis coordinate 11 represents FKA02 and 12 represents FKA01.

Figure 5 displays recognition rates of FKA02 on different dimensions and the results are from Test C. Horizontal axes represent the row and column dimensions of the final low dimensional matrix used for face recognition.

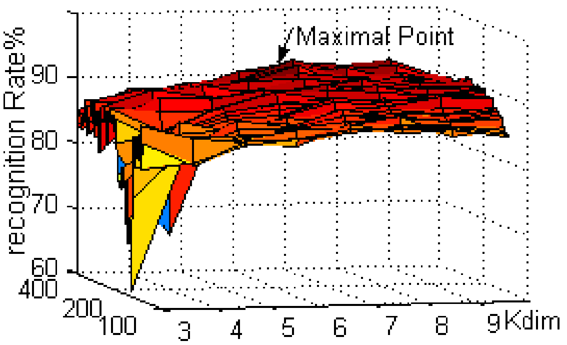

Figure 5. Accuracy of FKA 02 vs. number of eigenvectors in Test C on the ORL database.

5. 2. FERET database

In this experiment, seventy persons of the FERET database are used and each person has six different face images. All images are aligned by fixing the location of the two eyes and resized to 46*56 pixels. There are facial expression, illumination and pose, facial details variances in the images. Figure 6 displays six examples of one person in FERET.

Figure 6. Six samples of one subject in FERRET database

In this experiment, we randomly partition the database into G4/P2 as test A and G2/P4 as test B. Figure 7 presents the results of these two experiments, which also demonstrates that the algorithm FKA02 outperforms traditional KDA algorithm with different kernel parameters.

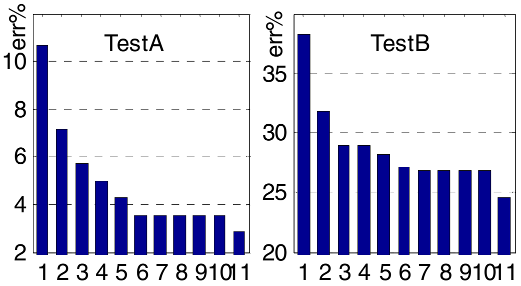

Figure 7. Error rate of FKA02 vs. KDAs on FERET database.

5.3. PIE database

The CMU PIE database contains more than 40,000 facial images of 68 people. The images were acquired across different poses, under variable illumination conditions and with different facial expressions. In our used database of PIE, five near frontal poses (C27, C05, C29, C09 and C07) and illumination 08 and 11 are chosen. The flash 08 and 11 are placed near the center and the illumination can be considered as the nearly frontal illumination. Each person has ten images and all the images are aligned by fixing the locations of two eyes, and the images are resized to 64*64 pixels. Figure 8 shows five examples of one person without preprocessing.

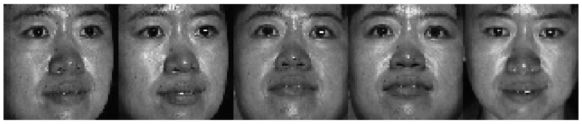

Figure 8. Five images of one person in the PIE database

Similar to experiments above, the data set is randomly partitioned into gallery and probe sets with G4/P6 in test A and G3/P7 in test B. We compare KFA01 with KDA in this experiment. The result in Figure 9 again shows that KFA01 can improve face recognition accuracy compared with traditional KDA with different kernel parameters.

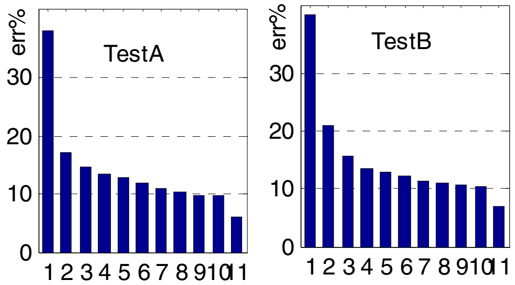

Figure 9. Error Rate of FKA01 vs. KDAs on the PIE database

From the results above, we can find that the parameter for KDA to obtain the best performance is discrepant on different data set; hence, it is difficult and unreasonable to select kernels by experience as in traditional KDA. While our algorithms can effectively combine different kernels and derive elegant representation for classification; thus are superior to traditional KDA in almost all cases.

6. Conclusion

In this paper, we proposed a novel criterion that fuses the process of kernel selection into the nonlinear discriminant analysis. Also we developed an efficient iterative procedure for the optimization of the criterion. The algorithm integrating these two contributions, called *Fisher+Kernel Analysis* (FKA), automatically combines all the kernels in user-defined kernel bank to search for the most discriminating features; moreover, it alleviates the overfitting existed in traditional KDA. Extensive experiments on different face databases validate the superiority of the proposed FKA compared with traditional KDA algorithm.

Appendix

Proof of overfitting in KDA mentioned in Section 4:

Lemma 1 The rank of matrix M_w is *n-p*.

Theorem 1 In KDA algorithms, when K is a nonsingular matrix, there exist p basis vectors that map the original data of the same class onto the same point

Proof: When K is nonsingular, Eq. (4) can be simplified as:

$$M_B K\alpha = \lambda M_W K\alpha \tag{13}$$

Set that $K\alpha = \beta, \beta \in \mathbb{R}^n$. Using lemma 1, we conclude that M_W only has p zero eigenvalues.

Define $\beta_l = (\overbrace{0,\cdots 0}^{\sum_{j=1}^{l-1} n_j}, \overbrace{1,\cdots 1}^{n_l}, 0,\cdots 0)$, we have

$$M_W \beta_l = diag(0,\cdots 0, A_l \beta_l, 0,\cdots 0) = 0 \tag{14}$$

So $\beta_l (l = 1 \cdots p)$ is the eigenvector corresponding to M_W's zero eigenvalues. Recalling the M matrices defined in Eq. (2) and (3), we can obtain that

$$M_B \beta_l = \left(M_T - M_W\right)\beta_l = M_T \beta_l \neq 0 \tag{15}$$

So $\alpha_l = K^{-1}\beta_l$ is chosen as the eigenvectors, and we have $\lambda = \infty$.

The image of x_i in higher dimensional feature space, *i.e.* $\phi(x_i)$, is projected to $g_l(x_i) = \phi(x_i)\psi_l$. Note that

$$g_l(x_i) = \phi(x_i)\sum_{j=1}^{n} \alpha_l(j)\phi_j = K_i \bullet \alpha_l = \beta_l$$

where $\beta_l(i)$ means the i th entry of the vector β_l . Let $\psi_l = \sum_{j=1}^{n} \alpha_l(j)\phi_j$ be a set of basis vectors of the feature space, where the vector α_l corresponds $K\alpha_l = \beta_l$. Note that when x_i, x_j are two different samples from the same class (for example, class c) in the original data set, they are projected to the coordinate 0 under the basis $\beta_l (l \neq c)$ and the coordinate 1 under the basis $\beta_i (i = c)$. Under this set of basis, the data of the same class is mapped to the same data point in the dimension-reduced subspace. ■

Acknowledgement:
This work was partially supported by NKBRPC (2004CB318005).

References

[1] K. Crammer, J. Keshet, and Y. Singer. "Kernel Design Using Boosting,'' Advances in Neural Information Processing Systems Vol. 15. Cambridge, MA, MIT Press.

[2] N. Cristianini, J. Kandola, A. Elisseeff and J. Shawe-Taylor. "On kernel-target alignment," in Neural Information Processing System, pp.367-373, British Columbia, Canada Dec. 2001.

[3] T. Evgeniou, M. Pontil and A. Elisseeff. "Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers," Machine Learning, Vol 55, Issue 1, pp. 71-97, Apr.2004.

[4] G. Fung, M. Dundar, J. Bi and B. Rao, "A Fast Iterative Algorithm for Fisher Discriminant using Heterogeneous Kernels," Proc. of 21st Int. Conf. on Machine Learning ,Banff, Canada, 2004.

[5] H. Jiang, C. Pong, W. Chen and J. Lai. "Kernel subspace LDA with optimized kernel parameters on face recognition," Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2004.

[6] P. Howland and H. Park, "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 26, No8, August 2004.

[7] R. Kondor and T. Jebara. "A Kernel between Sets of Vectors," in Proc. of the 20th Int. Conf. on Machine Learning, Washington DC, 2003.

[8] G. Lanckriet, N. Cristianini, L. Ghaoui, P. Bartlett and M. Jordan, "Learning the kernel matrix with semi-definite programming," Proc. of the 19th Int. Conf. on Machine Learning 2002, Sydney, Australia.

[9] J. Lu, K. Plataniotis and N. Venetsanopoulos. "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms," IEEE Trans. On Neural Networks, 2002.

[10] Q. Liu, H. Lu, S. Ma. "Improving kernel Fisher discriminant analysis for face recognition," IEEE Trans. On Circuits and Systems for Video Technology, Vol. 14 pp42-49, Jan. 2004.

[11] Olivetti & Oracle Research Laboratory, The Olivetti & Oracle Research Laboratory Face Database of Faces, http://www.cam-orl.co.uk/facedatabase.html.

[12] I. Philips, 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, 1998.

[13] B. Scholkopf and Alex Smola. "Learning with Kernels," chapter 2.3, Th2.18, MIT Press, Cambridge, MA, 2002.

[14] T. Sim, S. Baker and M. Bsat. "The CMU Pose, Illumination, and Expression (PIE) Database," in Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, May, 2002.

[15] J. Taylor and N. Cristianini. "Kernel Methods for Pattern Analysis," Cambridge University Press, 2004.

[16] M. Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods", in Proc. of the 5th Int. Conf. on Automatic Face and Gesture Recognition, Washington D. C., May 2002.