

Face Recognition Using Kernel Based Fisher Discriminant Analysis

Qingshan Liu Rui Huang Hanqing Lu Songde Ma
National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Science, P.O.Box 2728, Beijing, China
qslu, rhuang, luhq, masd@nlpr.ia.ac.cn

Abstract

Fisher Linear Discriminant Analysis (FLDA) has been successfully applied to face recognition, which is based on a linear projection from the image space to a low dimensional space by maximizing the between-class scatter and minimizing the within-class scatter. But face image data distribution in practice is highly complex because of illumination, facial expression and pose variations. In this paper, we present to use Kernel based Fisher Discriminant Analysis for face recognition. The kernel trick is used firstly to project the input data into an implicit space called feature space by nonlinear kernel mapping, then Fisher Linear Discriminant Analysis is adopted to this feature space, thus a nonlinear discriminant can be yielded in the input data. Another similar Kernel-based method is Kernel PCA, in which PCA is used in the feature space. The experiments in this paper are performed with the polynomial kernel, and this method is compared with Kernel PCA and FLDA. Extensive experimental results show that the correct recognition rate of this method is higher than that of Kernel PCA and FLDA.

1. Introduction

Face recognition has received extensive attention within the past 20 years because of the potential applications in many fields, such as identity authentication, surveillance and human-computer interface. Numerous algorithms have been proposed and a detailed survey was given in reference [17].

Linear subspace analysis has been used for face recognition by many researchers because of its simplicity and efficiency, such as linear Principal Component Analysis (PCA) or Fisher Linear Discriminant Analysis (FLDA). It seeks to project the input data into a low dimensional space through a linear transformation. Linear PCA^[1,11] is optimal for representation and reconstruction.

But it is inadequate for discriminant purpose between different faces because PCA seeks to maximize the total scatter across all classes. The idea of FLDA^[5,12,16] seeks to find discriminant projectors which can maximize the between-class scatter and minimize the within-class scatter. P.Belhumeur^[12] compared the linear PCA and FLDA methods for face recognition on the *Harvard* and *Yale* face databases, and experimental results showed that FLDA was better than linear PCA, especially with illumination variation. However, FLDA is still a linear techniques in nature, so it is also inadequate to describe the complex variation of the face images in practice, such as illumination, facial expression and pose variations. M.Bartlett^[9] proposed to use Independent Component Analysis (ICA) for face recognition, which separates the high-order moments of the input in addition to second-order moments adopted in linear PCA. But B.Moghaddam^[2] compared it with linear PCA on the *FERET* face database and found that it gave the same recognition accuracy as linear PCA.

Recently kernel-based nonlinear analysis has been given more attention in pattern recognition, because the kernel trick can efficiently construct nonlinear relations of the input data in an implicit feature space obtained by nonlinear kernel mapping, and they only depend on inner products in the feature space but need not to compute the feature space explicitly. The kernel trick was first used in Support Vector Machines (SVM)^[7], which looks for an optimal separating hyperplane in the feature space; and was applied to a wide class of problems including face recognition^[3,9]. SVM is based on the margin maximization and sensitive only to the extreme values. Kernel PCA^[4] combines the kernel trick with PCA to find nonlinear principal components in the feature space. M.Yang^[10] demonstrated that it outperformed PCA in face recognition. However, as the same as PCA, Kernel PCA captures the overall variance of all patterns which are inadequate for discriminating purposes.

In this paper, Kernel-based Fisher Discriminant Analysis^[7,14] (Kernel FDA) is proposed for face recognition. Similar to kernel PCA, Kernel FDA combines the kernel trick with FLDA. Firstly, the kernel

trick is used to project the input data into an implicit feature space through nonlinear kernel mapping, then FLDA is performed in this feature space, thus a nonlinear discriminant can be yielded in the input data. Recently a similar idea has been successfully used in modeling multi-view face^[19] and gesture recognition^[20]. Our experiments are performed with the polynomial kernel and Kernel FDA is compared with Kernel PCA and FLDA. Extensive experimental results show that it can give higher recognition rate than Kernel PCA and FLDA.

The rest part of this paper is arranged as follows: Kernel FDA is introduced in Section 2, Experiments are described in Section 3 and discussions in Section 4, finally, we will give our conclusion.

2. Kernel FDA

The idea of Kernel FDA is to yield a nonlinear discriminant in the input space through the kernel trick and FLDA. At first, the input data is projected into an implicit feature space F by nonlinear mapping, $\phi: x \in R^N \rightarrow f \in F$, then seeks to find a linear transformation in F which can maximize the between-class scatter and minimize the within-class scatter in F . Moreover, it is unnecessary to compute ϕ explicitly but compute the inner product of two vectors in F with an inner product kernel function:

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

Let x be a vector of the input set X with n elements.

X_i designs subsets 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. Define

between-class scatter matrix S_B and within-class matrix S_W in the feature space F as:

$$S_B = \frac{1}{C(C-1)} \sum_{i=1}^C \sum_{j=1}^C (u_i - u_j)(u_i - u_j)^T \quad (2)$$

$$S_W = \frac{1}{C} \sum_{i=1}^C \frac{1}{n_i} \sum_{j=1}^{n_i} (\phi(x_j) - u_i)(\phi(x_j) - u_i)^T \quad (3)$$

where $u_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(x_j)$ denotes the sample mean of class i in F .

To perform FLDA in F , it is equal to maximize expression (4).

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (4)$$

Because any solution $w \in F$ must lie in the span of all the samples in F , there exist coefficients α_i , $i=1, 2, \dots, n$, such that

$$w = \sum_{i=1}^n \alpha_i \phi(x_i) \quad (5)$$

By the expression (5), the projection of each class means u_i onto w can be written:

$$w^T u_i = \alpha^T \begin{pmatrix} \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_1, x_j) \\ \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_2, x_j) \\ \dots \\ \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_n, x_j) \end{pmatrix} = \alpha^T m_i \quad (6)$$

It follows that

$$w^T S_B w = \alpha^T K_B \alpha \quad (7)$$

where $K_B = \frac{1}{C(C-1)} \sum_{i=1}^C \sum_{j=1}^C (m_i - m_j)(m_i - m_j)^T$, and a similar transformation as in (7), it can be found that

$$w^T S_W w = \alpha^T K_W \alpha \quad (8)$$

where $K_W = \frac{1}{C} \sum_{i=1}^C \frac{1}{n_i} \sum_{j=1}^{n_i} (\zeta_j - m_i)(\zeta_j - m_i)^T$,

$\zeta_j = (k(x_1, x_j), k(x_2, x_j), \dots, k(x_n, x_j))^T$.

Thus, maximizing expression (4) is converted to maximize:

$$J(\alpha) = \frac{\alpha^T K_B \alpha}{\alpha^T K_W \alpha} \quad (9)$$

Similar to FLDA, this problem can be solved by finding the leading eigenvectors of $K_W^{-1} K_B$, and the projection of a point x onto w in F is given by

$$(w \cdot \phi(x)) = \sum_{i=1}^n \alpha_i k(x_i, x) \quad (10)$$

3. Experiments

In the experiments, we compare Kernel FDA with Kernel PCA and FLDA. For kernel methods, the polynomial kernel function, $k(x, y) = (x \cdot y)^d$, is selected. The selection of kernel function is lack of theoretic selection scheme. In fact, how to select kernel function for special tasks is still an open problem.

The experiments are performed on two benchmarks:

one is the *ORL* database and the other is a dataset of the *FERET* database^[13]. The Nearest Neighbor (NN) classifier is used for all the methods. On each benchmark, we reduce the dimensions to $C-1$ for Kernel FDA and FLDA, where C is the number of classes. Because there is often no enough samples for training to guarantee within-class scatter nonsingular in the application of face recognition, there is a numerical problem for Kernel FDA and FLDA. In order to deal with this numerical problem, we replace K_W with $K_W + \mu I$ for Kernel FDA, where μ is very small constant and I is the identity matrix^[14]. We adopt the method^[17] that uses PCA firstly to remove the null space of the within-class scatter for FLDA. As for Kernel PCA, we follow the rule^[18] that the ratio between the sum of the first s selected eigenvalues and the sum of all the eigenvalues is greater than 0.9, that is:

$$ratio = \frac{\sum_{i=1}^s \lambda_i}{\sum_{i=1}^{all} \lambda_i} \geq 0.9, \text{ and the number of selected}$$

eigenvectors (principal components) are empirically determined to achieve the highest recognition rate corresponding to each degree d of the polynomial kernel.

In addition, in the experiments the training set is obtained randomly each time, so perhaps there exists some fluctuation among the results. In order to reduce the fluctuation, we do each experiment more than 10 times and all the result data given in the paper is an average of them.

3.1 FERET Dataset

There are 70 persons in this dataset of the *FERET* face database. Each person has 6 different frontal-view images. There are three different illuminations and two different facial expressions in each illumination. The eye locations are fixed by geometric normalization. The images are resampled to 92×112 . Figure 1 shows 12 samples of two persons randomly selected from the dataset.



Figure 1. Samples from the *FERET* dataset

4 images per person are used for training and the other 2 images are used for testing each time. The following Table 1 is given the experimental results of Kernel FDA and Kernel PCA at different degrees $d=2,3,4$, of the polynomial kernel. From the results, it can be found that Kernel FDA outperforms Kernel PCA obviously at

each d , and Kernel FDA gives the higher recognition rate (88.93%) at $d=2$ than $d=3,4$.

Table 1. Recognition rate (%) on the *FERET* database

Degree (d)	Kernel FDA	Kernel PCA
2	88.93	79.86
3	84.90	78.15
4	80.72	74.79

The following Figure 2 is the comparison results between Kernel FDA at degree $d=2$ and FLDA. The horizontal axis represents that k images per person are selected randomly for training and the other $6-k$ images for testing.

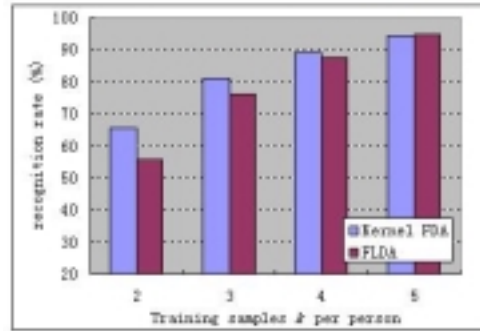


Figure 2. Kernel FDA vs FLDA on the *FERET* dataset

Figure 2 shows that Kernel FDA is better than FLDA. For example, when training samples of each person $k=2$, it gives the accuracy of 65.67%, while FLDA just has 55.82%. With k increasing to 4, it is still above 2% higher than that of FLDA (88.93% vs 87.5%), but combining with Table 1, we find that FLDA is better than Kernel PCA.

3.2 ORL Database

There are 40 persons in the *ORL* database and 10 different images with each person, including variations in pose, facial expression (open or closed eyes, smiling or non-smiling) and with glasses or no-glasses, but there is little illumination variation. The size of images is 92×112 , and no preprocessing is done. Figure 3 shows 20 randomly selected samples of two persons.



Figure 3. Samples from the *ORL* database

Each time, 5 images per person are selected randomly from the database for training and the other 5 images for testing. Kernel FDA is still compared with

Kernel PCA at degree $d=2,3,4$. Experimental results are given in the Table 2. It is similar to the experimental results on the *FERET* dataset that Kernel FDA is always better than kernel PCA and the recognition rate of Kernel FDA at $d=2$ is a little higher than $d=3,4$.

Table 2. Recognition rate (%) on the *ORL* database

Degree (d)	Kernel FDA	Kernel PCA
2	95.35	93.75
3	94.30	92.75
4	92.25	91.70

In addition, Kernel FDA at degree $d=2$ is also compared with FLDA on the *ORL* database. Each time k images per person are selected randomly for training and the other $10 - k$ images for testing. Experimental results are given in Figure 4. Similar to the experimental results of the *FERET* dataset, Kernel FDA gives higher recognition rates than FLDA at each k . When the training samples of each person $k=2$, the recognition rate of Kernel FDA is above 6% higher than FLDA (81.88% vs 75.75%). At $k=8$, Kernel FDA can achieve the recognition rate of 99.38%, even higher than the 97.50% of FLDA.

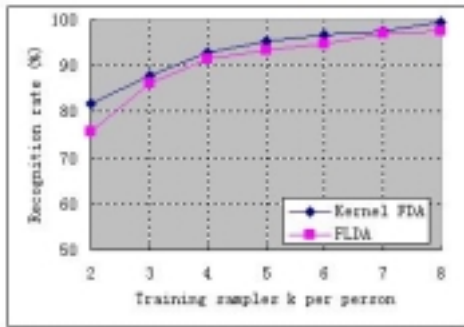


Figure 4. Kernel FDA vs FLDA on the *ORL* database

4. Discussions

From above experimental results, we find that Kernel FDA can obtain a higher recognition rate than Kernel PCA and FLDA. It shows that nonlinear discriminant with kernel is more efficient for describing the practical face images. With the different degree d of the polynomial kernel, Kernel FDA and Kernel PCA will give the different performance. In our experiments, degree $d=2$ is better than $d=3,4$, and at degree $d=4$ Kernel FDA gives recognition rate lower than FLDA when the training samples $k=4$ on the *FERET* dataset (80.72% vs 87.5%) and the training samples $k=5$ on the *ORL* database (92.25% vs 93.3%). In addition, It is also found that Kernel FDA, Kernel PCA and FLDA all perform well on the *ORL* database, which has little illumination variation, but even Kernel PCA is not as good as FLDA on the *FERET* dataset. It seems that Fisher Discrimination

Analysis is better than PCA to deal with illumination variation, which is as same as the experimental result of P.Belhumeur^[12].

5. Conclusion

Kernel FDA is a technique that combines the kernel trick with FLDA. The kernel trick is used to map the input data into an implicit feature space by nonlinear kernel functions, and then FLDA is performed in the feature space to yield a nonlinear discriminant in the input data. In this paper, we present to use it for face recognition, and compare its performance with Kernel PCA and FLDA on two benchmarks. Extensive experiments show that it can give the recognition rate higher than Kernel PCA and FLDA.

Acknowledgements:

We show our gratitude to the *FERET* program (USA) and the Olivetti Research Laboratory in Cambridge (UK) for their devotion of *FERET* and *ORL* face databases for public research.

References

- [1] Baback Moghaddam and Alex Pentland. "Probabilistic Visual Learning for Object Representation". IEEE Trans. PAMI, Vol 19, No 7, pp 696-710, 1997.
- [2] Baback Moghaddam. "Principal Manifolds and Bayesian Subspaces for Visual Recognition". Tech Report 99-35, A Mitsubishi Electric Research Laboratory, 1999.
- [3] Bernd Heisele, Purdy Ho and Tomaso Poggio. "Face Recognition with Support Vector Machines: Global versus Component-Based Approach". ICCV 2001.
- [4] B.Scholkopf, A.Smola and K.R.Muller. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". Neural Computation, vol 10, pp1299-1319, 1998.
- [5] Chengjun Liu and Harry Wechsler. "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases". IEEE Trans. Image Processing, Vol 9, No 1, 2000.
- [6] E.Osuna, R.Freund and F.Girosi. "Support Vector Machines: Training and Applications". Tech Report, AI Lab, MIT, 1997.
- [7] G.Baudat and F.Anouar. "Generalized Discriminant Analysis Using a Kernel Approach". Neural Computation, 12(10): 2385-2404, 2000.
- [8] Guodong Guo, Stan Z. Li and C.Kapluk. "Face Recognition by Support Vector Machines". In Proc. Int. Conf. Automatic Face and Gesture Reconition, 2000.
- [9] Marian Stewart Bartlett, H.Martin Lades and Terrence J.Sejnowski. "Independent Component Representations for Face Recognition". Proc. Of SPIE, 2399 : 528-539, 1998.
- [10] Ming-Hsuan Yang, Narendra Ahuja and David Kriegman. "Face Recognition Using Kernel Eigenfaces". Int. Conf. On Image Processing, 2000.
- [11] M.Turk and Pentland. "Eigenfaces for Recognition".

- J.Cognitive Neuroscience, Vol 3, No 1, 1991.
- [12] Peter N.Belhumeur, Joao P.Hespanha and David Kriegman. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection". IEEE Trans. PAMI, 19(7), pp 711-720, 1997.
- [13] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," Image and Vision Computing J, Vol. 16, No. 5, pp 295-306, 1998.
- [14] Sebastian Mika, Gunnar Ratsch, Jason Weston. "Fisher Discriminant Analysis with Kernels". Neural Networks for Signal Processing IX, pp 41-48, 1999.
- [15] Stan Z.Li, Qingdong Fu, Lie Gu, ect. "Kernel Machine Based Learning For Multi-View Face Detection and Pose Estimation". ICCV 2001.
- [16] W.Zhao, R.Chellappa and P.J.Phillips. "Subspace Linear Discriminant Analysis for Face Recognition". Tech Report CAR-TR-914, Center for Automation Research, University of Maryland, 1999.
- [17] W.Zhao, R.Chellappa, A.Rosenfeld and P.J.Phillips. "Face Recognition: A Literature Survey". CS-Tech Report-4167, University of Maryland, 2000.
- [18] W.S.Yambor, B.A.Draper and J.R.Beveridge. "Analyzing PCA-based Face Recognition Algorithms: Eigenvectors Selection and Distance Measures". 2nd Workshop on Empirical Evaluation in Computer Vision, 2000.
- [19]Yongmin Li, Shaogang Gong and Heather Liddell. "Recognising Trajectories of Facial Identities Using Kernel Discriminant Analysis", BMVC 2001.
- [20]Y.Wu, T.S.Huang, K.Toyama. "Self-Supervised Learning for Object based on Kernel Discriminant-EM Algorithm". In the proc. of ICCV 2001.