

Kernel-Based Optimized Feature Vectors Selection and Discriminant Analysis for Face Recognition

Qingshan Liu Rui Huang Hanqing Lu Songde Ma
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P.O.Box 2728, Beijing 100080, China
qslu, rhuang, luhq, masd@nlpr.ia.ac.cn

Abstract

In practice, the face image data distribution is very complex because of pose, illumination and facial expression variation, so it is inadequate to describe it just by Fisherface or Fisher linear discriminant analysis (FLDA). In this paper, a new method is presented for face recognition using kernel-based optimized feature vectors selection and discriminant analysis. With the kernel technique, an optimized data set is selected from the data and mapped into the feature space based on a geometrical approach, to form a subspace that can capture the structure of the entire data into the feature space. Then all the data are projected into this subspace and FLDA is performed in this subspace to extract nonlinear discriminant features of the data for face recognition. Another similar nonlinear discriminant analysis is Kernel-based Fisher Discriminant Analysis (KFDA), which transformed all the data into the feature space and FLDA was performed in the feature space, but the computational complexity in the proposed method is significantly reduced against KFDA. We test our method on two benchmarks, and experimental results demonstrate that it outperforms Fisherface and can give the same recognition accuracy as KFDA.

1. Introduction

In the past 20 years, face recognition has received extensive attention because of the potential applications in many fields, such as biometrics, surveillance, access control and human-computer interface. Numerous algorithms have been proposed; a detailed review was given in reference [1].

For face recognition, how to select the discriminating features, which can make the between-class scatter maximized and the within-class scatter minimized, is a key problem. Eigenface^[7] used linear Principal

Component Analysis (PCA) to extract features. However, PCA is actually optimal to representation and reconstruction for all face classes, and unnecessarily optimal to discrimination for one face class from others. Fisherface^[6] used Fisher Linear Discriminant Analysis (FLDA) to extract features, which seeks to find a linear transformation to maximize the between-class scatter and minimize the within-class scatter, and it was demonstrated to be better than Eigenface^[6], but it is still a linear technique in nature, so it is also inadequate to describe the complexity of face image data distribution in practice because of illumination, facial expression and pose variations. Bartlett^[8] 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 Moghaddam^[9] compared it with linear PCA on the Feret face database and found that it gave the same recognition accuracy as linear PCA.

Kernel-based technique is an efficient trick to analysis the complex relations of the data through mapping the data into a high dimensional implicit feature space F , $\phi: x \in R^N \rightarrow \phi(x) \in F$, with a nonlinear dot product kernel function, $k(x, y) = (\phi(x) \cdot \phi(y))$. The kernel trick was first used in Support Vectors Machine^[10] (SVM) which looks for an optimal separating hyperplane in F ; and was applied to a wide class of problems such as face recognition^[5,11]. But SVM is based on the margin maximization and sensitive only to the extreme values. Kernel-based Principal Component Analysis^[4,12] (KPCA) was developed to extract nonlinear principal components in F . However, as the same as PCA, KPCA captures the overall variance of all patterns which are inadequate for discriminating purposes. Kernel-based Fisher Discriminant Analysis^[3,13] (KFDA) combines the kernel trick with FLDA to extract nonlinear discriminant features of the data. The kernel trick is used to map all the data into the feature space, and FLDA is performed in the feature space. It's efficiency has been demonstrated in

applications^[14,15]. In KFDA, the dot product matrix K with all the training samples need to be computed, $K = (k_{i,j})_{1 \leq i,j \leq M}$, $k_{i,j} = \phi(x_i) \cdot \phi(x_j)$, (M is the number of training samples), for FLDA which performs in the implicit space F to extract nonlinear discriminant features is actually based on the matrix K .

It is known that the transformed data set of all the training data into F forms a subspace in F with a dimension up to M . But in practice the dimension of this subspace is often lower than M and equal to the rank of the matrix K , $rank(K) < M$. For this consideration, in this paper, a new method is presented with kernel-based optimized feature vectors selection and discriminant analysis for face recognition. With the kernel technique, an optimized data set is selected from the data and mapped into the feature space based on a geometrical approach^[2], to form a subspace that can capture the structure of the entire data into the feature space. We call it Kernel-based optimized feature vectors selection. Then all the data are projected into this subspace and FLDA is performed in this subspace to extract nonlinear discriminant features of the data for face recognition. Similar to KFDA, nonlinear discriminant features are extracted by the kernel trick, but the computational complexity is reduced against KFDA, because it just needs to compute and store the sufficient sub-matrix of K with the selected data. We compare it with Fisherface and KFDA on two benchmarks, and experimental results demonstrate that it outperforms Fisherface and almost gives the same recognition accuracy as KFDA.

The rest of paper is organized as follows: Kernel-based optimized feature vectors selection is introduced in Section 2, and nonlinear discriminant analysis in Section 3. Experiments are described in Section 4, and finally conclusions are given.

2. Kernel-Based Optimized Feature Vectors Selection

Because the dimensionality of the data subspace into F is given by the rank of K . In practice, the rank of the matrix K is often inferior to M , $rank(K) < M$, especially when the training data set is larger, $rank(K) \ll M$. Recently Baudat^[2] presented a idea based on geometrical approach to select a subset of the data into F to describe the structure of the data into F for reducing the memory for storage and used it in the kernel function approximation application. In this paper, we adopt this idea to select an optimized data set from the data into F to form a subspace that can capture the structure of all the data into F , before nonlinear discriminant analysis is performed.

The idea is to look for the vectors that are sufficient

to express all the data as a linear combination of those selected vectors in the transformed space F according to a geometrical consideration.

Define the mapping of x_i as $\phi(x_i) = \phi_i$, and assume $S = \{x_{s_1}, x_{s_2}, \dots, x_{s_L}\}$ to be the set of selected vectors into F , where L is the number of selected vectors. The estimation of the mapping of any vectors x_i can be regarded as a linear combination of S , given by:

$$\hat{\phi}_i = \Phi_S \cdot \alpha_i \quad (1)$$

where $\Phi_S = (\phi_{s_1}, \phi_{s_2}, \dots, \phi_{s_L})$ is the matrix of the selected vectors into F and $\alpha_i = (\alpha_i^1, \alpha_i^2, \dots, \alpha_i^L)$ is the coefficient vectors that weighted this Matrix.

We want to find the coefficients α_i such that the estimated mapping $\hat{\phi}_i$ is as close as possible to the real mapping ϕ_i , so it is equal to minimize the following expression (2):

$$\min \delta_i = \frac{\|\phi_i - \hat{\phi}_i\|^2}{\|\phi_i\|^2} \quad (2)$$

Rewriting (2) in a matrix form and putting the derivatives to zero according to α_i , expression (2) can be converted to expression (3):

$$\min \delta_i = 1 - \frac{K_{Si}^T K_{SS}^{-1} K_{Si}}{k_{ii}} \quad (3)$$

where K_{SS} is a square matrix of dot products of the selected vectors: $K_{SS} = (k_{s_p, s_q})_{1 \leq p, q \leq L}$, and $K_{Si} = (k_{s_p, i})_{1 \leq p \leq L}$ is the vectors of dot product between x_i and the selected vector set S .

The goal is to find the set S that minimizes expression (3) over all the samples x_i . Removing the constant, it is converted to maximize the expression (4):

$$\max_S J_S = \frac{1}{M} \sum_{x_i \in X} \left(\frac{K_{Si}^T K_{SS}^{-1} K_{Si}}{k_{ii}} \right) \quad (4)$$

Following above discussions, we can use an iterative process to select the optimized vector set S^* with number L^* , which stops when K_{SS} is no longer invertible^[2].

3. Nonlinear Discriminant Analysis

Once the optimized feature vectors are selected, they define a subspace Φ_{S^*} in F that captures the structure of the entire data. The transformation of a sample x_i is projected into this subspace given by:

$$z_i = \Phi_{S^*}^T \phi_i = (k_{s_p,i})_{1 \leq p \leq L^*} \quad (5)$$

By expression (5), transform all the data into the subspace. If define the matrix $Z = (z_i)_{1 \leq i \leq M}$, then Z is actually just the sub-matrix of K . Now we perform FLDA in this subspace, thus nonlinear discriminant features of input data are extracted.

Assuming input data set X with c classes, and class i has m_i samples: $\sum_{i=1}^c m_i = M$. Define the between-class scatter S_B and within-class scatter S_W in this subspace as the following:

$$S_B = \sum_{i=1}^c m_i (u_i - u)(u_i - u)^T \quad (6)$$

$$S_W = \sum_{i=1}^c \sum_{j=1}^{m_i} (z_j - u_i)(z_j - u_i)^T \quad (7)$$

where $u = \frac{1}{M} \sum_{i=1}^M z_i$ is the total mean and class means

are given by $u_i = \frac{1}{m_i} \sum_{j=1}^{m_i} z_j$, $i=1,2,\dots,c$.

To find the linear discriminant in this subspace, just need to maximize the expression (8):

$$J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega} \quad (8)$$

It is equal to find the leading eigenvector of $S_W^{-1} S_B$. The projection of a new pattern x onto ω is given by:

$$y = \omega^T z(x) \quad (9)$$

where $z(x) = (k(x, x_{s_1}), k(x, x_{s_2}), \dots, k(x, x_{s_L}))^T$.

4. Experiments

The proposed method is tested by two group experiments corresponding to two benchmarks, and it is compared with Fisherface and KFDA. A polynomial kernel function $k(x, y) = (x \cdot y)^n$ is selected with $n=2$ for our method and KFDA in the experiments. The selection is empirical. The nearest neighbor classifier is used for all the three methods.

First Group: The experimental dataset consists of 420 facial images corresponding to 70 persons selected from the US Army *FERET*^[16] 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 and the images are resized to 92×112 . Figure 1 shows 12 samples of two persons randomly

selected from the dataset.



Figure 1. Samples from the *FERET* dataset

We use different training samples and testing samples to test our method and compare with Fisherface and KFDA. $k=2,3,4,5$ samples of each person are randomly selected from the dataset for training and the other $6-k$ samples of each person for testing. For each k , the algorithms run more than 10 times and the following table 1 is given the average recognition rate.

Table 1. Recognition rate on the *FERET* dataset

k	Our method	KFDA	Fisherface
2	65.50	65.67	55.82
3	80.96	80.86	76.10
4	88.79	88.93	87.50
5	94.43	94.29	94.57

Table 1 showed that our method was better than Fisherface. For example, when training samples of each person $k=2$, our method gave the accuracy of 65.5%, while Fisherface just had 55.82%. With k increased to 4, our method also had a higher recognition rate than Fisherface (88.79% vs 87.50%). At $k=5$, the recognition rate of Fisherface was similar to our method, because there had enough training samples corresponding to testing samples (5:1). In addition, our method almost has the same recognition rate as KFDA at each k .

Second Group: There are 40 persons in the *ORL* database and 10 different images with each person, including variations in pose, facial expression (open/closed eyes, smiling/non-smiling) and with glasses/no-glasses, but there is little illumination variation. The size of images is 92×112 and no preprocess is done. Figure 2 shows 20 randomly selected samples of two persons.



Figure 2. Samples from the *ORL* database

In this group experiments, we test the methods with different training samples and testing samples corresponding the training number $k=2,3,4,5,6,7,8$ of each subject. Each time randomly select k samples from each subject to train and the other $10-k$ to test. The following table 2 is given an average recognition rate of more than 10 times for each k .

Table 2: Recognition rate on the ORL database

k	Our method	KFDA	Fisherface
2	81.88	81.88	73.75
3	87.82	87.82	85.14
4	92.59	92.70	89.38
5	95.25	95.35	91.65
6	96.63	96.68	93.88
7	97.50	97.50	94.83
8	99.38	99.38	97.50

In this group experiments, our method gave higher recognition rates than Fisherface at each k and almost gave the same recognition rate as KFDA at each k too. When the training samples of each person $k=2$, the recognition rate of our method is above 7% higher than Fisherface (81.88% vs 73.75%). At $k=8$, our method could achieve the recognition rate of 99.38%, even higher than the 97.50% of Fisherface. These experimental results are similar to the experimental results of the first group.

From above two group experimental results, we could conclude that nonlinear discriminant features achieved by kernel technique and discriminant analysis were better than linear discriminant features used in Fisherface to describe the complex variations of pose, illuminate and facial expression. Though our method and KFDA all could extract nonlinear discriminant features and almost gave the same recognition rate, from analysis in Section 2 and Section 3, we knew that our method just needed to compute and store the sufficient sub-matrix of K , thus, the complexity of computation was reduced. It also showed that feature vectors selection process in our method did not lost any important information against KFDA.

5. Conclusions

It was inadequate to describe well the complication of face images in practice just by linear discriminant features, such as Fisherface. In this paper, a novel method was presented for face recognition. The kernel trick was used to extract an optimized feature vectors into the feature space F and formed a subspace to capture the structure of the whole data into F , then nonlinear discriminant features were extracted from this subspace. Another similar nonlinear discriminant analysis was KFDA, but our method reduced significantly the computational complexity without losing any important information against KFDA. Experimental results demonstrated that this method outperformed Fisherface and almost gave the same recognition accuracy as KFDA.

Acknowledgments:

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.

We would also like to thank Gaston Baudat for useful discussions.

Reference

- [1] W.Zhao, R.Chellappa, A.Rosenfeld and P.J.Phillips. "Face Recognition: A Literature Survey". *CS-Tech Report-4167*, University of Maryland, 2000.
- [2] G.Baudat and F.Anouar. "Kernel-based Methods and Function Approximation" pp. 1244 -- 1249, Washington, DC July 15 - 19, 2001.
- [3] G.Baudat and F.Anouar. "Generalized Discriminant Analysis Using a Kernel Approach". *Neural Computation*, 12(10): 2385-2404, 2000.
- [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] Guodong Guo, Stan Z. Li and C.Kapluk. "Face Recognition by Support Vector Machines". *In Proc. Int. Conf. Automatic Face and Gesture Reconition*, 2000.
- [6] Peter N.Belhumeur, Joao P.Hespanha and David J.Kriegman. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection". *IEEE Trans. PAMI*, 19(7), pp 711-720, 1997.
- [7] M.Turk and Pentland. "Eigenfaces for Recognition". *J.Cognitive Neuroscience*, Vol 3, No 1, 1991.
- [8] Marian Stewart Bartlett, H.Martin Lades and Terrence J.Sejnowski. "Independent Component Representations for Face Recognition". *Proc. Of SPIE*, 2399 : 528-539, 1998.
- [9] Baback Moghaddam. "Principal Manifolds and Bayesian Subspaces for Visual Recognition". *Tech Report 99-35*, A Mitsubishi Electric Research Laboratory, 1999.
- [10] E.Osuna, R.Freund and F.Girosi. "Support Vector Machines: Training and Applications". *Tech Report*, AI Lab, MIT, 1997.
- [11] Bernd Heisele, Purdy Ho and Tomaso Poggio. "Face Recognition with Support Vector Machines: Global versus Component-Based Approach". *ICCV 2001*.
- [12] Ming-Hsuan Yang, Narendra Ahuja and David Kriegman. "Face Recognition Using Kernel Eigenfaces". *Int. Conf. On Image Processing*, 2000.
- [13] Sebastian Mika, Gunnar Ratsch, Jason Weston. "Fisher Discriminant Analysis with Kernels". *Neural Networks for Signal Processing IX*, pp 41-48, 1999.
- [14] Yongmin Li, Shaogang Gong and Heather Liddell. "Recognising Trajectories of Facial Identities Using Kernel Discriminant Analysis", *Bmvc* 2001.
- [15] Qingshan Liu, Rui huang, Hanqing Lu and Songde Ma. "Face Recognition Using Kernel-based Fisher Discriminant Analysis". *Tech Report*, National Lab of Pattern Recognition, China. September, 2001.
- [16] 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.