
Comparing Kernel-based Learning Methods for Face Recognition

Zhiguo Li

Department of Computer Science, Rutgers University, NJ 08854 USA

ZHLI@PAUL.RUTGERS.EDU

Abstract

Principal Component Analysis (PCA) and Fisher Discriminant Analysis (FDA) have been successfully applied to face recognition, and both are based on the second order statistics of the image set. Kernel-based subspace methods try to capture the higher order statistics of the image set and thus may provide better results for recognition purposes. In this paper, we try to compare different algorithms for face recognition and find out if kernel-based methods are better than linear version methods for face recognition, and we also explore the experimental results.

1. Introduction

Face recognition has received extensive attention because of the potential applications in many fields, including identity authentication, surveillance and human-computer interaction.

Linear subspace analysis methods have been used for face recognition by many researchers because of their simplicity and efficiency. [such as PCA (Turk, 1991) and FDA (Belhumeur, 1997)] The representations in these subspace methods are based on second order statistics of the image set, and they do not address higher order statistical dependencies such as the relationships among three or more pixels, which may capture important information for recognition.

The kernel trick was first used in support vector machines (SVMs), and a review of kernel-based learning algorithms was given by Müller (2001). The kernel trick is to project the input lower dimensional data into an implicit higher dimensional space called feature space by nonlinear kernel mapping. In this way, data which is not linearly separable in low dimensional space could be linearly separable in higher dimensional feature space. Kernel PCA (Schölkopf, 1998) and Kernel FDA (Mika, 1999) are two such algorithms. They only depend on inner products in the feature space and need not compute the feature space explicitly. Recently kernel-based subspace methods have been used in face recognition studies (Liu, 2002; Yang, 2002).

In this paper, we compare the kernel-based subspace methods for face recognition with the classical linear subspace methods. In the next two sections, we will first review PCA, Kernel PCA, FDA and Kernel FDA algorithms. Experimental results based on these algorithms on a face recognition problem are given in Section 4. In Section 5 we discuss the implications and then conclude in Section 6.

2. PCA and Kernel PCA

For PCA, let us consider a set of N samples \mathbf{x}_i , $\mathbf{x}_i = [x_{i1}, \dots, x_{in}]^T \in R^n$. PCA aims to find a linear transformation mapping the original n -dimensional space into an m -dimensional space, where $m < n$. Denoting $W \in R^{n \times m}$, a matrix with orthogonal columns, the new vectors \mathbf{y}_k are defined by $\mathbf{y}_k = W^T \mathbf{x}_k$. Let the total scatter matrix S_t be defined

as $S_t = \sum_{k=1}^N (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T$, where $\mathbf{m} \in R^n$ is the mean

image of all samples. After applying the linear transformation, the scatter of the transformed feature vectors $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ is $W^T S_t W$. In PCA, the optimal projection W_{opt} is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e. $W_{opt} = \arg \max_W |W^T S_t W| = [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_m]$, where $\{\mathbf{w}_i | i = 1, 2, \dots, m\}$ is the set of n -dimensional eigenvectors of S_t corresponding to the set of decreasing eigenvalues. Since these eigenvectors have the same dimension as the original images, they are referred to as Eigenfaces (Turk, 1991).

In Kernel PCA, each vector \mathbf{x} is first projected from the input space, R^n , to a high dimensional feature space, R^f , by a nonlinear mapping function: $\Phi: R^n \rightarrow R^f$, $f > n$. In R^f , the corresponding eigenvalue problem

is $I\mathbf{w}^\Phi = C^\Phi \mathbf{w}^\Phi$, where C^Φ is the covariance matrix. It's easy to show that all solutions \mathbf{w}^Φ with $I \neq 0$ lie in the span of $\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}_N)$, and there exists

coefficients a_i such that $\mathbf{w}^\Phi = \sum_{i=1}^N a_i \Phi(\mathbf{x}_i)$. Denoting

an $N \times N$ matrix K by $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, the Kernel PCA problem becomes $INK\mathbf{a} = K^2\mathbf{a} \equiv mI\mathbf{a} = K\mathbf{a}$, where \mathbf{a} denotes a column vector with entries a_1, \dots, a_N .

We can now project the vectors in R^f to a lower dimensional space spanned by the eigenvectors \mathbf{w}^Φ . Let \mathbf{x} be a test sample whose image is $\Phi(\mathbf{x})$ in R^f , then the projection of $\Phi(\mathbf{x})$ onto the eigenvectors \mathbf{w}^Φ is the nonlinear principal components corresponding to $\Phi(\mathbf{x})$:

$$\mathbf{w}^\Phi \cdot \Phi(\mathbf{x}) = \sum_{i=1}^m a_i (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) = \sum_{i=1}^m a_i k(\mathbf{x}_i, \mathbf{x}) .$$

In other words, we can extract the first q ($1 \leq q \leq m$) nonlinear principal components using the kernel function without the expensive operation that explicitly projects samples to a high dimensional space R^f . Classical PCA is a special case of Kernel PCA with first order polynomial kernel. In other words, Kernel PCA is a generalization of classical PCA since different kernels can be used for different nonlinear projections. For a detailed derivation, refer to the paper by Schölkopf (1998).

3. FDA and Kernel FDA

It should be noted that PCA is inadequate for discriminant purposes between different faces because PCA seeks to maximize the total scatter across all classes. Some unwanted variations (due to changes in illumination, facial expressions, pose, etc.) may be retained. It has been observed that in face recognition the variations between the face images of the same person due to illumination and pose are almost always larger than image variations due to the changes in face identity. Therefore, while the PCA projections are optimal in a correlation sense, these eigenvectors are not optimal from the classification viewpoint.

FDA is used to seek a projection W from the original space to a lower-dimensional space that maximizes the between-class scatter while minimizing the within-class scatter. Figure 1 illustrates the difference between PCA and FDA. For the two class artificial data, after projecting

onto the projection direction found by PCA, the two classes are totally messed up. But for FDA, after projecting onto the projection direction found by FDA, the two classes are separated perfectly.

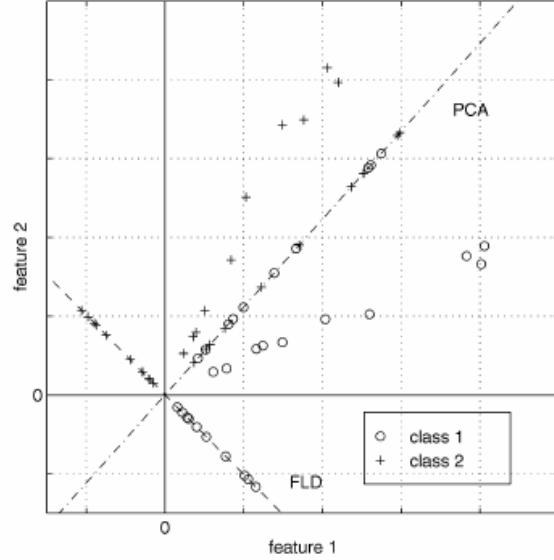


Figure 1. PCA vs. FDA

A typical way to achieve the goal of FDA is to maximize

the ratio: $\frac{|W^T S_b W|}{|W^T S_w W|}$, where

$$S_b = \frac{1}{C(C-1)} \sum_{i=1}^C \sum_{j=1}^C (u_i - u_j)(u_i - u_j)^T, \text{ and}$$

$$S_w = \frac{1}{C} \sum_{i=1}^C \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_j - u_i)(\mathbf{x}_j - u_i)^T .$$

S_b is the between-class scatter matrix and S_w is the within-class scatter matrix. The ratio is maximized when the column vectors of W are the eigenvectors of $S_w^{-1} S_b$.

For the derivation of KFDA, it's similar to the case of KPCA. The key point is to write the algorithm in the form of dot products in high dimensional space. For a detailed derivation, see the paper by Mika (1999).

4. Experiments

In our experiments, we compare the face recognition rate of the four algorithms: PCA, Kernel PCA, FDA and Kernel FDA on the publicly available AT&T, Yale and FERET face databases (Phillips, 1998). The statistics for these face databases are indicated in Table 1. Figure 2 shows sample face images from the AT&T database. We test the recognition rates with different numbers of

training examples. k images of each subject are randomly selected for training and the remaining images for that subject are used for testing. For each value of k , 10 runs are performed with different random partitions between the training and testing sets. Recognition was performed based on a nearest neighbor classifier. Figures 3 shows the change in recognition rate with respect to the number of training images, and “degree” in these figures means the degree of the polynomial kernel function. Figures 4-6 and Table 2 compare the recognition rate for different methods, where with KFDA and KPCA we use degree 2 polynomial kernel functions, with FDA and PCA, we use the largest $C - 1$ eigenvectors, and with “PCA all” we use all $C \cdot k - 1$ eigenvectors. We make the following observations based on the experimental results. (1) Normally lower order polynomial kernel functions achieve higher recognition rates. (2) The recognition rate increases when the number of training images for each subject increases. (3) FDA methods always outperform PCA methods. (4) The PCA method with all eigenvectors is better than PCA with only $C - 1$ eigenvectors, but with the cost of more computation.



Figure 2. Sample face images from the AT&T database

Table 1. Face databases we used in our experiments.

DATABASE	NUM OF FACES	NUM OF SUBJECTS	VARIATIONS
AT&T	400	40	Pose, expression, makeup
FERET	420	70	Illumination, expression
YALE	165	15	Illumination, expression

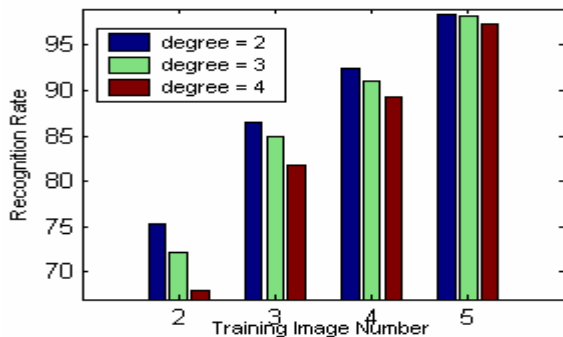


Figure 3. Results of Kernel FDA on FERET face database

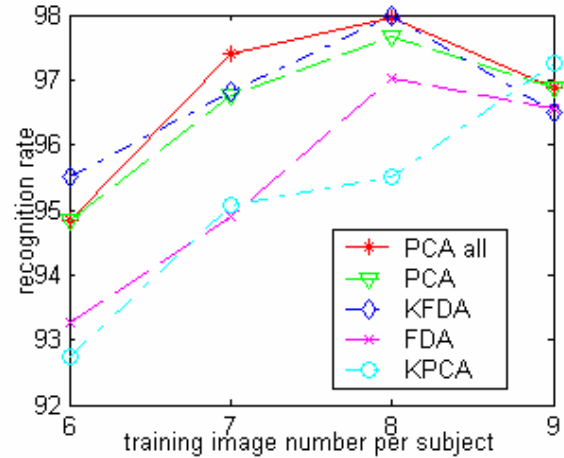


Figure 4. Recognition rate on the AT&T face dataset

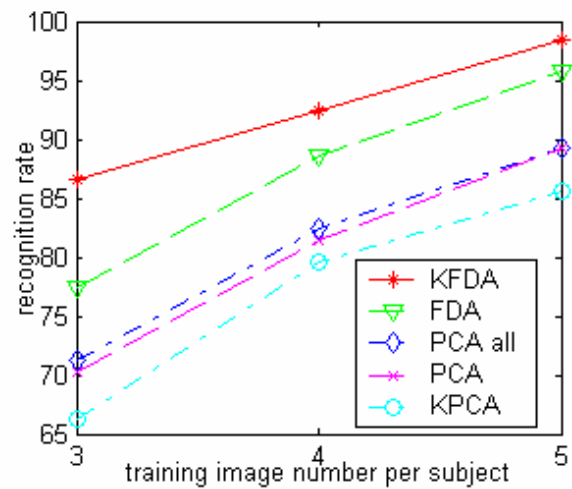


Figure 5. Recognition rate on the FERET face dataset.

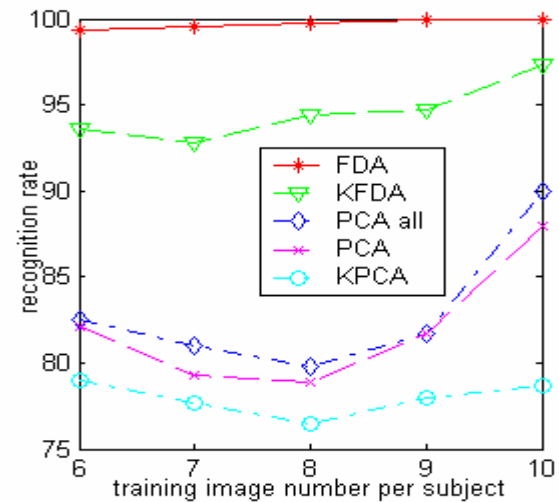


Figure 6. Recognition rate on the Yale face dataset.

Table 2. Recognition results on FERET database.

METHOD	TRAIN NUM = 2	TRAIN NUM = 3	TRAIN NUM = 4
PCA	70.3333	81.3571	89.2857
PCA WITH ALL EIGENVECTORS	71.1905	82.3571	89.2857
FDA	77.4286	88.5714	95.7143
KFDA	86.5714	92.3571	98.4286
KPCA	66.3333	79.5714	85.5714

5. Discussions

- As we can see from the experimental results, kernel methods are not necessarily better than the linear subspace methods. Why is not there advantage for kernel methods over linear methods? One possible explanation is that although nonlinear kernel mapping can make data more linear separable in high dimensional feature space than in the input space, it does not make the between-class distances far enough so that we can improve recognition rate using NN. The other possible reason is that for face data, the original dimension is already very high, maybe it does not make much sense to mapping the face data to a even higher dimensional space
- One concern about the kernel methods is that they could be computationally expensive, but in fact it is very feasible. In the test on FERET database, when using 210 face images as the training data, it takes 18.2s to train kernel FDA and 14.1s for kernel PCA on a P4 2.6GHz machine, and most of the training time are spent on finding the eigenvectors. Why kernel methods are efficient? Firstly we are just working in the subspace of the full space R^f spanned by the sample images. Secondly, we do not need to compute dot products explicitly between vectors in high dimensional feature space R^f , as we are using kernel functions.
- Experiments tell us FDA based methods are better than PCA based methods, especially when the dataset contains variations on illumination. It has been observed that in face recognition the variations between the face images of the same person due to illumination and pose are almost always larger than image variations due to the changes in face identity. Thus for PCA, if the first few principal components capture the variation due to lighting, then better clustering of projected samples is achieved by ignoring them.

6. Conclusions and future work

While good results have been reported (Schölkopf, 1998; Mika, 1999) using Kernel PCA and Kernel FDA with application to handwriting character recognition on the USPS dataset, we found that for face recognition, there is no big difference between kernel methods and linear methods. Due to variations in facial expression, pose and illuminations, it is really hard to estimate the true distribution of face dataset in high dimensional space. Furthermore, although Mercer's theorem gives the condition under which we can construct kernel functions, the selection of kernel functions under particular situation lacks theoretic basis. Future work will try to use the Artificial Neural Network (ANN) and multi-class SVM instead of Nearest Neighbor (NN) as the classifier.

Acknowledgements

We thank the FERET program, AT&T and Yale for their devotion of face database for public research.

References

- Belhumeur, P., Hespanha, J. & Kriegman, D. (1997). Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720.
- Liu, Q., Huang, R., Lu, H. & Ma, S. (2002). Face recognition using kernel based fisher discriminant analysis. *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Page(s): 187-191.
- Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Muller, K. (1999). Fisher discriminant analysis with kernels. *IEEE Neural Networks for Signal Processing Workshop*, pages 41-48.
- Müller, K. R., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001). An introduction to Kernel-Based learning algorithms. *IEEE Trans. Neural Networks*, 12(2):181-201.
- Phillips, P.J., Wechsler, H., Huang, J. & Rauss, P. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295-306.
- Schölkopf, B., Smola, A., & Muller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299-1319.
- Turk, M., & Pentland, A. (1991). Eigen Faces for Recognition. *Journal of Cognitive Neuroscience*, 3(1).
- Yang, M. H. (2002). Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods. *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*.