PERGAMON

Pattern Recognition 36 (2003) 563–566

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

Why can LDA be performed in PCA transformed space?

Jian Yang*, Jing-yu Yang

Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China

Received 24 December 2001; accepted 22 January 2002

Abstract

PCA plus LDA is a popular framework for linear discriminant analysis (LDA) in high dimensional and singular case. In this paper, we focus on building a theoretical foundation for this framework. Moreover, we point out the weakness of the previous LDA based methods, and suggest a complete PCA plus LDA algorithm. Experimental results on ORL face image database indicate that the proposed method is more effective than the previous ones. © 2002 Published by Elsevier Science Ltd on behalf of Pattern Recognition Society.

Keywords: Linear discriminant analysis (LDA); PCA plus LDA; Complete PCA plus LDA algorithm; Feature extraction; Face recognition

1. Introduction

Linear discriminant analysis (LDA) has been successfully applied in many classification problems such as image recognition, multimedia information retrieval and so on. However, for the high-dimensional and small sample size problem such as face identification, the traditional LDA encounters two aspects of difficulties [1,2]. First, the traditional algorithm cannot be used directly in that the within-class scatter matrix is always singular. Second, the high-dimensional image vectors lead to computationally difficulty.

In order to avoid these difficulties, a very popular technique usually called PCA plus LDA [1–3] is proposed and widely utilized subsequently. In this method, the principal component analysis (PCA) is first used for dimensionality reduction before the application of LDA. Although PCA plus LDA approach has been verified effective by experience, the theoretical foundation of this method is still not clear. Why select PCA for dimensionality reduction beforehand? Is there any important discriminatory information lost in the PCA process since the criterion of PCA is not identical to that of LDA? These essential problems still remain unsolved.

In this paper, we intend to solve these problems and build a theoretical foundation for the PCA plus LDA method. Moreover, we point out the weakness of the previous LDA based methods, and suggest a complete PCA plus LDA algorithm. Experimental results indicate that the proposed method is more effective.

2. The essence of LDA in singular case: PCA plus LDA

Suppose there are c known pattern classes, S_b, S_w and S_t denote the between-class scatter matrix, within-class scatter matrix and total scatter matrix, respectively. As we know, they are all semi-positive definite, and satisfy $S_t = S_b + S_w$.

The classical Fisher criterion function is generally defined by

$$J_f(X) = \frac{X^{\mathrm{T}} S_b X}{X^{\mathrm{T}} S_w X} \quad \text{or} \quad J(X) = \frac{X^{\mathrm{T}} S_b X}{X^{\mathrm{T}} S_t X}. \tag{1}$$

In the singular case, the latter one is usually adopted. And a set of optimal discriminant vectors (projection axes) based on this criterion is required. Now, the problem is where to

* Corresponding author. Tel.: +86-25-4316840; fax: +86-25-4315510.

E-mail addresses: tuqingh@mail.njust.edu.cn (J. Yang), yangjy@mail.njust.edu.cn (J.-yu Yang).

0031-3203/02/$22.00 © 2002 Published by Elsevier Science Ltd on behalf of Pattern Recognition Society.
PII: S0031-3203(02)00048-1

find them. Naturally, we can find them in R^n. But it is too difficult in that the dimension is very high and S_t is always singular. Fortunately, they can be derived from a much lower dimensional subspace of R^n by the following theory.

Suppose $\beta_1, \beta_2, \ldots, \beta_n$ are n orthonormal eigenvectors of S_t, and the first m ($m = \operatorname{rank} S_t$) ones are corresponding to positive eigenvalues. Define the subspace $\Phi_t = \operatorname{span}\{\beta_1, \beta_2, \ldots, \beta_m\}$, and its orthogonal complement can be denoted by $\Phi_t^{\perp} = \operatorname{span}\{\beta_{m+1}, \ldots, \beta_n\}$. Obviously, $\Phi_t^{\perp}$ is the null space of S_t.

Since S_b, S_w and S_t are all semi-positive definite and $S_t = S_b + S_w$, it is easy to get

Lemma 1. *If S_t is singular, $X^{\mathrm{T}} S_t X = 0$ if and only if $X^{\mathrm{T}} S_w X = 0$ and $X^{\mathrm{T}} S_b X = 0$.*

Theorem 1. *For any arbitrary $\varphi \in R^n$, φ can be denoted by $\varphi = X + \xi$, where, $X \in \Phi_t$ and $\xi \in \Phi_t^{\perp}$, and satisfies $J(\varphi) = J(X)$.*

Proof. Since $R^n = \operatorname{span}\{\beta_1, \beta_2, \ldots, \beta_n\}$, by the definition of Φ_t and $\Phi_t^{\perp}$, for any arbitrary $\varphi \in R^n$, φ can be denoted by $\varphi = \underbrace{\lambda_1\beta_1 + \cdots + \lambda_m\beta_m}_{m} + \underbrace{\lambda_{m+1}\beta_{m+1} + \cdots + \lambda_n\beta_n}_{n-m} = X + \xi$, where $X \in \Phi_t$ and $\xi \in \Phi_t^{\perp}$.

Since $\xi \in \Phi_t^{\perp}$, it follows that $\xi^{\mathrm{T}} S_t \xi = 0$.

By Lemma 1, we have $\xi^{\mathrm{T}} S_b \xi = 0$, which imply that $S_b \xi = 0$ since S_b is semi-positive definite. Hence $\varphi^{\mathrm{T}} S_b \varphi = \xi^{\mathrm{T}} S_b \xi + 2X^{\mathrm{T}} S_b \xi + X^{\mathrm{T}} S_b X = X^{\mathrm{T}} S_b X$.

Similarly, $\varphi^{\mathrm{T}} S_t \varphi = X^{\mathrm{T}} S_t X$.

So $J(\varphi) = J(X)$. □

According to Theorem 1, we can conclude that all optimal discriminant vectors can be derived from Φ_t without any loss of the optimal discriminatory information with respect to Fisher criterion $J(X)$.

Now, the problem is how to find the optimal discriminant vectors in Φ_t. By linear algebra theory, Φ_t is isomorphic to m-dimensional Euclidean space R^m. And the corresponding isomorphic mapping is

$$X = PY, \quad \text{where} \quad P = (\beta_1, \beta_2, \ldots, \beta_m), \quad X \in \Phi_t \quad \text{and} \quad Y \in R^m. \tag{2}$$

By the isomorphic mapping $X = PY$, the criterion function $J(X)$ becomes

$$J(X) = \frac{Y^{\mathrm{T}}(P^{\mathrm{T}} S_b P) Y}{Y^{\mathrm{T}}(P^{\mathrm{T}} S_t P) Y} = \frac{Y^{\mathrm{T}} \tilde{S}_b Y}{Y^{\mathrm{T}} \tilde{S}_t Y} = \tilde{J}(Y), \tag{3}$$

where $\tilde{S}_b = P^{\mathrm{T}} S_b P$, $\tilde{S}_t = P^{\mathrm{T}} S_t P$. It is easy to prove that $\tilde{S}_b$ is semi-positive definite and $\tilde{S}_t$ is positive definite. That means $\tilde{J}(Y)$ can act as a criterion like Fisher criterion. By the property of isomorphic mapping and Eq. (3), we have

Proposition 1. *Suppose $Y_1, Y_2, \ldots, Y_d$ are optimal discriminant vectors based on $\tilde{J}(Y)$, then, $X_1 = PY_1$, $X_2 = PY_2, \ldots, X_d = PY_d$ are the required optimal discriminant vectors based on $J(X)$.*

Then, the linear discriminant transformation can be defined as follows:

$$Z = W^{\mathrm{T}} X, \tag{4}$$

where $W^{\mathrm{T}} = (X_1, X_2, \ldots, X_d)^{\mathrm{T}} = (PY_1, PY_2, \ldots, PY_d)^{\mathrm{T}} = (Y_1, Y_2, \ldots, Y_d)^{\mathrm{T}} P^{\mathrm{T}}$.

The transformation in Eq. (4) can be divided into two items

$$Y = P^{\mathrm{T}} X, \quad \text{where } P = (\beta_1, \beta_2, \ldots, \beta_m) \tag{5}$$

and

$$Z = V^{\mathrm{T}} Y, \quad \text{where } V = (Y_1, Y_2, \ldots, Y_d). \tag{6}$$

Since the column vectors of P are eigenvectors corresponding to nonzero eigenvectors of S_t, the transformation in Eq. (5) is exactly PCA which transform R^n into R^m. In the transformed space R^m, it is easy to get that the total scatter matrix is $\tilde{S}_t = P^{\mathrm{T}} S_t P$ and the between-class scatter matrix is $\tilde{S}_b = P^{\mathrm{T}} S_b P$. Thus, the criterion $\tilde{J}(Y)$ is exactly the Fisher criterion in PCA transformed space, and $Y_1, Y_2, \ldots, Y_d$ are the corresponding Fisher optimal discriminant vectors.

Now, the essence of LDA in singular case is revealed. That is, PCA is first used to reduce the dimension of image space to m (the rank of the total scatter matrix). Then, LDA is performed in the transformed space.

3. How to perform LDA in the PCA transformed space

Although the Fisherfaces [1] and EFM [3] methods both follow the PCA plus LDA strategy, they are imperfect in that some small principal components are thrown away in the PCA step. So some potential and valuable discriminatory information is lost in this step. Rather, in this section, we propose a complete LDA method that is capable of deriving all discriminatory information. In PCA step, we use all positive principal components and transform the image space into R^m, where $m = \operatorname{rank} S_t$. Then, we use the OFLD [4] method for the second feature extraction. The idea of the algorithm is described as follows. In PCA transformed space R^m, split the within-class scatter matrix $\tilde{S}_w$ into its null space $\tilde{\Phi}_w^{\perp} = \operatorname{span}\{\gamma_{q+1}, \ldots, \gamma_m\}$ and its orthogonal complement $\tilde{\Phi}_w = \operatorname{span}\{\gamma_1, \ldots, \gamma_q\}$, where $\gamma_1, \ldots, \gamma_m$ are orthonormal eigenvectors of $\tilde{S}_w$, and the first q ones are corresponding to positive eigenvalues. In fact, it can be verified all discriminatory information with respect to Fisher criterion is contained in these two subspaces [4]. Since for any

Table 1
Comparison of the maximal recognition rates of the five LDA based methods with a minimum distance classifier

Number of training sample	Proposed	Fisherface [1]	EFM [3] $m = 50$	NLDA [5]	DLDA [2]
3	92.5%(52)	87.5%(38)	87.9%(36)	91.8%(39)	87.9%(35)
4	95.4%(55)	88.7%(39)	92.1%(30)	94.6%(39)	90.8%(22)
5	97.0%(41)	88.5% (39)	93.5%(39)	96.0%(39)	94.0%(34)

Note: In this table, $m = 50$ is the number of the selected principal components in PCA step of EFM, and the value in ( ) denotes features number as the maximal recognition rate is achieved.

Table 2
Classification errors of the proposed method as axes number varying

Classifier	39	40	41	42	43	44	45	46	47	48	49
Minimum distance	8	7	6	6	6	6	7	7	8	7	8
Nearest neighbor	8	8	8	6	6	6	7	7	7	8	8

nonzero vector Y in $\tilde{\Phi}_w^{\perp}$, the within-class scatter $Y^{\mathrm{T}}\tilde{S}_w Y = 0$ and the between-class scatter $Y^{\mathrm{T}}\tilde{S}_b Y > 0$, so the Fisher criterion $\tilde{J}(Y)$ can be replaced by $\tilde{J}_b(Y) = Y^{\mathrm{T}}\tilde{S}_b Y$. While, for any nonzero vector Y in $\tilde{\Phi}_w$, $Y^{\mathrm{T}}\tilde{S}_w Y > 0$, so the Fisher criterion $\tilde{J}(Y)$ is still applicable. The isomorphic mapping technique mentioned above is employed again for the calculation of the Fisher optimal discriminant vectors based on $\tilde{J}_b(Y)$ (or $\tilde{J}(Y)$) in $\tilde{\Phi}_w^{\perp}$ (or $\tilde{\Phi}_w$). The detailed algorithm is as follows.

Step 1. In PCA transformed space R^m, work out the within-class scatter matrix $\tilde{S}_w$'s orthonormal eigenvectors $\gamma_1, \ldots, \gamma_m$, suppose the first q ones are corresponding to positive eigenvalues.

Step 2. Let $P_1 = (\gamma_{q+1}, \ldots, \gamma_m)$ and $\bar{S}_b = P_1^{\mathrm{T}}\tilde{S}_b P_1$, work out $\bar{S}_b$'s orthonormal eigenvectors $Z_1, \ldots, Z_l$, then, the optimal discriminant vectors contained in $\tilde{\Phi}_w^{\perp}$ are $Y_j = P_1 Z_j$, $j = 1, \ldots, l$. Generally, $l = c - 1$, c is the number of classes.

Step 3. Let $P_2 = (\gamma_1, \ldots, \gamma_q)$ and $\hat{S}_b = P_2^{\mathrm{T}}\tilde{S}_b P_2$, $\hat{S}_t = P_2^{\mathrm{T}}\tilde{S}_t P_2$, work out d–l generalized eigenvectors $Z_{l+1}, \ldots, Z_d$ of $\hat{S}_b$ and $\hat{S}_t$ corresponding to the first d–l largest eigenvalues. Then, the optimal discriminant vectors derived from $\tilde{\Phi}_w$ are $Y_j = P_2 Z_j, j = l + 1, \ldots, d$.

Step 4. Let $Y_j = P_1 Z_j (j = 1, \ldots, l)$ and $Y_j = P_2 Z_j (j = l + 1, \ldots, d)$ act as projection axes to form the feature extractor $\Phi = (Y_1, \ldots, Y_l, Y_{l+1}, \ldots, Y_d)$.

4. Experiment

The proposed method is tested on the ORL face image database (http://www.cam-orl.co.uk). There are 10 different images of 40 distinct subjects. There are variations in facial expression (open/closed eyes, smiling/nonsmiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an up-right, frontal position, with tolerance for some tilting and rotation of up to about 20°. There is some variation in scale of up to about 10%. The size of each image is 92×112.

In this experiment, we use the first k ($k = 3, 4, 5$, respectively) images of each person for training and the remaining for testing. The Fisherfaces [1], EFM [3], NLDA [5], DLDA [2] and the proposed algorithm are, respectively, used for feature extraction. In the transformed space, a minimum distance classifier is employed. The recognition accuracy is listed in Table 1. And, as the axes numbers varying from 39 to 49, the classification errors of the proposed method with a minimum distance classifier and a nearest neighbor classifier are shown in Table 2. Table 1 shows the performance of the proposed method is better than the others'. Table 2 indicates the classification results are very robust with the variation of axes number.

Acknowledgements

We wish to thank National Science Foundation of China under Grant No. 60072034 for supporting.

References

[1] P.N. Belhumeur, et al., Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[2] Hua Yu, Jie Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognition 34 (11) (2001) 2067–2070.

[3] Chengjun Liu, Harry Wechsler, Robust coding schemes for indexing and retrieval from large face databases, IEEE Trans. Image Process. 9 (1) (2000) 132–137.
[4] Jian Yang, J.Y. Yang, Optimal FLD algorithm for facial feature extraction, SPIE Proceedings of the Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision, October, Vol. 4572, 2001, pp. 438–444.
[5] H.-Y. Li-Fen Chen, et al., A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

About the Author—JIAN YANG was born in Jiangsu, China, on 3rd June 1973. He received his M.S. degree in Applied Mathematics from Changsha Railway University in 1998. Now, he is a teacher in the Department of Applied Mathematics of Nanjing University of Science and Technology (NUST). At the same time, he is working for his Ph.D. degree in Pattern Recognition and Intelligence Systems. He is the author of over 10 scientific papers in pattern recognition and data fusion. His current interests include face recognition and detection, handwritten character recognition and data fusion.

About the Author—JING-YU YANG received the B.S. degree in Computer Science from NUST, Nanjing, China. From 1982 to 1984 he was a visiting scientist at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. From 1993 to 1994 he was a visiting professor at the Department of Computer Science, Missuria University. And in 1998, he acted as a visiting professor at Concordia University in Canada. He is currently a professor and Chairman in the Department of Computer Science at NUST. He is the author of over 100 scientific papers in computer vision, pattern recognition, and artificial intelligence. He has won more than 20 provincial awards and national awards. His current research interests are in the areas of pattern recognition, robot vision, image processing, data fusion, and artificial intelligence.