PERGAMON

Pattern Recognition 35 (2002) 2665–2668

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

What's wrong with Fisher criterion?

Jian Yang[a,b,*], Jing-yu Yang[a], David Zhang[b]

[a]*Department of Computer Science, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, People's Republic of China*
[b]*Centre for Multimedia Signal Processing, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong*

Received 28 February 2002; accepted 11 March 2002

Abstract

In this paper, a seemingly contradictory experimental result concerning Fisher criterion is first exhibited. Then, we analyze this result from the statistical correlation point of view and give a reasonable explanation. More importantly, we emphasize that Fisher criterion is not an absolute criterion, and, it should be associated with the statistical correlation together to assess the discrimination of a set of discriminant vectors. © 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Fisher criterion; Linear discriminant analysis (LDA); Feature extraction; Handwritten digit recognition

1. Introduction

It is well known that linear discriminant analysis (LDA) is a very important approach for linear feature extraction. Generally speaking, there are two typical LDA techniques. One is known as Foley–Sammon linear discriminant analysis (FSLDA) [1], and the other is the uncorrelated linear discriminant analysis (ULDA), which was recently developed by Jin et al. [3,4]. They both try to find a set of optimal discriminant vectors $\varphi_1, \ldots, \varphi_d$ by maximizing the Fisher criterion:

$$J(\varphi) = \frac{\varphi^{\mathrm{T}} S_b \varphi}{\varphi^{\mathrm{T}} S_w \varphi}, \tag{1}$$

where, S_b is the between-class scatter matrix and S_w is the within-class scatter matrix. Their main difference is the optimal discriminant vectors of FSLDA are subject to orthogonal constraint:

$$\varphi_i^{\mathrm{T}} \varphi_j = 0, \quad \forall i \neq j, \quad i, j = 1, \ldots, d. \tag{2}$$

While, those of ULDA are subject to S_t-orthogonal constraint [3]:

$$\varphi_i^{\mathrm{T}} S_t \varphi_j = 0, \quad \forall i \neq j, \quad i, j = 1, \ldots, d. \tag{3}$$

where, $S_t = S_w + S_b$ denotes the total scatter matrix.

Recently, Yang [5] proved that ULDA is the further development of the classical LDA [2], and, specially, ULDA is equivalent to the classical LDA when the positive generalized eigenvalues of S_b and S_w are unequal [4]. As to the detailed algorithms of FSLDA and ULDA, please refer to Refs. [3,5], respectively.

2. A contradictory experimental result

The experiment is performed on Concordia University CENPARMI handwritten numeral database. In this database, there are 600 samples for each of 10 digits (from 0 to 9), in which 400 ones for training and the others for testing. We do experiment based on two kinds of original features: 256-dimensional Gabor transformation feature and 121-dimensional Legendre moment feature [4,5].

The algorithm in Ref. [3] is used to work out $c - 1 = 9$ ($c=10$ is the number of classes) optimal discriminant vectors of FSLDA, and, the algorithm in Ref. [5] is exploited to find 9 optimal discriminant vectors of ULDA. At the same time, the value of Fisher criterion corresponding to each

[*] Corresponding author. Department of Computer Science, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, People's Republic of China. Tel.: +86-25-431-6840; fax: +86-25-431-5510.

E-mail addresses: tuqingh@mail.njust.edu.cn, csjyang@comp.polyu.edu.hk (Jian Yang), yangjy@mail.njust.edu.cn (Jing-yu Yang), csdzhang@comp.polyu.edu.hk (David Zhang).

0031-3203/02/$22.00 © 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(02)00071-7

Table 1
The value of Fisher criterion corresponding to each discriminant vector of ULDA and FSLDA

Fisher criterion value of the *i*th axe	1	2	3	4	5	6	7	8	9
Gabor									
FSLDA	3.90	3.81	3.69	3.57	3.38	3.18	2.94	2.76	2.60
ULDA	3.90	2.64	1.87	1.63	1.38	0.95	0.69	0.57	0.49
Legendre									
FSLDA	4.83	4.67	4.41	4.29	4.05	3.87	3.69	3.46	3.30
ULDA	4.83	2.61	2.18	1.62	1.02	0.95	0.69	0.43	0.42

Table 2
Comparison of classification error rates in FSLDA transformed space and ULDA transformed space as the number of selected projection axes varying from 1 to 9

Number of axes	1	2	3	4	5	6	7	8	9
Gabor									
FSLDA	0.660	0.576	0.549	0.569	0.561	0.538	0.528	0.501	0.457
ULDA	0.660	0.448	0.303	0.246	0.199	0.166	0.157	0.153	0.153
Legendre									
FSLDA	0.606	0.595	0.563	0.553	0.539	0.524	0.509	0.501	0.483
ULDA	0.606	0.357	0.272	0.172	0.141	0.114	0.097	0.097	0.097

discriminant vector of ULDA and FSLDA is worked out and listed in Table 1.

Then the obtained discriminant vectors of ULDA and FSLDA are, respectively, used to form feature extractor and transform 256-dimensional Gabor feature space and 121-dimensional Legendre feature space into k-dimensional (k varies from 1 to 9) feature spaces. At last, in each transformed space, a Bayes quadratic classifier is employed. Assuming that the distribution is normal and the prior probability of each class is equal, the Bayes discriminant function [5] can be defined by

$$g_l(x) = \tfrac{1}{2}\ln|\boldsymbol{\Sigma}_l| + \tfrac{1}{2}(x-\mu_l)^{\mathrm{T}}\boldsymbol{\Sigma}_l^{-1}(x-\mu_l), \tag{4}$$

where μ_l and $\boldsymbol{\Sigma}_l$ denote the mean vector and covariance matrix of class l, respectively. Based on this function, the following rule is used to make a decision: if sample x satisfies $g_k(x) = \min_l g_l(x)$, then $x \in \omega_k$. The classification error rates are shown in Table 2.

Table 2 indicates that the ULDA is more effective than FSLDA as the axe number varying from 2 to 9. To our surprise, in Table 1, the value of the Fisher criterion corresponding to each discriminant vector of FSLDA (except for the first one) is much larger than that of ULDA. A similar contradictory experimental result is shown in Tables 1 and 2 in Refs. [4]. According to the physical meaning of Fisher criterion, the larger it is, the more discriminatory the corresponding discriminant vector should be. It seems that FSLDA should do better, but the fact is not. Why is it? What's wrong with the Fisher criterion?

3. Analysis of the experimental result

To explain the above seemingly contradictory phenomenon, we first introduce some related analysis tools. After the linear discriminant transformation $Z = \Phi^{\mathrm{T}}Y$, where $\Phi = (\varphi_1, \dots, \varphi_d)$, the original feature vector Y is transformed into $Z = (Z_1, \dots, Z_d)^{\mathrm{T}}$, and the ith feature component is $Z_i = \varphi_i^{\mathrm{T}}Y$, $i = 1, \dots, d$. Then, the covariance between Z_i and Z_j is

$$\begin{aligned}\operatorname{Cov}(Z_i, Z_j) &= E(Z_i - EZ_i)(Z_j - EZ_j)\\ &= \varphi_i^{\mathrm{T}}\{E(Y-EY)(Y-EY)^{\mathrm{T}}\}\varphi_j\\ &= \varphi_i^{\mathrm{T}}S_t\varphi_j.\end{aligned}$$

Accordingly, their correlation coefficients is

$$\rho(Z_i, Z_j) = \frac{\varphi_i^{\mathrm{T}}S_t\varphi_j}{\sqrt{\varphi_i^{\mathrm{T}}S_t\varphi_i}\sqrt{\varphi_j^{\mathrm{T}}S_t\varphi_j}}. \tag{5}$$

Since the discriminant vectors of ULDA are S_t-orthogonal, for the projected features $Z_i = \varphi_i^{\mathrm{T}}Y$ ($i = 1, \dots, d$), we have $\rho(Z_i, Z_j) = 0$, $i \neq j$. That is to say, the components of ULDA transformed pattern vector are mutually uncorrelated. This property is very desirable for feature extraction.

Now, using Eq. (5), we work out the correlation coefficients between arbitrary two feature components in each FSLDA transformed space, respectively, based on the original Gabor feature and Legendre feature, and list them in Tables 3 and 4.

Table 3
Correlation between FSLDA transformed features based on original Gabor features

$\rho\ (Z_i, Z_j)$	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
Z_1	1.00	1.00	0.99	0.99	0.98	0.97	0.96	0.95	0.29
Z_2	1.00	1.00	1.00	0.99	0.99	0.98	0.96	0.95	0.32
Z_3	0.99	1.00	1.00	1.00	0.99	0.98	0.97	0.96	0.31
Z_4	0.99	0.99	1.00	1.00	0.99	0.98	0.97	0.96	0.30
Z_5	0.98	0.99	0.99	0.99	1.00	0.99	0.98	0.97	0.34
Z_6	0.97	0.98	0.98	0.98	0.99	1.00	0.99	0.97	0.26
Z_7	0.96	0.96	0.97	0.97	0.98	0.99	1.00	0.98	0.30
Z_8	0.95	0.95	0.96	0.96	0.97	0.97	0.98	1.00	0.42
Z_9	0.29	0.32	0.31	0.30	0.34	0.26	0.30	0.42	1.00

Table 4
Correlation between FSLDA transformed features based on original Legendre features

$\rho\ (Z_i, Z_j)$	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
Z_1	1.00	1.00	0.99	0.99	0.98	0.98	0.97	0.96	0.96
Z_2	1.00	1.00	0.99	0.99	0.98	0.98	0.97	0.97	0.96
Z_3	0.99	0.99	1.00	1.00	0.99	0.99	0.98	0.97	0.96
Z_4	0.99	0.99	1.00	1.00	0.99	0.99	0.98	0.98	0.97
Z_5	0.98	0.98	0.99	0.99	1.00	0.99	0.99	0.98	0.97
Z_6	0.98	0.98	0.99	0.99	0.99	1.00	0.99	0.99	0.98
Z_7	0.97	0.97	0.98	0.98	0.99	0.99	1.00	0.99	0.98
Z_8	0.96	0.97	0.97	0.98	0.98	0.99	0.99	1.00	0.99
Z_9	0.96	0.96	0.96	0.97	0.97	0.98	0.98	0.99	1.00

Tables 3 and 4 both show the high correlation between FSLDA transformed features. The high correlation leads to much information redundancy within the resulting feature vector. So, the effective discriminatory information contained in the FSLDA transformed feature vectors is insufficient despite the corresponding Fisher criterion ratios of FSLDA are much larger than those of ULDA. This is the key reason why FSLDA performs not as well as ULDA.

4. In-depth comprehend of Fisher criterion

In this paper, the experimental results and analysis make us comprehend the Fisher criterion in depth. When it is adopted to obtain only one discriminant vector, the larger the ratio is, the more powerful the projective vector's discrimination is. That is unquestioned. But, once it is used to select a set of projection vectors, some questions come into being. In the above experiment, we find that despite the Fisher criterion value corresponding to each discriminant vector of FSLDA is much larger than that of ULDA, yet the discriminatory power of FSLDA is much weaker than that of ULDA. The reason is that after the sample being projected onto the axes of FSLDA, the resulting features are highly correlated, which lead to high redundancy of information. Whereas, the really effective discriminatory information that contained in the FSLDA transformed feature vector is deficient. Conversely, after the projection of sample onto the axes of ULDA, the resulting features are uncorrelated so that their formed feature vector contains enough discriminatory information.

Accordingly, Fisher criterion is not an absolute criterion when used for deriving a set of discriminant vectors in that the correlation between the projected features is also a crucial and considerable factor. And, we cannot assess the effectiveness of a set of projection vectors merely based on their corresponding values of Fisher criterion, rather, Fisher criterion and the statistical correlation should be combined together to assess the discrimination of a set of projection vectors. More specifically, in order to obtain a set of most discriminatory discriminant vectors, Fisher criterion should be associated with the S_t-orthogonal constraint in Eq. (3), which can make sure the resulting features to be uncorrelated.

In fact, the following criterion (another version of Fisher criterion), adopted by the classical LDA [2], faces the same problem like Fisher criterion.

$$J_c(W) = \frac{|W^{\mathrm{T}} S_b W|}{|W^{\mathrm{T}} S_w W|}. \tag{6}$$

This criterion is not unconditional as well, since it cannot determine the relationship between the column vectors of projection matrix W. Strictly speaking, we cannot get an optimal solution merely by maximizing this criterion without any constraint. If the column vectors of projection matrix are required to satisfy orthogonal constraint, we can obtain one optimal projection matrix. And, if they are required to satisfy S_t-orthogonal constraint, we can obtain another optimal projection matrix, which is different from the former. Considering the factor of statistical correlation, the S_t-orthogonal constraint is more preferable. If this constraint is adopted, the resulting optimal extractor (projection matrix) is just same as that of ULDA.

Acknowledgements

The work is partially supported with Centre for Multimedia Signal Processing, The Hong Kong Polytechnic University.

References

[1] D.H. Foley, J.W. Sammon Jr., An optimal set of discriminant vectors, IEEE Trans. Comput. 24 (3) (1975) 281–289.

[2] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[3] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face recognition based on uncorrelated discriminant transformation, Pattern Recognition 34 (7) (2001) 1405–1416.

[4] Z. Jin, J.Y. Yang, Z.M. Tang, Z.S. Hu, A theorem on uncorrelated optimal discriminant vectors, Pattern Recognition 34 (10) (2001) 2041–2047.

[5] Jian Yang, J.Y. Yang, An optimal *K–L* transform method for feature extraction, SPIE Proceedings of Image Extraction, Segmentation, and Recognition, October 2001, Vol. 4550, pp. 239–244.

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. After graduation, he acts as a teacher in the Department of Applied Mathematics of Nanjing University of Science and Technology (NUST). At the same time, He is pursuing the Ph.D. degree in Pattern Recognition and Intelligence System. Now, he does research work in Hong Kong Polytechnic University. 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 200 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.

About the Author—DAVID ZHANG graduated in Computer Science from Peking University and received his M.S. and Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology (HIT) in 1983 and 1985, respectively. He received his second Ph.D. in Electrical and Computer Engineering at University of Waterloo, Ontario, Canada, in 1994. Currently, he is a professor at Hong Kong Polytechnic University. He is a founder and director of both Biometrics Technology Centres supported by UGC/CRC, Hong Kong Government, and National Nature Scientific Foundation (NSFC) of China, respectively. In addition, he is a founder and Editor-in-Chief, International Journal of Image and Graphics, and an Associate Editor, Pattern Recognition, International Journal of Pattern Recognition and Artificial Intelligence, and International Journal of Robotics and Automation. So far, he has published over 170 papers including four books in his research areas. Prof. Zhang is a senior member of the IEEE.