Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Pattern Recognition Letters 24 (2003) 147–158

Pattern Recognition Letters

www.elsevier.com/locate/patrec

A generalized Foley–Sammon transform based on generalized fisher discriminant criterion and its application to face recognition

Yue-Fei Guo [a,*], Shi-Jin Li [b], Jing-Yu Yang [c], Ting-Ting Shu [c], Li-De Wu [a]

[a] *Department of Computer Science, Information System Integration Lab., Fudan University, Shanghai 200433, China*
[b] *Department of Computer, HoHai University, Nanjing 210098, China*
[c] *Department of Computer, NUST, Nanjing 210094, China*

Received 11 September 2000; received in revised form 28 November 2001

Abstract

As the generalization of Fisher discriminant criterion, in this paper, the conception of the generalized Fisher discriminant criterion is presented. On the basis of the generalized Fisher discriminant criterion, the generalized Foley–Sammon transform (GFST) is proposed. The main difference between the GFST and the Foley–Sammon transform (FST) is that the sample set has the minimum within-class scatter and the maximum between-class scatter in the subspace spanned by all discriminant vectors constituting GFST while the sample set has these properties only on the one-dimensional subspace spanned by each discriminant vector constituting FST, that is, the transformed sample set by GFST has the best discriminant ability in global sense while FST has this property only in part sense. To calculate the GFST, an iterative algorithm is proposed, which is proven to converge to the precise solution. The speed and errors of the iterative procedure are also analyzed in detail. Lastly, our method is applied to facial image recognition, and the experimental results show that present method is superior to the existing methods in terms of correct classification rate.

Keywords: Pattern recognition; Feature extraction; Fisher discriminant criterion; Generalized Fisher discriminant criterion; Foley–Sammon transform; Generalized Foley–Sammon; Generalized optimal set of discriminant vectors

1. Introduction

It is well known that the linear feature extraction is an efficient way of reducing the dimensionality of feature vectors. Up to now, a lot of linear feature extraction methods have been proposed (Tian et al., 1988a,b), and the Foley–Sammon transform (FST) (Foley and Sammon, 1975) has been considered as one of the best methods in terms of discriminant ability. FST is based on the Fisher discriminant criterion that was first used for the linear discriminant problem (Fisher, 1936). In 1970, Sammon proposed the optimal discriminant plane (Sammon, 1970) based on the Fisher linear discriminant method. In 1975, Foley and Sammon extended Sammon's method and presented the

* Corresponding author.
E-mail address: yfguo@fudan.edu.cn (Y.-F. Guo).

0167-8655/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-8655(02)00207-6

Nomenclature

S_b between-class scatter matrix
S_w within-class scatter matrix
S_t population scatter matrix
$J_f(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi}$ Fisher discriminant function, where φ is an arbitrary n-dimensional vector
$J(\Phi) = \frac{\sum_{i=1}^{r} \varphi_i^T S_b \varphi_i}{\sum_{i=1}^{r} \varphi_i^T S_w \varphi_i}$ generalized Fisher discriminant function, where $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_r)$
$\widetilde{J}(\Phi) = \frac{\sum_{i=1}^{r} \varphi_i^T S_b \varphi_i}{\sum_{i=1}^{r} \varphi_i^T S_t \varphi_i}$ substitute of $J(\Phi)$
$E(\bullet)$ the expectation of "$\bullet$"
$\mathrm{tr}(\bullet)$ the trace of the matrix "$\bullet$"
$\mathrm{span}\{\bullet\}$ the subspace spanned by the set of vectors "$\bullet$"
$\bullet^T$ the transpose of the matrix "$\bullet$"

result on the optimal set of discriminant vectors by which FST can be constituted. Their important result attracted many researcher's attention in the field of pattern recognition (Tian et al., 1986; Hong, 1991; Okada et al., 1982; Hamamoto et al., 1989; Kittler, 1977; Hong and Yang, 1991; Cheng et al., 1992; Liu et al., 1992b; Belhumeur et al., 1997; Etemad and Chellappa, 1997). For example, FST has been applied to image classification (Tian et al., 1986) and human facial image recognition (Hong, 1991), and the solving methods of FST on various conditions have been developed, among which the method in (Liu et al., 1992b) is the most effective.

From the viewpoint of algebra, each vector of the set of the Foley–Sammon optimal discriminant vectors can be calculated step by step using the following method: first construct the orthogonal complementary space of the subspace spanned by the discriminant vectors calculated before (let the subspace be null space at the first step), then choose the vector that maximizes the Fisher criterion function as the present discriminant vector from the orthogonal complementary space. It represents that the sample set has the minimum within-class scatter and the maximum between-class scatter in the one-dimensional (1D) subspace spanned by the discriminant vector among all vectors in the complementary space. In spite of good quality of the scatter matrices in the 1D space spanned by each discriminant vector, it cannot conclude that the scatter matrices in the subspace spanned by all discriminant vectors of the Foley–Sammon optimal set have properties such as the minimum within-class scatter and the maximum between-class scatter. However, these properties are very important for designing a classifier.

Liu et al. (1992a) proposed a generalized optimal set of discriminant vectors. The main difference between their method and the methods calculating FST is that at each step for solving the discriminant vector, the criterion of selecting the vector in the complementary space as the discriminant vector is that the projected set of the training sample set in the subspace spanned by the vector and the other discriminant vectors previously calculated has the maximum ratio between the between-class distance and the within-class distance. Due to calculating the discriminant vectors step by step, the projected set on vectors of the generalized optimal set has not the best separable ability in global sense yet.

This paper presents the definition of the generalized Fisher discriminant criterion. On the basis of the generalized Fisher discriminant criterion, the generalized Foley–Sammon transform (GFST) is proposed. The main difference between the discriminant vectors constituting GFST and the discriminant vectors calculated with existing methods is that the projected set on the discriminant vectors constituting GFST has the best separable ability in global sense. It is clear that the properties of the scatter matrices of the sample set in the subspace spanned by the vectors of GFST are good in terms of separable ability. However, the calculation of the discriminant vectors constituting GFST is very difficult. To solve the vectors of GFST, an iterative algorithm is derived, which is proven to converge to the precise solution. The speed and errors of the iterative procedure are also analyzed in detail. The experimental results have shown that our method is superior to the methods in (Liu et al., 1992a,b), which were considered as the most effective in the existing methods, in terms of correct classification rate.

The remainder of the paper is organized as follows: Section 2 gives a brief review of FST and concepts of the generalized Fisher discriminant criterion and GFST. Section 3 introduces some basic theorems and discusses the iterative algorithm based on the basic theorems. Section 4 provides the classification results of the method. Also, the present method and the methods of Liu et al. (1992a,b) are compared in terms of correct classification rate. Finally, Section 5 gives a brief summary of the present method.

2. Foley–Sammon transform and the generalized Foley–Sammon transform

Let $w_1, w_2, \dots, w_m$ be m known pattern classes, $X = \{x_i\}$ $i = 1, 2, \dots, N$ be the set of n-dimensional samples. Each x_i in X belongs to a class w_j, i.e., $x_i \in w_j$, $i = 1, 2, \dots, N$, $j = 1, 2, \dots, m$. Suppose the mean vector, the covariance matrix and a priori probability of class w_i are m_i, c_i, $P(w_i)$, respectively. Then, the between-class scatter matrix S_b, the within-class scatter matrix S_w, and the population scatter matrix S_t are determined by the following formulae:

$$S_b = \sum_{i=1}^{m} P(w_i)(m_i - m_0)(m_i - m_0)^T \tag{1}$$

$$S_w = \sum_{i=1}^{m} P(w_i) E\{(x - m_i)(x - m_i)^T / w_i\} = \sum_{i=1}^{m} P(w_i) C_i \tag{2}$$

$$C_i = E\{(x - m_i)(x - m_i)^T / w_i\} \tag{3}$$

$$S_t = S_b + S_w = E\{(x - m_0)(x - m_0)^T\} \tag{4}$$

$$m_0 = E\{x\} = \sum_{i=1}^{m} P(w_i) m_i \tag{5}$$

where m_0 is the mean vector of the population distribution of samples defined by (5). The Fisher criterion can be defined as follows:

$$J_f(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \tag{6}$$

where φ is an arbitrary n-dimensional vector. Let φ_1 be the unit vector which maximize $J_f(\varphi)$, then φ_1 is the first vector of Foley–Sammon optimal set of discriminant vectors (the between-class distance in the direction of φ_1 will be maximum while the within-class distance will be minimum), the ith vector of Foley–Sammon optimal discriminant vectors will be calculated by optimizing the following problem:

$$\max_{\varphi_j^T \varphi_i = 0, \|\varphi_i\| = 1} \{J_f(\varphi_i)\} \quad j = 1, 2, \dots, i - 1 \tag{7}$$

Let $S = \{\varphi_i\}$, $i = 1, 2, \dots, r$, then the following linear transform is called FST:

$$y = \Phi^T x \tag{8}$$

where $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_r)$. Let Y be the transformed version of X by (8), then the ratio between the between-class distance and the within-class distance of Y is:

$$J(\Phi) = \frac{\mathrm{tr}(\Phi^T S_b \Phi)}{\mathrm{tr}(\Phi^T S_w \Phi)} = \frac{\sum_{i=1}^{r} \varphi_i^T S_b \varphi_i}{\sum_{i=1}^{r} \varphi_i^T S_w \varphi_i} \tag{9}$$

It is clear that the transformed set has the best separable ability in global sense when $J(\Phi)$ reaches maximum. Because the $\varphi_1, \varphi_2, \dots, \varphi_r$ constituting FST are solved step by step, FST will not guarantee the maximum of $J(\Phi)$.

Definition 1. Let

$$J(\widetilde{\Phi}) = \max_{\Phi} J(\Phi) \tag{10}$$

$$y = \widetilde{\Phi}^T x \tag{11}$$

where $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_r)$, $\widetilde{\Phi} = (\tilde{\varphi}_1, \tilde{\varphi}_2, \dots, \tilde{\varphi}_r)$, $\varphi_1, \varphi_2, \dots, \varphi_r$ and $\tilde{\varphi}_1, \tilde{\varphi}_2, \dots, \tilde{\varphi}_r$ are unit orthogonal column vectors in n-dimensional space. Then $J(\Phi)$ is called the generalized Fisher discriminant criterion, and (11) is the generalized FST (GFST).

The discriminant vectors constituting GFST can be calculated by solving the following problem:

$$\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} (J(\Phi)), \quad i,j=1,2,\ldots,r,$$
$$\Phi=(\varphi_1,\varphi_2,\ldots,\varphi_r).$$

3. Solving method

3.1. Basic theorems

An iterative algorithm of calculating GFST will be provided in this section, which will converge to the theoretical precise solution. Some conclusions will be given first before we present the detailed algorithm.

Theorem 1. *Suppose* $\forall x_i \in R^n,\ i=1,\ldots,r,\ f(x_1,\ldots,x_r)\geqslant 0,\ g(x_1,\ldots,x_r)\geqslant 0,\ f+g>0$, *and let* $h_1(x_1,\ldots,x_r)=f/g, h_2(x_1,\ldots,x_r)=(f/f+g)$, *then* h_1 *will reach its maximum at* $x_1^0,\ldots,x_r^0$ *iff* h_2 *reaches its maximum at* $x_1^0,\ldots,x_r^0$.

Proof. Because $f\geqslant 0,\ g\geqslant 0$, so $0\leqslant h_1\leqslant +\infty$, $0\leqslant h_2\leqslant 1$. And if $g>0$, then

$$h_2=\frac{f/g}{1+f/g}=\frac{h_1}{1+h_1},$$

hence h_2 will increase iff h_1 increases. If $g=0$, then $h_1=+\infty,\ h_2=1$. According to the two points above, the theorem is proven. □

Corollary 1. $J(\Phi)$ *in Definition* 1 *can be replaced by the following*:

$$\widetilde{J}(\Phi)=\frac{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{b}}\Phi)}{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{t}}\Phi)}=\frac{\sum_{i=1}^{r}\varphi_i^{\mathrm{T}}S_{\mathrm{b}}\varphi_i}{\sum_{i=1}^{r}\varphi_i^{\mathrm{T}}S_{\mathrm{t}}\varphi_i} \tag{12}$$

The proof procedure is omitted since it is the same as that of corollary in (Liu et al., 1992a).

Note: According to this corollary, we can obtain an equivalent criterion to replace the generalized Fisher discriminant criterion.

Theorem 2. *Suppose A is a real symmetric matrix of n order*, *B is a positive-definite matrix of n order*, *then*:

$$\lambda_0=\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}B\tilde{\varphi}_l}=\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}}\left(\frac{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l}{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}B\varphi_l}\right)$$
$$i,j=1,\ldots,r,\ i\neq j \tag{13}$$

iff

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(A-\lambda_0B)\tilde{\varphi}_l=\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}}\left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda_0B)\varphi_l\right)$$
$$=0\quad i,j=1,\ldots,r,\ i\neq j \tag{14}$$

where, $\tilde{\varphi}_i^{\mathrm{T}}\tilde{\varphi}_j=0,\ i\neq j,\ i,j=1,2,\ldots,r.$

Proof. To prove the necessity:

Suppose

$$\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}B\tilde{\varphi}_l}=\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}}\left(\frac{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l}{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}B\varphi_l}\right)=\lambda_0$$

then

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(A-\lambda_0B)\tilde{\varphi}_l=0. \text{ ①}$$

By the assumption of the theorem, $\forall\varphi_1,\ldots,\varphi_r\in R^n$, which are orthogonal unit vectors, we have:

$$\frac{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l}{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}B\varphi_l}\leqslant\lambda_0$$

and because B is positive-definite, so

$$\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda_0B)\varphi_l\leqslant 0. \text{ ②}$$

From ①, ② and $\varphi_1,\ldots,\varphi_r$ is arbitrary, we have:

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(A-\lambda_0B)\tilde{\varphi}_l=\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}}\left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda_0B)\varphi_l\right)$$
$$=0,\quad i,j=1,\ldots,r,\ i\neq j.$$

To prove the sufficiency:

Suppose

$$\sum_{l=1}^{r} \tilde{\varphi}_l^{\mathrm{T}}(A-\lambda_0 B)\tilde{\varphi}_l = \max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\sum_{l=1}^{r} \varphi_l^{\mathrm{T}}(A-\lambda_0 B)\varphi_l\right) = 0, \quad i,j=1,\ldots,r,\ i\neq j.$$

Because B is positive definite, so

$$\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}B\tilde{\varphi}_l} = \lambda_0. \text{ ③}$$

Let $\varphi_1,\ldots,\varphi_r$ be a group of arbitrary orthogonal unit vectors, then from the supposition we have $\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda_0 B)\varphi_l \leqslant 0$.

Hence

$$\frac{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l}{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}B\varphi_l}\lambda_0. \text{ ④}$$

So from ③ and ④ we have

$$\lambda_0 = \frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}B\tilde{\varphi}_l} = \max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\frac{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l}{\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}B\varphi_l}\right),$$

$$i,j=1,\ldots,r,\ i\neq j. \qquad \square$$

Theorem 3. *Under the assumption of Theorem 2, it holds that*:

(1) $\lambda<\lambda_0$ iff $\max\limits_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda B)\varphi_l\right) > 0.$

(2) $\lambda>\lambda_0$ iff $\max\limits_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda B)\varphi_l\right) < 0.$

Proof. we only prove (1). To prove the necessity:

Because

$$\lambda<\lambda_0=\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}B\tilde{\varphi}_l} \Rightarrow \sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(A-\lambda B)\tilde{\varphi}_l > 0,$$

so

$$\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda B)\varphi_l\right) > 0.$$

To prove the sufficiency:

Suppose

$$\sum_{l=1}^{r}\hat{\varphi}_l^{\mathrm{T}}(A-\lambda B)\hat{\varphi}_l = \max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda B)\varphi_l\right) > 0,$$

then

$$\frac{\sum_{l=1}^{r}\hat{\varphi}_l^{\mathrm{T}}A\hat{\varphi}_l}{\sum_{l=1}^{r}\hat{\varphi}_l^{\mathrm{T}}B\hat{\varphi}_l} > \lambda, \quad \text{so } \lambda_0 \geqslant \frac{\sum_{l=1}^{r}\hat{\varphi}_l^{\mathrm{T}}A\hat{\varphi}_l}{\sum_{l=1}^{r}\hat{\varphi}_l^{\mathrm{T}}B\hat{\varphi}_l} > \lambda.$$

Note: From Theorem 3, we can know the scope λ_0

Theorem 4. *Suppose A is a real symmetric matrix, then it holds that*:

$$\max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l = \lambda_1+\cdots+\lambda_r,$$

$$\min_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \sum_{l=1}^{r}\varphi_l^{\mathrm{T}}A\varphi_l = \lambda_{n-r+1}+\cdots+\lambda_n,$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ are the n eigenvalues of matrix A. And suppose $\tilde{\varphi}_1,\tilde{\varphi}_2,\ldots,\tilde{\varphi}_n$ are the orthogonal unit eigenvectors corresponding to $\lambda_1,\lambda_2,\ldots,\lambda_n$, then

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l = \lambda_1+\cdots+\lambda_r, \tag{15}$$

$$\sum_{l=n-r+1}^{n}\tilde{\varphi}_l^{\mathrm{T}}A\tilde{\varphi}_l = \lambda_{n-r+1}+\cdots+\lambda_n. \tag{16}$$

Proof. The conclusion can be made from the result of Rayleigh quotient (Wang and Shi, 1988, Th 12.1 and Th 12.3). $\square$

Theorem 5 (Sun, 1987). *Let A, E be Hermite matrices of n order respectively, $\widetilde{A}=A+E$, α_i, β_i, $\tilde{\alpha}_i$, $i=1,2,\ldots,n$, are the eigenvalues of A, E, $\widetilde{A}$ in decreasing order, then $\alpha_i+\beta_n \leqslant \tilde{\alpha}_i \leqslant \alpha_i+\beta_1$.*

Note: From Theorem 5, we can provide the result which could be used to estimate the errors of the procedure.

Theorem 6. *Let*

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)(A-\lambda B)\tilde{\varphi}(\lambda) = \max_{\substack{\varphi_i^{\mathrm{T}}\varphi_j=0\\ \|\varphi_i\|=1}} \left(\sum_{l=1}^{r}\varphi_l^{\mathrm{T}}(A-\lambda B)\varphi_l\right)$$

then

$$\lim_{\lambda\to\lambda_0}\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)A\tilde{\varphi}_l(\lambda)}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)}=\lambda_0,$$

where λ_0, A, B *are the same as those in Theorem* 2, $\tilde{\varphi}_i^{\mathrm{T}}\tilde{\varphi}_j=0,\ i\neq j,\ i,j=1,2,\ldots,r.$

Proof. Let $\lambda=\lambda_0+\varepsilon$, ①

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)(A-\lambda B)\tilde{\varphi}(\lambda)=\varepsilon_1,$$

then

$$f(\lambda)=\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)A\tilde{\varphi}_l(\lambda)}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)}=\lambda+\frac{\varepsilon_1}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)},$$

$$|f(\lambda)-\lambda_0|\leqslant|\varepsilon|+\frac{|\varepsilon_1|}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)}.\ ②$$

According to Theorem 4, we have

$$\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)\geqslant\lambda_{n-r+1}+\cdots+\lambda_n>0$$

(B is positive-definite)

where $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$ are the eigenvalues of B. Let

$$\Delta=\lambda_{n-r+1}+\cdots+\lambda_n>0,$$

then

$$|f(\lambda)-\lambda_0|\leqslant|\varepsilon|+\frac{|\varepsilon_1|}{\Delta}.\ ③$$

From ① we have $A-\lambda B=A-\lambda_0B+(-\varepsilon)B$, and let σ_i, $\tilde{\sigma}_i$, $i=1,2,\ldots,n$ are the eigenvalues of $A-\lambda_0B$, $A-\lambda B$ in decreasing order, then $\varepsilon_1=\sum_{i=1}^{r}\tilde{\sigma}_i$. Let $\varepsilon>0$, according to Theorem 5, $\sigma_i+(-\varepsilon)\lambda_1\leqslant\tilde{\sigma}_i\leqslant\sigma_i+(-\varepsilon)\lambda_n$ holds. So

$$\sum_{i=1}^{r}\sigma_i+r(-\varepsilon)\lambda_1\leqslant\varepsilon_1=\sum_{i=1}^{r}\tilde{\sigma}_i\leqslant\sum_{i=1}^{r}\sigma_i+r(-\varepsilon)\lambda_n.$$

From the definition of λ_0, we have $\sum_{i=1}^{r}\sigma_i=0$. And because $\lim_{\lambda\to\lambda_0}r(-\varepsilon)\lambda_i=\lim_{\varepsilon\to0}r(-\varepsilon)\lambda_i=0,\ i=1,2,\ldots,n$. So

$$\lim_{\lambda\to\lambda_0}\varepsilon_1=0.\ ④$$

In the case of $\varepsilon<0$, $\sum_{i=1}^{r}\sigma_i+r(-\varepsilon)\lambda_n\leqslant\varepsilon_1=\sum_{i=1}^{r}\tilde{\sigma}_i\leqslant\sum_{i=1}^{r}\sigma_i+r(-\varepsilon)\lambda_1$, ④ also holds. Based on ②, ③, ④. We have $\lim_{\lambda\to\lambda_0}f(\lambda)=\lambda_0$. i.e.,

$$\lim_{\lambda\to\lambda_0}\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)A\tilde{\varphi}_l(\lambda)}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)}=\lambda_0.\qquad\square$$

Note: This theorem guarantees that our iterative algorithm given in 3.2 converges to the theoretical solution.

As a byproduct of the proof of Theorem 6, we can get:

Corollary 2

$$\left|\frac{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)A\tilde{\varphi}_l(\lambda)}{\sum_{l=1}^{r}\tilde{\varphi}_l^{\mathrm{T}}(\lambda)B\tilde{\varphi}_l(\lambda)}-\lambda_0\right|\leqslant\left(1+\frac{r\mu}{\Delta}\right)|\lambda-\lambda_0|,$$

where $\Delta=\lambda_{n-r+1}+\cdots+\lambda_n$, $\mu=\max\{|\lambda_1|,|\lambda_n|\}$, $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$ are the eigenvalues of matrix B.

Note: From this corollary, we can estimate the errors of the iterative procedure.

3.2. The iterative algorithm

The following algorithm is designed to solve the optimal discriminant vectors constituting GFST based on the above theorems.

Case 1: S_{t} is nonsingular

In this case, $S_{\mathrm{t}}^{-1}(0)=\phi$, $\overline{S_{\mathrm{t}}^{-1}(0)}=R^n$, and S_{t} is a positive-definite matrix.

(1) It is obvious that $0\leqslant\widetilde{J}(\Phi)\leqslant1$. Let $a=0, b=1$ and $\lambda=(a+b/2)=1/2$. We can calculate $\lambda_1,\ldots,\lambda_r$, the first r eigenvalues of $S_{\mathrm{b}}-\lambda S_{\mathrm{t}}$, $\tilde{\varphi}_1(\lambda),\ldots,\tilde{\varphi}_r(\lambda)$, the orthonormal eignvectors corresponding to $\lambda_1,\ldots,\lambda_r$, and ε_1, the sum of $\lambda_1,\ldots,\lambda_r$.

If $\varepsilon_1=0$, then $\tilde{\varphi}_1(\lambda),\ldots,\tilde{\varphi}_r(\lambda)$ are the optimal discriminant vectors constituting GFST according to Theorem 2; If $\varepsilon_1<0$, then the values of a and b can be taken as a and λ, respectively, because $\lambda_0<\lambda$ holds in this case according to Theorem 3; If $\varepsilon_1>0$, then the values of a and b can be taken as λ and b, respectively, because $\lambda_0>\lambda$ holds in this case according to Theorem 3; It is obvious that $|\lambda-\lambda_0|\leqslant|a-b|/2$ holds.

(2) Repeat step (1) until $(1 + r\mu/\Delta)|\lambda - \lambda_0| \leqslant (1 + r\mu/\Delta)|a - b| < \delta$, where μ, Δ are the same as those in the corollary of Theorem 6, δ is a given small positive value.

According to the corollary of Theorem 6, we can get:

$$\left| \frac{\sum_{l=1}^{r} \tilde{\varphi}_l^{\mathrm{T}}(\lambda) A \tilde{\varphi}_l(\lambda)}{\sum_{l=1}^{r} \tilde{\varphi}_l^{\mathrm{T}}(\lambda) B \tilde{\varphi}_l(\lambda)} - \lambda_0 \right| \leqslant \delta.$$

Then $\tilde{\varphi}_1(\lambda), \ldots, \tilde{\varphi}_r(\lambda)$ are the optimal discriminant vectors constituting GFST.

It can be noticed that $|\lambda - \lambda_0| \leqslant |a - b|/2 < 1/2^l \to 0$ holds at the lth iteration. So the iterative procedure is convergent at the exponential rate.

Case 2: S_t is singular

Suppose $S_t^{-1}(0) = \mathrm{span}\{\alpha_1, \ldots, \alpha_k\}$, $\overline{S_t^{-1}(0)} = \mathrm{span}\{\beta_1, \ldots, \beta_{n-k}\}$, where $\alpha_1, \ldots, \alpha_k$; $\beta_1, \ldots, \beta_{n-k}$ are both orthogonal unit vectors. Because $\forall \alpha \in S_t^{-1}(0)$, $\alpha^{\mathrm{T}} S_b \alpha = \alpha^{\mathrm{T}} S_t \alpha = 0$, that is, the between-class distance of the projected set on α is zero, so, the vectors in $S_t^{-1}(0)$ contribute nothing to classifying, hence the optimal discriminant vectors should be selected from $\overline{S_t^{-1}(0)}$. $\forall \beta \in \overline{S_t^{-1}(0)}$, $\beta = a_1\beta_1 + a_2\beta_2 + \cdots + a_{n-k}\beta_{n-k} = P\hat{\beta}$, where $P = (\beta_1, \beta_2, \ldots, \beta_{n-k})$, $\hat{\beta} = (a_1, a_2, \ldots, a_{n-k})^{\mathrm{T}}$, and in the function of $\widetilde{J}(\Phi)$, let $\varphi_l = P\hat{\varphi}_l$, $l = 1, 2, \ldots, r$, then in the subspace of $\overline{S_t^{-1}(0)}$, we have

$$\widetilde{J}(\Phi) = \frac{\sum_{l=1}^{r} \hat{\varphi}_l^{\mathrm{T}} (P^{\mathrm{T}} S_b P) \hat{\varphi}_l}{\sum_{l=1}^{r} \hat{\varphi}_l^{\mathrm{T}} (P^{\mathrm{T}} S_t P) \hat{\varphi}_l} \equiv \widetilde{J}(\hat{\Phi}), \text{ where}$$

$$\hat{\Phi} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_r),$$

and it is obvious that $P^{\mathrm{T}} S_t P$ is a positive-definite matrix. Analogous to the *case* 1, $\tilde{\hat{\Phi}} = (\tilde{\hat{\varphi}}_1, \tilde{\hat{\varphi}}_2, \ldots, \tilde{\hat{\varphi}}_r)$ can be calculated. The following two are easy to prove:

$$\|\varphi_l\| = \|P\hat{\varphi}_l\| = 1 \quad \text{iff } \|\hat{\varphi}_l\| = 1,$$

$$\varphi_i^{\mathrm{T}} \varphi_j = 0 \quad i \neq j \quad \text{iff } \hat{\varphi}_i^{\mathrm{T}} \hat{\varphi}_j = 0 \quad i \neq j,$$

So the optimal discriminant vectors constituting GFST are $\tilde{\varphi}_l = P\tilde{\hat{\varphi}}_l$, $l = 1, 2, \ldots, r$.

4. Experimental results

4.1. Experiment 1

We generate three classes of six-dimension pattern vectors randomly. There are three vectors per class. Seven sets of data are generated. Table 1 shows $J(\Phi)$'s, the ratio between the between-class distance and the within-class distance in the subspace spanned by discriminant vectors (two discriminant vectors), obtained by Liu et al. (1992b) (FST), Belhumeur et al. (1997), Etemad and Chellappa (1997) and Liu et al. (1992a) (GODV) and our method (GFST) for seven different sets of data. We can see that the ratio $J(\Phi)$ obtained by our method is the maximum in four methods in all cases.

Table 2 shows seventh set of data. Table 3 presents the discriminant vectors calculated by four methods with seventh set of data and their corresponding ratio $J(\Phi)$'s.

4.2. Experiment 2

Face recognition, which is one of the most active areas in pattern recognition, is a very difficult problem. To demonstrate the effectiveness of the algorithm presented in this paper, a series of experiments have been conducted on the recognition

Table 1
The ratio $J(\Phi)$'s obtained by Liu et al. (1992b) (FST), Belhumeur et al. (1997), Etemad and Chellappa (1997) and Liu et al. (1992a) (GODV) and our method (GFST)

	The ratio between the between-class and within-class in discriminant subspaces						
GFST	20.20	8.43	17.51	43.47	37.95	37.70	11.48
Belhumeur et al. (1997) and Etemad and Chellappa (1997)	16.67	2.54	16.83	38.93	27.30	26.16	5.85
FST	19.76	6.49	17.31	41.45	35.38	23.43	8.61
GODV	19.95	6.49	17.38	42.61	36.03	36.03	9.53

Table 2
Seventh set of data

Three column sample vectors of the class 1		
1.5087	4.9655	3.4197
6.9790	8.9977	2.8973
3.7837	8.2163	3.4119
8.6001	6.4491	5.3408
8.5366	8.1797	7.2711
5.9356	6.6023	3.0929
Three column sample vectors of the class 2		
16.7699	13.8913	3.4591
11.3614	12.4262	19.5949
7.4083	15.8964	5.4289
14.0548	19.1369	5.0466
10.9314	10.4518	17.5148
8.8976	17.6028	14.7461
Three column sample vectors of the class 3		
4.0956	8.5323	15.4654
0.3527	14.0767	10.0185
26.8169	1.9434	12.9872
5.9741	29.6500	6.7785
8.9617	17.4838	17.3942
19.8433	12.7049	22.8110

Table 3
Panels (a)–(d) contains two column discriminant vectors calculated with the method of our paper, Belhumeur et al. (1997) and Etemad and Chellappa (1997), FST, GODV with seventh set of data and their corresponding $J(\Phi)$'s, respectively

Discriminant vectors	
Panel (a)	
0.1425	0.0272
0.5154	−0.3501
0.4761	0.2493
−0.0808	0.1808
0.1429	0.8807
−0.6786	0.0786
$J(\Phi)=\dfrac{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{b}}\Phi)}{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{w}}\Phi)}=11.48$	
Panel (b)	
0.0297	−0.2310
−0.2760	−0.7565
0.3115	−0.4616
0.0959	0.0397
0.8688	0.0724
−0.2487	0.3929
$J(\Phi)=\dfrac{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{b}}\Phi)}{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{w}}\Phi)}=5.85$	
Panel (c)	
0.0297	−0.0595
−0.2760	−0.4195
0.3115	−0.2429
0.0959	0.2120
0.8688	0.1698
−0.2487	0.8293
$J(\Phi)=\dfrac{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{b}}\Phi)}{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{w}}\Phi)}=8.61$	
Panel (d)	
0.0297	−0.1166
−0.2760	−0.5646
0.3115	−0.3533
0.0959	0.1492
0.8688	0.1375
−0.2487	0.7083
$J(\Phi)=\dfrac{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{b}}\Phi)}{\mathrm{tr}(\Phi^{\mathrm{T}}S_{\mathrm{w}}\Phi)}=9.53$	

of human facial images, and two experiments are described below. The images for the two experiments are extracted from ORL face image database, collected by the Olivetti Research Laboratory in Cambridge. 120 human face images of 12 persons, with 10 facial images per person, are used in the experiments. All the images are the size of 112 by 92, which are shown by Fig. 1. First of all, a 112-dimensional algebraic feature vector is extracted from each facial image by the method presented in (Liu et al., 1993). Therefore, there are 12 classes each having 10 feature vectors. For comparison purposes, the method of solving discriminant vectors presented in this paper and the methods in (Liu et al., 1992a,b), are tested using the above extracted algebraic feature vectors.

In the first experiment, in order to compare the time-cost calculating the discriminant vectors of these three methods, we record the time of each method in various situations under the same programming environment: PII 400 and MATLAB 5.2. Some of results are illustrated in Table 4, where FST, GODV, and GFST indicate the methods of Liu et al. (1992a,b), and our new method, respectively.

The second experiment aims to show the discriminant performances of the features by the methods in (Liu et al., 1992a,b), and our method. Therefore, the minimum distance classification criterion and the whole set of algebraic feature vectors are used for the recognition. We first fix an integer m $(1 \leqslant m \leqslant 10)$, and randomly choose m algebraic feature vectors per class as training data to calculate the scatter matrices and the mean

Fig. 1. Facial images of 12 persons.

vectors of individual classes. Then, we compute r discriminant vectors using three methods, respectively, and classify the projected data by the minimum distance classification criterion. Some of the results are also illustrated in Table 4.

According to Table 4, we have the following conclusions:

(1) The time-cost of computing the discriminant vectors by our new method is much lower than that by the methods in (Liu et al., 1992a,b). With the increase of the number of the discriminant vectors, our algorithm is more efficient.
(2) In general cases, the error classification number of the present method is lower than that

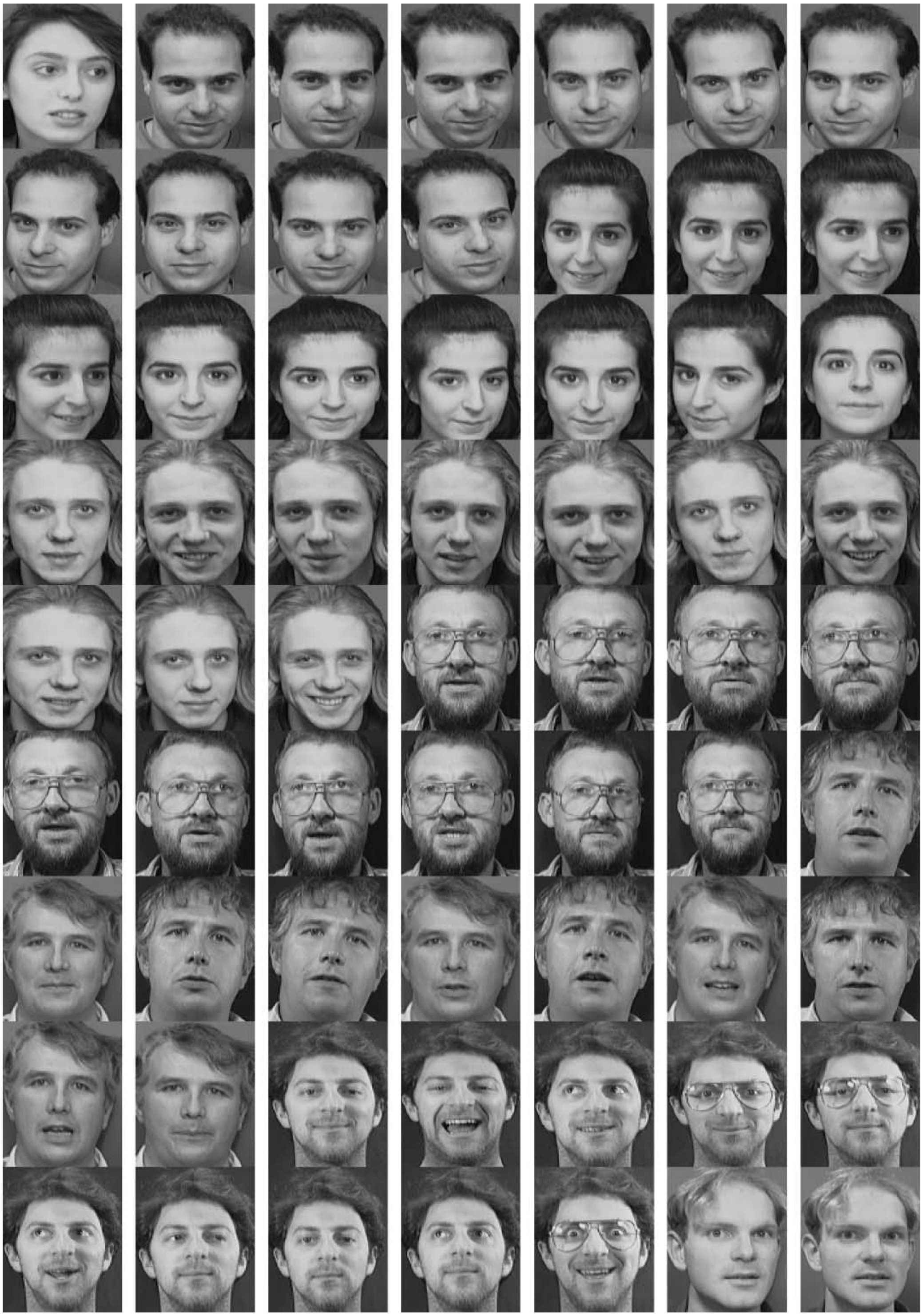

Fig. 1 (*continued*)

Table 4
Classification performances and time-cost by methods in (Liu et al., 1992a,b), and our method, p denotes the number of persons, m number of the training samples per class, r number of the discriminant vectors calculated

p	m	r	Error classification number			Time-cost (s)		
			GODV	FST	GFST	GODV	FST	GFST
5	4	3	3	0	0	18.68	37.57	14.22
5	4	4	3	0	0	21.09	46.08	14.06
6	4	4	9	1	0	21.80	44.44	13.78
6	4	5	6	0	0	24.66	52.40	13.78
7	4	5	5	0	0	25.98	50.37	13.84
7	4	6	5	0	0	29.11	58.17	13.79
8	4	6	7	2	0	30.76	56.13	13.40
8	4	7	6	1	0	34.21	63.60	13.40
9	4	7	7	2	0	37.02	60.36	13.23
9	4	8	7	1	0	40.37	67.65	13.30
10	4	8	5	3	3	42.51	65.80	13.24
10	4	9	5	3	3	46.74	72.72	13.40
11	4	9	7	3	1	49.76	70.20	13.40
11	4	10	7	2	1	54.87	76.79	13.12
12	4	10	7	8	2	58.27	73.98	13.18
12	4	11	7	5	2	62.67	80.63	13.19

of the methods in (Liu et al., 1992a,b) which were considered as the most effective methods in the existing methods.

(3) The performance of the minimum distance classifier designed based on the present method is stable as the number of discriminant vectors and sample classes varies.

4.3. Experiment 3

Table 5 provide the comparison results of classification performances based on the GFSTs calculated by applying different error precision to the case of 11 persons, 4 training samples per class and 10 discriminant vectors (see Table 4).

From the Table 5 we can see that the classification performances with our method are quite well even if the error precision is not very small.

Table 5
The comparison results of classification performances based on the GFSTs calculated by applying different error precision, δ denotes the error precision

δ	Ratio $J(\Phi)$	Error classification number
10^{-1}	0.99	2
10^{-3}	1	1
10^{-5}	1	1
10^{-7}	1	1

5. Summary

The linear feature extraction is an efficient way of reducing the dimensionality of feature vectors. This paper presents the concepts of the generalized Fisher discriminant criterion and the GFST for linear feature extraction. The main idea of the present method can be described as follows: the criterion of selecting the discriminant vectors constituting GFST is that the projected set of the training sample set on the subspace spanned by all discriminant vectors constituting GFST has the maximum ratio between the between-class distance and the within-class distance, that is, the discriminant vectors constituting GFST possess the most discriminant power in global sense. The iterative algorithm of solving the discriminant vectors constituting GFST is discussed in detail. A lot of experiments have been done and the results showed that the correct classification rate of the present method is higher than that of the methods of Liu et al. (1992a,b), which were considered as the most effective methods in existing methods, and the performance of the minimum distance classifier designed based on the present method is stable as the number of discriminant vectors and training sample classes varies. The experimental results also reveal that the time-cost of our method is much lower than that of the methods of Liu et al. (1992a,b). With the increase of the number of the discriminant vectors, our algorithm is more efficient.

Acknowledgement

This work was supported by the National Natural Science Foundation of P.R. China (no. 69935010 and no. 60072034).

References

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell. 19 (7), 711–720.

Cheng, Y.Q., Zhuang, Y.M., Yang, J.Y., 1992. Optimal Fisher discriminant analysis using the rank decomposition. Pattern Recognit. 25 (1), 101–111.

Etemad, K., Chellappa, R., 1997. Discriminant analysis for recognition of human face images. J. Opt. Soc. Am. A 14 (8), 1724–1733.

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

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, 178–188.

Wang, G.L., Shi, R.C., 1988. Theory of Matrix. Publishing House of Defense Industry, Beijing (in Chinese).

Hong, Z.Q., 1991. Algebraic feature extraction of image for recognition. Pattern Recognit. 24 (3), 211–219.

Hamamoto, Y. et al., 1989. A note on the orthogonal discriminant vector of pattern recognition. Trans. IECE (A) J72-A (2), 414–419 (in Japanese).

Hong, Z.Q., Yang, J.-Yu., 1991. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognit. 24 (4), 317–324.

Sun, Y.G., 1987. Analysis of the Perturbation of Matrix. Scientific Press, Beijing (in Chinese).

Kittler, J., 1977. On the discriminant vector method of feature selection. IEEE Trans. Comput. 26 (6), 604–606.

Liu, K., Cheng, Y.Q., Yang, J.Y., 1992a. A generalized optimal set of discriminant vectors. Pattern Recognit. 25 (1), 731–739.

Liu, K., Cheng, Y.Q., Yang, J.Y., 1992b. An efficient algorithm for Foley–Sammon optimal set of discriminant vectors by algebraic method. Int. J. Pattern Recognit. Artif. Intell. 6 (5), 817–829.

Liu, K., Cheng, Y.Q., Yang, J.Y., 1993. Algebraic feature extraction for image recognition based on an optimal discriminant criterion. Pattern Recognit. 26 (6), 903–911.

Okada, T. et al., 1982. Theory of feature extraction by orthogonal discriminant vectors. Trans. IECE (A) J65-A (8), 767–771 (in Japanese).

Sammon, J.W., 1970. An optimal discriminant plane. IEEE Trans. Comput. C 19, 826–829.

Tian, Q. et al., 1986. Image classification by the Foley–Sammon transform. Opt. Eng. 25 (7), 834–839.

Tian, Q. et al., 1988a. Comparison of statistical pattern-recognition algorithms for hybrid processing, I. Linear-mapping algorithm.. J. Opt. Soc. Am. A 5 (10), 1655–1669.

Tian, Q. et al., 1988b. Comparison of statistical pattern-recognition algorithms for hybrid processing, II: eigenvector-based algorithm. J. Opt. Soc. Am. A 5 (10), 1670–1682.