PERGAMON

Pattern Recognition 33 (2000) 1713–1726

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

A new LDA-based face recognition system which can solve the small sample size problem

Li-Fen Chen[a], Hong-Yuan Mark Liao[b,*], Ming-Tat Ko[b], Ja-Chen Lin[a], Gwo-Jong Yu[c]

[a]*Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan*
[b]*Institute of Information Science, Academia Sinica, Taiwan*
[c]*Institute of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan*

Received 16 June 1998; received in revised form 21 June 1999; accepted 21 June 1999

Abstract

A new LDA-based face recognition system is presented in this paper. Linear discriminant analysis (LDA) is one of the most popular linear projection techniques for feature extraction. The major drawback of applying LDA is that it may encounter the *small sample size problem*. In this paper, we propose a new LDA-based technique which can solve the small sample size problem. We also prove that the most expressive vectors derived in the null space of the within-class scatter matrix using principal component analysis (PCA) are equal to the optimal discriminant vectors derived in the original space using LDA. The experimental results show that the new LDA process improves the performance of a face recognition system significantly. © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Face recognition; Feature extraction; Linear discriminant analysis; Linear algebra

1. Introduction

Face recognition has been a very hot research topic in recent years [1–4]. A complete face recognition system includes two steps, i.e., face detection [5,6] and face recognition [7,8]. In this paper, attention will be focused on the face recognition part. In the last 10 years, a great number of successful face recognition systems have been developed and reported in the literature [7–13]. Among these works, the systems reported in Refs. [7,10–13] all adopted the linear discriminant analysis (LDA) approach to enhance class separability of all sample images for recognition purposes. LDA is one of the most popular linear projection techniques for feature extraction. It finds the set of the most discriminant projection vectors which can map high-dimensional samples onto a low-dimensional space. Using the set of projection vectors determined by LDA as the projection axes, all projected samples will form the maximum between-class scatter and the minimum within-class scatter simultaneously in the projective feature space. The major drawback of applying LDA is that it may encounter the so-called *small sample size problem* [14]. This problem arises whenever the number of samples is smaller than the dimensionality of the samples. Under these circumstances, the sample scatter matrix may become singular, and the execution of LDA may encounter computational difficulty.

In recent years, many researchers have noticed this problem and tried to solve it using different methods. In Ref. [11], Goudail et al. proposed a technique which calculated 25 local autocorrelation coefficients from each sample image to achieve dimensionality reduction. Similarly, Swets and Weng [12] applied the PCA approach to accomplish reduction of image dimensionality. Besides image dimensionality reduction, some researchers have

* Corresponding author. Tel.: +886-2-788-3799x1811; fax: +886-2-782-4814.
E-mail address: liao@iis.sinica.edu.tw (H.-Y.M. Liao).

PII: S0031-3203(99)00139-9

tried to overcome the computational difficulty directly using linear algebra. Instead of calculating eigenvalues and eigenvectors from an $n \times n$ matrix, Fukunaga [14] proposed a more efficient algorithm and calculated eigenvalues and eigenvectors from an $m \times m$ matrix, where n is the dimensionality of the samples and m is the rank of the within-class scatter matrix S_w. In Ref. [15], Tian et al. used a positive pseudoinverse matrix S_w^+ instead of calculating the matrix S_w^{-1}. For the same purpose, Hong and Yang [16] tried to add the singular value perturbation in S_w and made S_w a nonsingular matrix. In Ref. [17], Cheng et al. proposed another method based on the principle of rank decomposition of matrices. The above three methods are all based on the conventional Fisher's criterion function. In 1992, Liu et al. [18] modified the conventional Fisher's criterion function and conducted a number of researches [10,18,19] based on the new criterion function. They used the total scatter matrix $S_t (= S_b + S_w)$ as the divisor of the original Fisher's function instead of merely using the within-class scatter matrix. They then proposed another algorithm based on the Foley–Sammon transform [20] to select the set of the most discriminant projection vectors. It is known that the purpose of an LDA process is to maximize the between-class scatter while simultaneously minimizing the within-class scatter. When the small sample size problem occurs, the within-class scatter matrix S_w is singular. The theory of linear algebra tells us that it is possible to find some projection vectors $\mathbf{q}$'s such that $\mathbf{q}^t S_w \mathbf{q} = 0$ and $\mathbf{q}^t S_b \mathbf{q} \neq 0$. Under the above special circumstances, the modified Fisher's criterion function proposed by Liu et al. [10] will definitely reach its maximum value, i.e., 1. However, an arbitrary projection vector $\mathbf{q}$ satisfying the maximum value of the modified Fisher's criterion cannot guarantee maximum class separability unless $\mathbf{q}^t S_b \mathbf{q}$ is further maximized. Liu et al.'s [10] approach also suffers from the stability problem because the eigenvalues determined using their method may be very close to each other. This problem will result in instability of the projection vector determination process. Another drawback of Liu et al.'s approach is that their method still has to calculate an inverse matrix. Most of the time, calculation of an inverse matrix is believed to be a bottleneck which reduces efficiency.

In this paper, a more efficient, accurate, and stable method is proposed to calculate the most discriminant projection vectors based on the modified Fisher's criterion. For feature extraction, a two-stage procedure is devised. In the first stage, the homogeneous regions of a face image are grouped into the same partition based on geometric characteristics, such as the eyes, nose, and mouth. For each partition, we use the mean gray value of all the pixels within the partition to represent it. Therefore, every face image is reduced to a feature vector. In the second stage, we use the feature vectors extracted in the first stage to determine the set of the most discriminant projection axes based on a new LDA process. The proposed new LDA process starts by calculating the projection vectors in the null space of the within-class scatter matrix S_w. This null space can be spanned by those eigenvectors corresponding to the set of zero eigenvalues of S_w. If this subspace does not exist, i.e., S_w is nonsingular, then S_t is also nonsingular. Under these circumstances, we choose those eigenvectors corresponding to the set of the largest eigenvalues of the matrix $(S_b + S_w)^{-1} S_b$ as the most discriminant vector set; otherwise, the small sample size problem will occur, in which case we will choose the vector set that maximizes the between-class scatter of the transformed samples as the projection axes. Since the within-class scatter of all the samples is zero in the null space of S_w, the projection vector that can satisfy the objective of an LDA process is the one that can maximize the between-class scatter. A similar concept has been mentioned in Ref. [13]. However, they did not show any investigation results, nor did they draw any conclusions concerning the concept. We have conducted a series of experiments and compared our results with those of Liu et al.'s approach [10] and the template matching approach. The experimental results have shown that our method is superior to both Liu et al.'s approach and the template matching approach in terms of recognition accuracy. Furthermore, we have also proved that our method is better than Liu et al.'s approach in terms of training efficiency as well as stability. This indicates that the new LDA process significantly improves the performance of a face recognition system.

The organization of the rest of this paper is as follows: In Section 2, the complete two-phase feature extraction procedure will be introduced. Experimental results including those of database construction, experiments on the small sample size problem, and comparisons with two well-known approaches, will be presented in Section 3. Finally, concluding remarks will be given in Section 4.

2. Feature extraction

In this section, we shall describe in detail the proposed feature extraction technique, which includes two phases: pixel grouping and generalized LDA based on the modified Fisher's function.

2.1. Pixel grouping

According to the conclusion drawn in Ref. [21], a statistics-based face recognition system should base its recognition solely on the "pure" face portion. In order to fulfill this requirement, we have built a face-only database using a previously developed morphology-based filter [6]. Using this morphological filter, the eye-analogue segments are grouped into pairs and used to locate potential face regions. Thus, every constituent of

the face-only database is the face portion containing only the eyes, nose and mouth. Some examples of this face database are shown in Fig. 1. In order to execute pixel grouping, the above-mentioned face-only images are transformed into normalized sizes. Let the training database be comprised of N normalized face-only images of size $P \times P$. We pile up these N images and align them into the same orientation, as shown in Fig. 2. Therefore, we obtain P^2 N-dimensional vectors whose elements are the gray values of the pixels. These P^2 N-dimensional vectors are then clustered into m groups using the k-means clustering method, where m is the resolution of the transformed images. After clustering, each image is partitioned into m groups, and each pixel is assigned to one of the groups. For each image, we calculate the average gray value of each group and use these m mean values to represent the whole image. Thus, the P^2-dimensional images are now reduced to m-dimensional with $m \ll P^2$. Fig. 3 shows some examples of the transformed images. The images in the leftmost column are the original images of size 60×60, and the others are the transformed images with increasing resolutions of 2^5, 2^6, 2^7, and 2^8, respectively, from left to right. After pixel grouping, we use the transformed images to execute the second phase – generalized LDA.

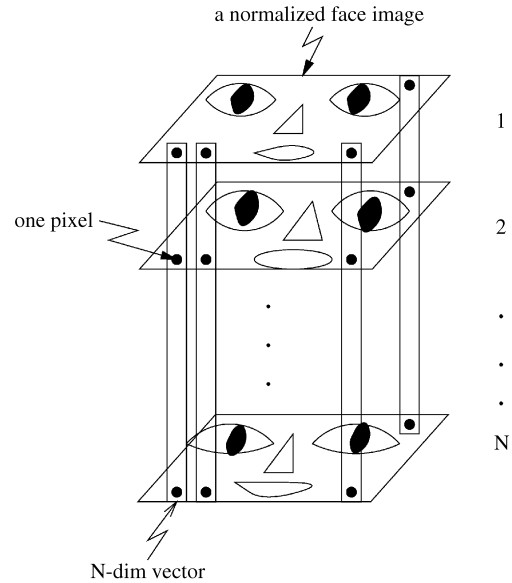

Fig. 2. Illustration of the pixel grouping process. N normalized face images are piled up and aligned into the same orientation. Suppose the image size is $P \times P$; then P^2 N-dimensional vectors are obtained, and the elements of a vector are the gray values of the pixels in N different images.

2.2. Generalized LDA

The purpose of pixel grouping is to reduce the dimensionality of the samples and to extract geometric features; however, it does not take class separability into consideration at all. In the literature [10–12], LDA is a well-known technique for dealing with the class separability problem. LDA can be used to determine the set of the most discriminant projection axes. After projecting all the samples onto these axes, the projected samples will form the maximum between-class scatter and the minimum within-class scatter in the projective feature space. In what follows, we shall first introduce the LDA approach and some related works. In the second subsection, we shall describe our approach in detail.

Fig. 1. Examples of normalized face-only images. The top two rows of images are of the same person, and the bottom two rows are of another person.

2.2.1. Conventional LDA and its potential problem

Let the training set comprise K classes, where each class contains M samples. In LDA, one has to determine the mapping

$$\tilde{\mathbf{x}}_m^k = A^t \mathbf{x}_m^k, \tag{1}$$

where $\mathbf{x}_m^k$ denotes the n-dimensional feature vector extracted from the mth sample of the kth class, and $\tilde{\mathbf{x}}_m^k$ denotes the d-dimensional projective feature vector of $\mathbf{x}_m^k$ transformed by the $n \times d$ transformation matrix A. One way to find the mapping A is to use Fisher's criterion [22]:

$$F(\mathbf{q}) = \frac{\mathbf{q}^t S_b \mathbf{q}}{\mathbf{q}^t S_w \mathbf{q}}, \tag{2}$$

where $\mathbf{q} \in \mathscr{R}^n$, $S_b = \sum_{k=1}^{K} (\bar{\mathbf{x}}^k - \bar{\mathbf{x}})(\bar{\mathbf{x}}^k - \bar{\mathbf{x}})^t$, and $S_w = \sum_{k=1}^{K} \sum_{m=1}^{M} (\mathbf{x}_m^k - \bar{\mathbf{x}}^k)(\mathbf{x}_m^k - \bar{\mathbf{x}}^k)^t$ are the between-class scatter matrix and within-class scatter matrix, respectively, where $\bar{\mathbf{x}}^k = 1/M \sum_{m=1}^{M} \mathbf{x}_m^k$ and $\bar{\mathbf{x}} = 1/KM \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbf{x}_m^k$.

Fig. 3. Results obtained after performing pixel grouping. The images in the leftmost column are the original images, and the others are the transformed images with the increasing resolutions of 2^5, 2^6, 2^7, and 2^8, from left to right.

The column vectors of A can be chosen from the set of $\tilde{\mathbf{q}}$'s, where

$$\tilde{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathscr{R}^n} F(\mathbf{q}). \tag{3}$$

After projecting all the $\mathbf{x}_m^k$'s (where $k = 1, \ldots, K$; $m = 1, \ldots, M$) onto the $\tilde{\mathbf{q}}$ axis, the projected samples, $\tilde{\mathbf{x}}_m^k$'s ($k = 1, \ldots, K$; $m = 1, \ldots, M$), will form the maximum between-class scatter and the minimum within-class scatter. The vector $\tilde{\mathbf{q}}$ is called the optimal discriminant projection vector. According to linear algebra, all $\tilde{\mathbf{q}}$'s can be eigenvectors corresponding to the set of largest eigenvalues of $S_w^{-1}S_b$. The major drawback of applying the LDA approach is that it may encounter the *small sample size problem* [14]. The small sample size problem occurs whenever the number of samples is smaller than the dimensionality of the samples. Whenever this happens, the matrix S_w becomes singular, and the computation of S_w^{-1} becomes complex and difficult. Liu et al. seriously addressed the problem in [10,18,19]. One of their efforts was to propose a modified Fisher's criterion function, $\hat{F}(\mathbf{q})$, to replace the original Fisher's function, $F(\mathbf{q})$. They have proved [19] that $\hat{F}(\mathbf{q})$ is exactly equivalent to $F(\mathbf{q})$. That is,

$$\arg\max_{q\in\mathscr{R}^n} \hat{F}(\mathbf{q}) = \arg\max_{\mathbf{q}\in\mathscr{R}^n} F(\mathbf{q}). \tag{4}$$

In what follows, we shall directly describe two theorems of Ref. [19] which are related to our work.

Theorem 1. *Suppose that R is a set in the n-dimensional space, $\forall \mathbf{x} \in R, f(\mathbf{x}) \geqslant 0, g(\mathbf{x}) \geqslant 0$, and $f(\mathbf{x}) + g(\mathbf{x}) > 0$. Let $h_1(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$, and $h_2(\mathbf{x}) = f(\mathbf{x})/(f(\mathbf{x}) + g(\mathbf{x}))$. Then, $h_1(\mathbf{x})$ has the maximum (including positive infinity) at point $\mathbf{x}_0$ in R iff $h_2(\mathbf{x})$ has the maximum at point x_0* [19].

Theorem 2. *The Fisher's criterion function $F(\mathbf{q})$ can be replaced by*

$$\hat{F}(\mathbf{q}) = \frac{\mathbf{q}^t S_b \mathbf{q}}{\mathbf{q}^t S_w \mathbf{q} + \mathbf{q}^t S_b \mathbf{q}} \tag{5}$$

in the course of solving the discriminant vectors of the optimal set [19].

From the above two theorems, we know that $F(\mathbf{q})$ and $\hat{F}(\mathbf{q})$ are functionally equivalent in terms of solving the optimal set of projection axes (or discriminant vectors). Therefore, one can choose either $F(\mathbf{q})$ or $\hat{F}(\mathbf{q})$ to derive the optimal projection axes. In this paper, we propose a new method to calculate the optimal projection axes based on $\hat{F}(\mathbf{q})$. According to the normal process of LDA, the solutions of $\max_{\mathbf{q}\in\mathscr{R}^n} \hat{F}(\mathbf{q})$ should be the eigenvectors corresponding to the set of the largest eigenvalues of the matrix $(S_b + S_w)^{-1}S_b$. If the small sample size problem occurs at this point, the eigenvectors of $(S_b + S_w)^{-1}S_b$ will be very difficult to compute due to the singularity problem. In order to avoid direct computation of $(S_b + S_w)^{-1}S_b$, Liu et al. [19] suggested deriving the discriminant vectors in the complementary subspace of the null space of $S_t (S_t = S_b + S_w$, which denotes a total scatter matrix), where the null space of S_t is spanned by the eigenvectors corresponding to the zero eigenvalues of S_t. Since the total scatter matrix S_t' in the complementary subspace is nonsingular, it is feasible to follow the normal LDA process to derive the discriminant vectors in this subspace. However, there are still some critical problems associated with this approach. The first problem with Liu et al.'s approach is the validity of the discriminant vector set problem. It is known that the purpose of LDA is to maximize the between-class scatter while minimizing the

within-class scatter simultaneously. In the special case where $\mathbf{q}^t S_w \mathbf{q} = 0$ and $\mathbf{q}^t S_b \mathbf{q} \neq 0$, Eq. (5) will definitely reach the maximum value of $\hat{F}(\mathbf{q})$. However, an arbitrary projection vector $\mathbf{q}$ satisfying the above conditions cannot guarantee derivation of the maximum $\mathbf{q}^t S_b \mathbf{q}$ value. Under these circumstances, a correct LDA process cannot be completed because only the within-class scatter is minimized while the between-class scatter is not surely maximized. The second problem associated with Liu et al.'s approach is the stability problem. In Ref. [23], the author stated that an eigenvector will be very sensitive to small perturbation if its corresponding eigenvalue is close to another eigenvalue of the same matrix. Unfortunately, in Ref. [18], the matrix used to derive the optimal projection vector suffers from the above-mentioned problem. In other words, their optimal projection vector determination process may be severely influenced whenever a small perturbation is added. The third problem associated with Liu et al.'s approach [18] is the singularity problem. This is because their approach still has to calculate the inverse of the matrix S_t'. In this paper, we propose a more efficient, accurate, and stable method to derive the most discriminant vectors from LDA based on the modified Fisher's criterion. In the proposed approach, we calculate the projection vectors in the null space of the within-class scatter matrix S_w because the projection vectors found in this subspace can make all the projected samples form zero within-class scatter. Furthermore, we will also prove that finding the optimal projection vector in the original sample space is equivalent to calculating the most expressive vector [12] (via principal component analysis) in the above-mentioned subspace. In what follows, we shall describe the proposed method in detail.

2.2.2. The proposed method

Let the database comprise K classes, where each class contains M distinct samples, and let $\mathbf{x}_m^k$ be an n-dimensional column vector which denotes the feature vector extracted from the mth sample of the kth class. Suppose S_w and S_b are, respectively, the within-class scatter matrix and the between-class scatter matrix of x_m^k's (where $k = 1, \ldots, K$; $m = 1, \ldots, M$), and suppose the total scatter matrix $S_t = S_w + S_b$. According to linear algebra [24] and the definitions of the matrices S_t, S_w, and S_b, $rank(S_t) \leqslant rank(S_b) + rank(S_w)$, where $rank(S_t) = \min(n, KM - 1)$, $rank(S_b) = \min(n, K - 1)$, and $rank(S_w) = \min(n, K \times (M - 1))$. In this paper, we shall determine a set of discriminant projection vectors from the null subspace of S_w. Therefore, the rank of S_w certainly is the major focus of this research. Suppose the rank of S_w is r, i.e., $r = \min(n, K \times (M - 1))$. If $r = n$, this implies that $K \times (M - 1) \geqslant n \Rightarrow KM \geqslant n + K \Rightarrow KM - 1 \geqslant n + K - 1 \geqslant n$. The above inequality means that the rank of S_t is equal to n. Consequently, if S_w is nonsingular, then S_t is nonsingular, too. Under these circumstances, there will be no singularity problem when the matrix $S_t^{-1} S_b$ is computed in the normal LDA process. On the other hand, if r is smaller than n, the small sample size problem will occur. For this case, we propose a new method to derive the optimal projection vectors.

Fig. 4 illustrates graphically the process of deriving the optimal projection vectors when $r < n$. In the top part of Fig. 4, V stands for the original sample space, and T represents a linear transformation: $T(\mathbf{x}) = S_w \mathbf{x}$, $x \in V$. Since the rank of S_w is smaller than the dimensionality of $V (r < n)$, there must exist a subspace $V_0 \subset V$ such that $V_0 = span\{\alpha_i \mid S_w \alpha_i = 0, \text{for } i = 1, \ldots, n - r\}$. V_0 here is called the null space of S_w. In the bottom part of Fig. 4, the flow chart of the discriminant vector determination process is illustrated. Let $Q = [\alpha_1, \ldots, \alpha_{n-r}]$. First, all samples X's are transformed from V into its subspace V_0 through the transformation QQ^t. Then, the eigenvectors corresponding to the largest eigenvalues of the between-class scatter matrix $\tilde{S}_b$ (a new matrix formed by the transformed samples) in the subspace V_0 are selected as the most discriminant vectors. In what follows, we shall describe our approach in detail.

First of all, Lemma 1 shows the subspace where we can derive the discriminant vectors based on maximizing the modified Fisher's criterion.

Lemma 1. *Suppose* $V_0 = span\{\alpha_i | S_w \alpha_i = 0,\ \alpha_i \in \mathscr{R}^n,\ i = 1, \ldots, n - r\}$, *where n is the dimensionality of samples,* S_w *is the within-class scatter matrix of the samples, and r is the rank of* S_w. *Let* S_b *denote the between-class scatter matrix of the samples. For each* $\tilde{\mathbf{q}} \in v_0$ *which satisfies* $\tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}} \neq 0$, *it will maximize the function* $\hat{F}(\mathbf{q}) = \mathbf{q}^t S_b \mathbf{q} / (\mathbf{q}^t S_b \mathbf{q} + \mathbf{q}^t S_w \mathbf{q})$.

Proof. 1. Since both S_b and S_w are real symmetric, $\mathbf{q}^t S_b \mathbf{q} \geqslant 0$ and $\mathbf{q}^t S_w \mathbf{q} \geqslant 0$, for all $\mathbf{q} \in \mathscr{R}^n$, it follows that

$$0 \leqslant \mathbf{q}^t S_b \mathbf{q} \leqslant \mathbf{q}^t S_b \mathbf{q} + \mathbf{q}^t S_w \mathbf{q} \Rightarrow 0 \leqslant \hat{F}(\mathbf{q}) = \frac{\mathbf{q}^t S_b \mathbf{q}}{\mathbf{q}^t S_b \mathbf{q} + \mathbf{q}^t S_w \mathbf{q}} \leqslant 1.$$

It is obvious that $\hat{F}(\mathbf{q}) = 1$ if and only if $\mathbf{q}^t S_b \mathbf{q} \neq 0$ and $\mathbf{q}^t S_w \mathbf{q} = 0$.

2. For each $\hat{q} \in V_0$, $\tilde{\mathbf{q}}$ can be represented as a linear combination of the set $\{\alpha_i\}$, i.e., $\tilde{\mathbf{q}} = \sum_{i=1}^{n-r} a_i \alpha_i$, where a_i is the projection coefficient of $\tilde{\mathbf{q}}$ with respect to α_i. Therefore, we have

$$S_w \tilde{\mathbf{q}} = S_w \sum_{i=1}^{n-r} a_i \alpha_i = \sum_{i=1}^{n-r} a_i S_w \alpha_i = 0 \Rightarrow \tilde{\mathbf{q}}^t S_w \tilde{\mathbf{q}} = 0.$$

From 1. and 2., we can conclude that for each $\tilde{\mathbf{q}} \in V_0$ which satisfies $\tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}} \neq 0$, the function $\hat{F}(\mathbf{q})$ will be maximized.

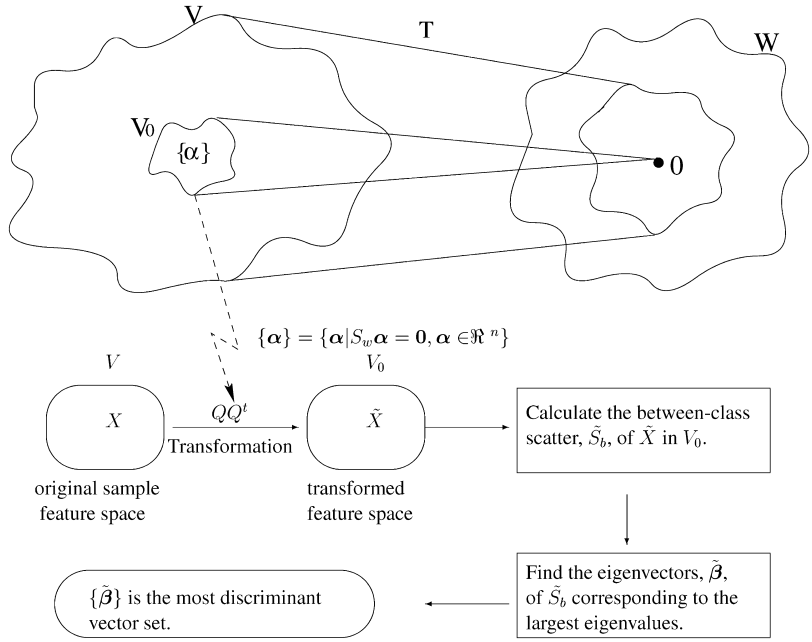

Fig. 4. Illustration of the projection vector set determination process. At the top of the figure, T is a linear transformation from V to W: $T(\boldsymbol{x}) = S_w\boldsymbol{x}, \boldsymbol{x} \in V$. V_0 is the null space of S_w. In the middle of the figure, X stands for the original sample set, and $\tilde{X}$ is the transformed sample feature set of X obtained through the transformation QQ^t, where $Q = [\alpha_1, \ldots, \alpha_{n-r}]$, n is the dimensionality of the samples, r is the rank of S_w, and $S_w\alpha_i = 0$ for each α_i. The most discriminant vectors for LDA can be computed from the between-class scatter matrix, $\tilde{S}_b$, of $\tilde{X}$.

Lemma 1 has a critical issue related to LDA. That is, when the small sample size problem occurs, an arbitrary vector $\tilde{\mathbf{q}} \in V_0$ that maximizes $\hat{F}(\mathbf{q})$ is not necessarily the optimal discriminant vector of LDA. This is because under the above situation, $\tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}}$ is not guaranteed to reach the maximal value. Therefore, one can conclude that it is not sufficient to derive the discriminant vector set simply based on the modified Fisher's criterion when the small sample size problem occurs. In what follows, Lemma 2 will show that the within-class scatter matrix of all the transformed samples in V_0 is a complete zero matrix. Lemma 2 is very important because once it is proved correct, determination of the discriminant vector set no longer depends on the total scatter matrix. Instead, the discriminant vector set can be derived directly from the between-class scatter matrix.

Lemma 2. *Let QQ^t be a transformation which transforms the samples in V into a subspace V_0, where $Q = [\alpha_1, \ldots, \alpha_{n-r}]$ is an $n \times (n-r)$ matrix and each α_i satisfies $S_w\alpha_i = 0$, for $i = 1, \ldots, n-r$; and where the subspace V_0 is spanned by the orthonormal set of α_i's. If all the samples are transformed into the subspace V_0 through QQ^t, then the within-class scatter matrix $\tilde{S}_w$ of the transformed samples in V_0 is a complete zero matrix.*

Proof. suppose $\mathbf{x}_m^k$ is the feature vector extracted from the mth sample of the kth class, and that the database comprised K classes, where each class contains M samples. Let $\mathbf{y}_m^k$ denote the transformed feature vector of $\mathbf{x}_m^k$ through the transformation QQ^t. That is, $\mathbf{y}_m^k = QQ^t\mathbf{x}_m^k$, $\bar{\mathbf{y}}^k = QQ^t\bar{\mathbf{x}}^k$, and $\bar{\mathbf{y}} = QQ^t\bar{\mathbf{x}}$, where $\bar{\mathbf{x}}^k = 1/M\sum_{m=1}^{M}\mathbf{x}_m^k$ and $\bar{\mathbf{x}} = 1/KM\sum_{k=1}^{K}\sum_{m=1}^{M}\mathbf{x}_m^k$ Thus,

$$\begin{aligned}\tilde{S}_w &= \sum_{k=1}^{K}\sum_{m=1}^{M}(\mathbf{y}_m^k - \bar{\mathbf{y}}^k)(\mathbf{y}_m^k - \bar{\mathbf{y}}^k)^t \\ &= \sum_{k=1}^{K}\sum_{m=1}^{M}(QQ^t\mathbf{x}_m^k - QQ^t\bar{\mathbf{x}}^k)(QQ^t\mathbf{x}_m^k - QQ^t\bar{\mathbf{x}}^k)^t \\ &= QQ^t\sum_{k=1}^{K}\sum_{m=1}^{M}(\mathbf{x}_m^k - \bar{\mathbf{x}}^k)(\mathbf{x}_m^k - \bar{\mathbf{x}}^k)QQ^t \\ &= QQ^tS_wQQ^t = \mathbf{0}, \text{ since } S_wQ = \mathbf{0}. \end{aligned} \tag{6}$$

We have mentioned earlier that the LDA process is used to determine the set of the most discriminant projection axes for all the samples. After projection, all the projected samples form the minimum within-class scatter and the maximum between-class scatter. Lemma 1 tells us that for any $\tilde{\mathbf{q}} \in V_0$, as long as it satisfies $\tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}} \neq 0$, the modified Fisher's criterion, $\hat{F}(\mathbf{q})$, will be maximized to 1.

However, Lemma 1 also tells us that we should add another criterion to perform LDA, not just depend on the Fisher's criterion. Lemma 2, on the other hand, tells us that the selection of $\tilde{\mathbf{q}} \in V_0$ enforces $\tilde{S}_w = \mathbf{0}$. That is to say: $\tilde{S}_t = \tilde{S}_w + \tilde{S}_b = \tilde{S}_b$. Since $\tilde{S}_w$ is consistently equal to $\mathbf{0}$, we have to select a set of projection axes that can maximize the between-class scatter in V_0. From the above two lemmas, we know that maximizing the between-class scatter in V_0 is equal to maximizing the total scatter in V_0. Under these circumstances, we can apply the principal component analysis (PCA) method [25] to derive the set of the most discriminant projection vectors and fulfill the requirement of LDA. The physical meaning of PCA is to find a set of the most expressive projection vectors such that the projected samples retain the most information about the original samples. The most expressive vectors derived from a PCA process are the l eigenvectors corresponding to the l largest eigenvalues of $\tilde{S}_t$, where $(\sum_{i=1}^{l}\lambda_i / \sum_{i=1}^{n}\lambda_i) \geqslant p$, n is the dimensionality of samples, and λ_i represents the eigenvalue ordered in the ith place in $\tilde{S}_t$. Basically, λ_i is in the decreasing order from 1 to n. If $p = 0.95$, a good enough representation is obtained [26]. In what follows, we shall show the proposed method in Theorem 3 based on the above two lemmas.

Theorem 3. *Suppose that $Q = [\alpha_1, \ldots, \alpha_{n-r}]$, and that α_i's are the eigenvectors corresponding to the zero eigenvalues of the within-class scatter matrix S_w in the original feature space V, where n is the dimensionality of the feature vectors and r is the rank of S_w. Let V_0 denote the subspace spanned by the set of eigenvectors $\alpha_1, \ldots, \alpha_{n-r}$. If r is smaller than n, the most expressive vector $\tilde{\mathbf{q}}$ in V_0 obtained through the transformation QQ^t will be the most discriminant vector in V.*

Proof. 1. From Lemma 2, we know that the within-class scatter matrix $\tilde{S}_w$ in V_0 is a complete zero matrix. Thus, the between-class scatter matrix $\tilde{S}_b$ in V_0 is equal to the total scatter matrix $\tilde{S}_t$ in V_0.

2. The most expressive projection vector $\tilde{\mathbf{q}}$ in V_0 satisfies $\tilde{\mathbf{q}}^t\tilde{S}_b\tilde{\mathbf{q}} > 0$. Suppose $S_b = \tilde{S}_b + \hat{S}_b$, where S_b, $\tilde{S}_b$, and $\hat{S}_b$ are all real symmetric. Then,

$$\tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}} = \tilde{\mathbf{q}}^t\tilde{S}_b\tilde{\mathbf{q}} + \tilde{\mathbf{q}}^t\hat{S}_b\tilde{\mathbf{q}} \geqslant \tilde{\mathbf{q}}^t\tilde{S}_b\tilde{\mathbf{q}} > 0 \Rightarrow \tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}} \neq 0.$$

3. We can show that $\tilde{\mathbf{q}}$ is the optimal solution within V_0 that can maximize $\hat{F}(\mathbf{q})$. Since the most expressive projection vector $\tilde{\mathbf{q}}$ in V_0 can maximize the value of $\tilde{\mathbf{q}}^t S_b \tilde{\mathbf{q}}$, and $\tilde{\mathbf{q}}^t S_w \tilde{\mathbf{q}} = 0$ is known, we can conclude that the most expressive projection vector in V_0 is the most discriminant projection vector in V for LDA.

After projecting all the samples onto the projective feature space based on Theorem 3, a Euclidean distance classifier is used to perform classification in the projective feature space.

The proposed algorithm

Input: N n-dimensional vectors.

Output: The optimal discriminant vector set of all N input vectors.

Algorithm:

Step 1. Calculate the within-class scatter S_w and the between-class scatter S_b.

Step 2. Suppose the rank of S_w is r. If $r = n$, then the discriminant set is the eigenvectors corresponding to the set of the largest eigenvalues of matrix $(S_b + S_w)^{-1}S_b$; otherwise, go on to the next step.

Step 3. Perform the singular value decomposition of S_w as $S_w = U\sum V^t$, where U = V because S_w is symmetric.

Step 4. Let $V = [v_1, \ldots, v_r, v_{r+1}, \ldots, v_n]$ and $Q = [v_{r+1}, \ldots, v_n]$. (It has been shown in Ref. [24] that the null space of S_w can be spanned by $v_{r+1}, \ldots, v_n$).

Step 5. Compute $\tilde{S}_b$, where $\tilde{S}_b = QQ^tS_b(QQ^t)^t$.

Step 6. Calculate the eigenvectors corresponding to the set of the largest eigenvalues of $\tilde{S}_b$ and use them to form the most discriminant vector set for LDA.

3. Experimental results

3.1. Database construction and feature extraction

The facial image database contained 128 persons (classes), in which for each person, 10 different face images with frontal views were obtained. The process for collecting facial images was as follows: after asking the persons to sit down in front of a CCD camera, with neutral expression and slightly head moving in frontal views, a 30-s period was recorded on videotape under well-controlled lighting condition. Later, a frame grabber was used to grab 10 image frames from the videotape and stored them with resolution of 155×175 pixels. According to the conclusion drawn in Ref. [21], which stated that a statistics-based face recognition system should base its recognition solely on the "pure" face portion, a face-only database was built using a previously developed morphology-based filter [6]. Part of the database is shown in Fig. 1. For pixel grouping, each database image was transformed into a normalized size, 60×60. Then, all the 1280 database images (128×10) were piled up and aligned into the same orientation. After this process, 3600 1280-dimensional vectors were obtained. These vectors were then clustered into m groups (where m stands for the required resolution) using the K-means clustering method. For each image, the average gray value of each group was calculated, and then these m mean values were used to represent the whole image. Therefore, the dimensionality of each image was reduced from 3600 to m dimensions. Since m is a variable which stands for the dimensionality of the feature vectors of experimentation, we designed an

experiment to decide the best value of m for subsequent experiments. For this experiment, we chose a training database containing 128 persons, with six frontal view samples for each person. For testing purposes, we used a 128-person testing database. Within the database, we obtained 10 samples for each person. Since the database used was a large database, the projection vectors for LDA could be directly computed from $S_t^{-1}S_b$. Table 1 showed a set of experimental results obtained by applying different m values ($m = 32$, 64, 128, and 256). The data shown in the second column of Table 1 are the number of projection axes used at a certain resolution. The number of projection axes adopted was decided by checking the p value mentioned in Section 2.2.2. If p reached 0.95, then we used its corresponding number of projection axes as the maximum number of projection axes. Therefore, for $m = 32$ and 64, the corresponding number of projection axes adopted as 29 and 53, respectively. From Table 1, we find that $m = 128$ was the most suitable number of features in terms of recognition rate and training efficiency. Therefore, in the subsequent sets of experiments, this number ($m = 128$) was globally used.

3.2. Experiments on the small sample size problem

In order to evaluate how our method interacts with the small sample size problem, including problems like the number of samples in each class and the total number of classes used, we conducted a set of experiments and show the results in Fig. 5. The horizontal axis in Fig. 5 represents the number of classes used for recognition, and the vertical axis represents the corresponding recognition rate. The '+', '×', and '○' signs in Fig. 5 indicate there were 2, 3, and 6 samples in each class, respectively, for experimentation. The results shown in Fig. 5 reflect that the proposed approach performed fairly well when the size of the database was small. However, when K (the number of classes) multiplied the $M - 1$ (the number of samples minus 1) was close to $n(n = 128)$, the performance dropped significantly. This phenomenon was especially true for the case where $M = 6$. In Section 2.2.2, we have mentioned that the information for deriving the most discriminant vectors depended on the null space of S_w, V_0. The dimension of V_0, $dim(V_0)$, was equal to $n - (KM - K)$, where n is equal to 128, K is the number of classes and M is the number of samples in each class. When $M = 6$ and K approached 25, $K(M - 1)$ was very close to $n(128)$. Under these circumstances, the recognition rate dropped significantly (see Fig. 5). The reason for this phenomenon emerged was the low value of $dim(V_0)$. When the $dim(V_0)$ value was small, not many spaces were available for deriving the discriminant projection axes; hence, the recognition rate dropped. Inspecting another curve (the '+' sign) in Fig. 5, it is seen that since there were only two samples in each class, the corresponding curve of the recognition rate is not as monotonous as those for the cases that contained 3 and 6 samples in a class. This part of the experiment provided a good guide for making better decisions regarding the number of samples in each class and the number of classes in a database. When one wants to solve the small sample size problem with good performance, the above experimental results can be used as a good reference.

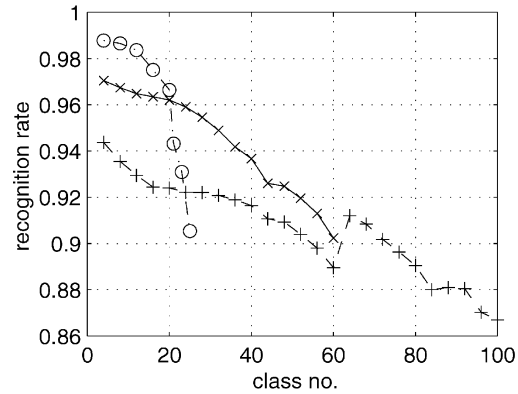

Fig. 5. Experimental results obtained using our method under the small sample size problem. The '+' sign means that each class in the database contains 2 samples. The '×' sign and '○' sign mean each class in the database contains three and six samples, respectively.

Table 1
Face recognition results obtained by applying different numbers of features extracted from images. The training database contains 128 persons, where each person contains six distinct samples

Number of features	Number of projection axes used	Recognition rate (%)	Training time (S)
$m = 32$	29	95.28	0.3039
$m = 64$	53	96.27	1.1253
$m = 128$	70	97.54	3.5746
$m = 256$	98	97.34	17.1670

3.3. Comparison with other approaches

In order to demonstrate the effectiveness of our approach, we conducted a series of experiments and compared our results with those obtained using two other well-known approaches. Fig. 6 shows the experimental results obtained using our approach, Liu et al.'s approach [10] and the template matching approach. The horizontal axes and vertical axes in Fig. 6 represent the number of classes in the database and the corresponding

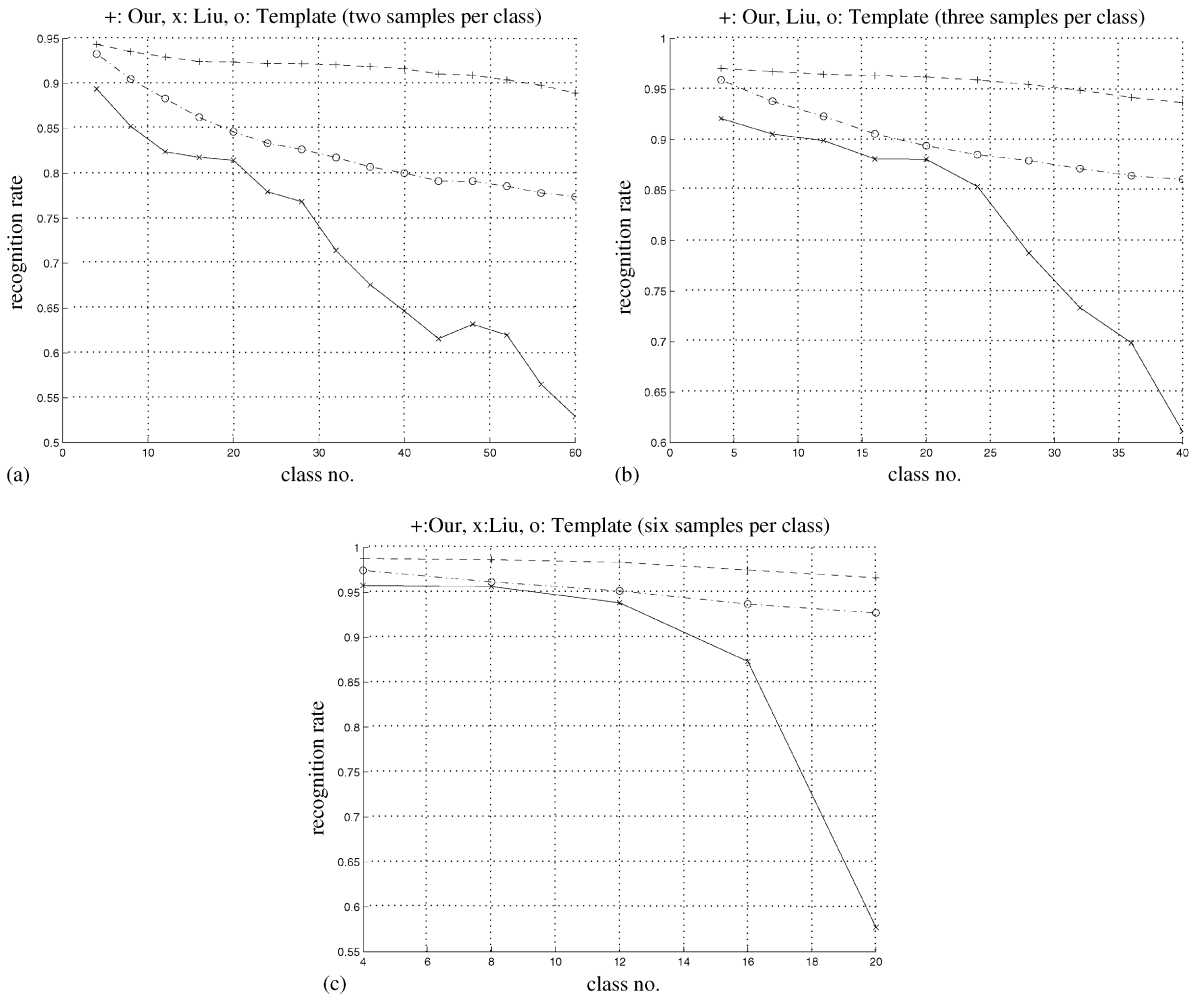

Fig. 6. The experimental results obtained using our method ('+' sign), Liu's method ('×' sign), and template matching ('○' sign). The horizontal axis represents the number of classes in the database, and the vertical axis stands for the recognition rate. (a) The results obtained when each class contains only two samples; (b) the results obtained when each class contains three samples; (c) the results obtained when each class contains six samples.

recognition rate, respectively. In addition, the '+','×' and '○' signs shown in Fig. 6 stand for the recognition results obtained using our approach, Liu et al.'s approach, and the template-matching approach, respectively. Furthermore, the data shown in Fig. 6(a)–(c) are the experimental results obtained when each class contained, respectively, 2, 3, and 6 samples. Among the three approaches, the template-matching approach performed recognition based solely on the original feature vectors. Therefore, there was no LDA involved in this process. Furthermore, from the data shown in Fig. 6, it is obvious that Liu et al.'s approach was the worst. Basically, the most serious problem which occurred in Liu's approach was the degraded discriminating capability. Although the derived discriminant vectors maximized the modified Fisher's criterion function, the optimal class separability condition, which is the objective of an LDA process, was not surely satisfied. Therefore, the projection axes determined by Liu et al.'s approach could not guarantee to provide the best class separability of all the database samples. Therefore, it is no wonder that the performance of Liu et al.'s approach was even worse than that of the template-matching approach. On the other hand, our approach was apparently superior because we forced the within-class scatter in the subspace to be zero. This constraint restricted the problem to a small domain, hence, it could be solved in a much easier way. Another advantage of our approach is that we do not need to

compute the inverse matrix. In Liu et al. [10], computation of the inverse matrix is indispensable. However, since we project all the samples onto an appropriate subspace, the computation of the inverse matrix, which is considered a time bottleneck, can be avoided.

Another advantage of our approach over Liu et al.'s approach is the training time requirement. Fig. 7 shows three sets of experiments; in each set of experiments we used different numbers of samples in a class (2 in (a), 3 in (b), and 6 in (c)). The '+' and '×' signs represent, respectively, the results obtained using our approach and Liu et al.'s approach. From Fig. 7(a)–(c), it is obvious that the training time required by Liu et al.'s approach grew exponentially when the database was augmented. The reason for this outcome was the projection axes determination process. In Liu et al.'s method, the projection axes are determined iteratively. In each iteration, their algorithm has to derive the projection vector in a recalculated subspace. Therefore, their training time is exponentially proportional to the number of classes adopted in the database. In comparison with Liu et al.'s approach, our approach requires a constant time for training. This is because our approach only has to calculate the subspace once and then derive all the projection vectors in this subspace.

The experimental results shown in Figs. 6 and 7 are comparisons between Liu et al.'s approach and ours in terms of accuracy and efficiency. In what follows, we shall compare our method with Liu et al.'s method using another important criterion – the stability criterion. Table 2 shows a set of experimental results regarding the stability test between our method and Liu et al.'s. In this

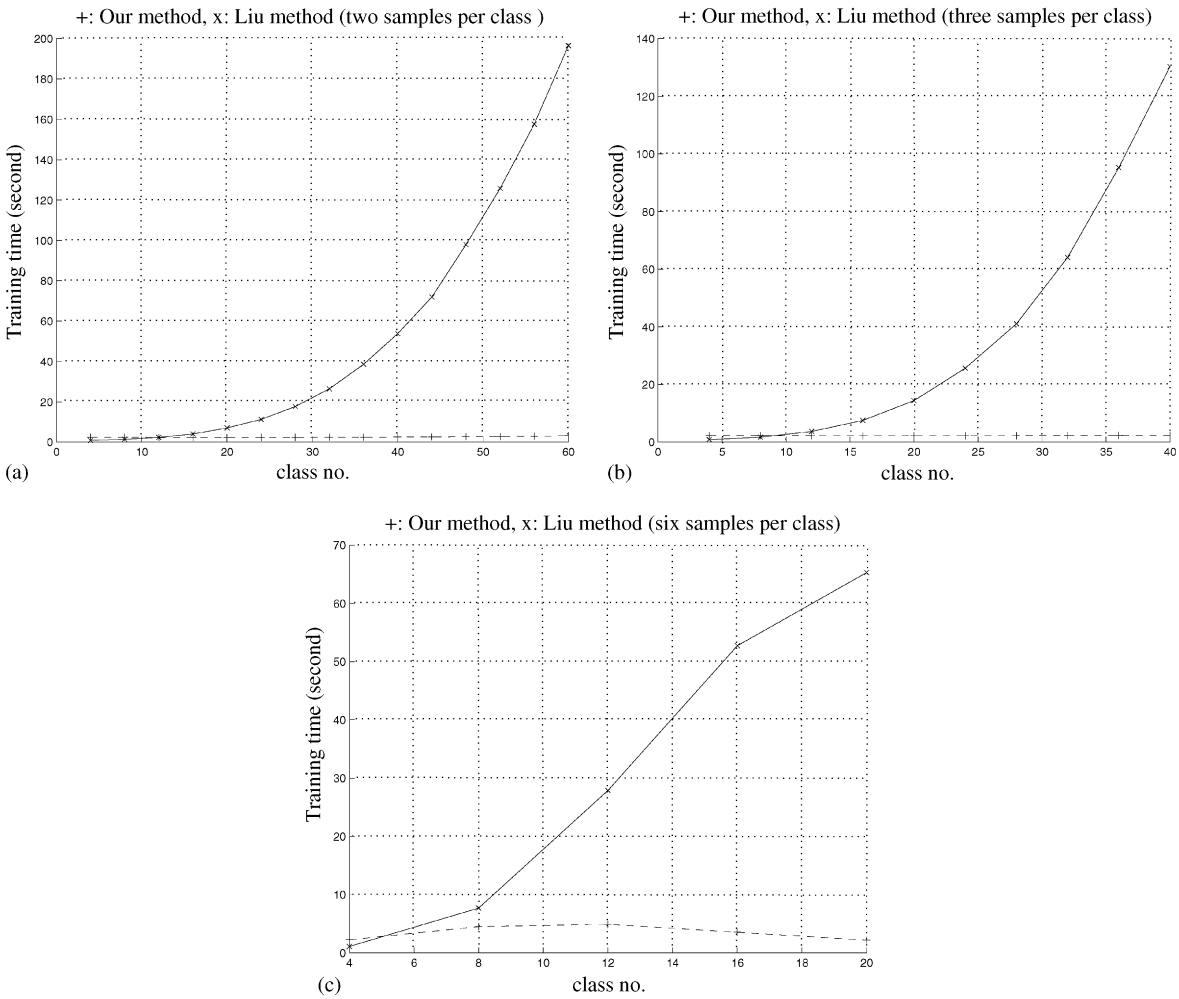

Fig. 7. Training time required by our method ('+' sign) and Liu's method ('×' sign). The horizontal axis represents the number of classes in the database, and the vertical axis stands for the training time. (a) The results obtained when each class contains only two samples; (b) the results obtained when each class contains three samples; (c) the results obtained when each class contains six samples.

Table 2
Stability test executed during the derivation of the first optimal projection vector. The training database comprised 10 classes, where each class contains three samples. The elements shown in the second and the fourth columns represent the orientation difference between the current optimal projection vector and the projection vector derived in the previous iteration

Iteration	Our method		Liu's method	
	Orientation difference (degrees)	Recognition rate (%)	Orientation difference (degrees)	Recognition rate (%)
1		90.56		90.56
2	0.0006	90.56	92.3803	90.56
3	0.0005	90.56	98.7039	90.56
4	0.0005	90.56	127.1341	90.56
5	0.0005	90.56	100.4047	90.56
6	0.0006	90.56	94.8684	90.56
7	0.0006	90.56	97.4749	88.33
8	0.0007	90.56	77.8006	90.56
9	0.0006	90.56	99.7971	90.56
10	0.0006	90.56	75.0965	90.56

set of experiments, we tried to compute the first optimal projection vector in 10 iterations. The leftmost column of Table 2 indicates the iteration number. The element shown in the second and the fourth column of Table 2 is the orientation difference (in degrees) between the current optimal projection vector and the projection vector derived in the previous iteration. The data shown in the second column were obtained by applying our method while the data shown in the fourth column were obtained by applying Liu et al.'s method. Theoretically, the optimal projection vector determined based on the same set of data should stay the same or only change slightly over 10 consecutive iterations. From Table 2, it is obvious that the projection vector determined by our method was very stable during the 10 consecutive iterations. On the other hand, the projection vector determined by Liu et al.'s method changed significantly between every two consecutive iterations. Linear algebra [23] tells us that an eigenvector will be very sensitive to small perturbation if its corresponding eigenvalue is close to another eigenvalue of the same matrix. Table 3 shows the eigenvalues obtained by our method and by Liu et al.'s. It is obvious that the eigenvalues obtained by our method are quite different from each other. However, the eigenvalues obtained by Liu et al.'s method are almost the same. These data confirm that our method was much more stable than Liu et al.'s.

Another important issue which needs to be discussed is the influence of the reserved percentage of $dim(V_0)$ on the recognition rate. Since the construction of V_0 is the most time consuming task in our approach, we would like to show empirically that by using only part of the space V_0, our approach can still obtain good recognition results. Fig. 8 illustrates the influence of the reserved percentage of $dim(V_0)$ on the recognition rate when the number of classes is changed. The '+', '×' and '○' signs indicate that there were 10, 20 and 30 classes in the database, respectively. In all of the above mentioned classes, each class contained three distinct samples. From the three curves shown in Fig. 8, it is obvious that by only reserving 10% of $dim(V_0)$, the recognition rate could still maintain 94%. Fig. 9, on the other hand, illustrates the influence of the reserved percentage of $dim(V_0)$ on the recognition rate when the number of samples in each class is changed. The '+','×' and '○' signs indicate that there were 2, 3, and 6 samples in each class, respectively. From Fig. 9, we can see that by only reserving 10% of $dim(V_0)$, the recognition rate could always reach 91%. Moreover, the results shown in Fig. 8 reflect that the information retained in the space V_0 (the null space of S_w) was more sensitive to the number of classes. This means

Table 3
The eigenvalues used to derive the first optimal projection vector. The elements shown in the left column are the eigenvalues determined using our method. The ones shown in the right column were determined using Liu et al.'s method

Eigenvalues determined using our method	Eigenvalues determined using Liu's method
3.31404839e + 04	1.00000000e + 00
2.39240384e + 04	1.00000000e + 00
1.67198579e + 04	1.00000000e + 00
1.01370563e + 04	1.00000000e + 00
6.88308959e + 03	1.00000000e + 00
7.41289737e + 03	1.00000000e + 00
2.70253079e + 03	1.00000000e + 00
5.53323313e + 03	1.00000000e + 00
3.46817376e + 03	1.00000000e + 00

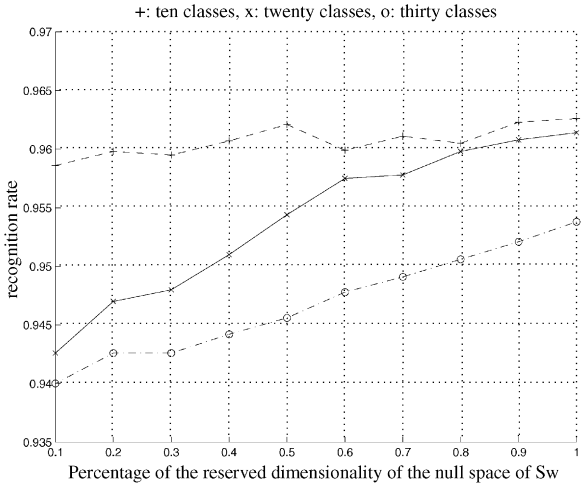

Fig. 8. Illustration of the influence of the reserved percentage of $dim(V_0)$ on the recognition rate. The '+', '×', and 'O' signs mean that there are 10, 20, and 30 classes in the database, respectively. Each class contains three distinct samples. This figure shows that the information contained in the null space of S_w was more sensitive to the number of classes in the database.

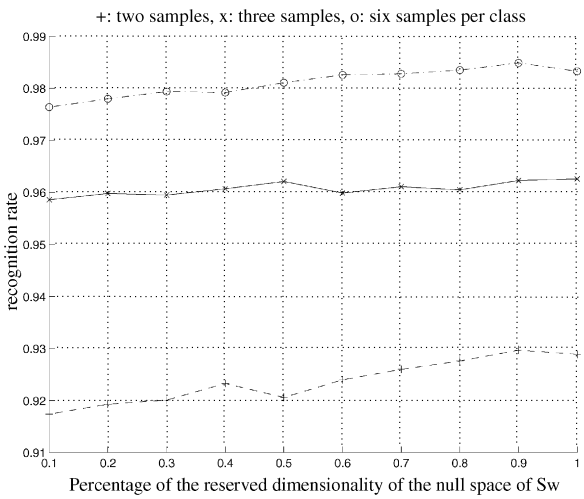

Fig. 9. Illustration of the influence of the reserved percentage of $dim(V_0)$ on the recognition rate. The '+', '×', and 'O' signs mean that there are two, three, and six samples in each class, respectively. The database comprised 10 classes. This figure shows that the information for the same person was uniformly distributed over the null space of S_w. Therefore, the percentage of $dim(V_0)$ did not influence the recognition results very much.

that when more classes are contained in the database, a higher percentage of V_0 should be reserved to obtain good recognition results. On the other hand, Fig. 9 shows that the information about the same person was uniformly distributed over the null space of S_w. Therefore, the percentage of $dim(V_0)$ did not influence the recognition results very much.

4. Concluding remarks

In this paper, we have proposed a new LDA-based face recognition system. It is known that the major drawback of applying LDA is that it may encounter the small sample size problem. When the small sample size problem occurs, the within-class scatter matrix S_w becomes singular. We have applied a theory from linear algebra to find some projection vectors $\mathbf{q}$'s such that $\mathbf{q}^t S_w \mathbf{q} = 0$ and $\mathbf{q}^t S_b \mathbf{q} \neq 0$. Under the above special circumstances, the modified Fisher's criterion function proposed by Liu et al. [10] can reach its maximum value, i.e., 1. However, we have found that an arbitrary projection vector $\mathbf{q}$ satisfying the maximum value of the modified Fisher's criterion cannot guarantee the maximum class separability unless $\mathbf{q}^t S_b \mathbf{q}$ is further maximized. Therefore, we have proposed a new LDA process, starting with the calculation of the projection vectors in the null space of the within-class scatter matrix S_w. If this subspace does not exist, i.e., S_w is nonsingular, then a normal LDA process can be used to solve the problem. Otherwise, the small sample size problem occurs, and we choose the vector set that maximizes the between-class scatter of the transformed samples as the projection axes. Since the within-class scatter of all the samples is zero in the null space of S_w, the projection vector that can satisfy the objective of an LDA process is the one that can maximize the between-class scatter. The experimental results have shown that our method is superior to Liu et al.'s approach [10] in terms of recognition accuracy, training efficiency, and stability.

References

[1] R. Chellappa, C. Wilson, S. Sirohey, Human and machine recognition of faces: a survey, Proc. IEEE 83 (5) (1995) 705–740.

[2] D. Valentin, H. Abdi, A. O'Toole, G. Cottrell, Connectionist models of face processing: a survey, Pattern Recognition 27 (9) (1994) 1209–1230.

[3] R. Brunelli, T. Poggio, Face recognition: features versus templates, IEEE Trans. Pattern Anal. Mach. Intell. 15 (10) (1993) 1042–1052.

[4] A. Samal, P. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, Pattern Recognition 25 (1) (1992) 65–77.

[5] S.H. Jeng, H.Y. Mark Liao, C.C. Han, M.Y. Chern, Y.T. Liu, Facial feature detection using geometrical face model: an efficient approach, Pattern Recognition 31 (3) (1998) 273–282.

[6] C.C. Han, H.Y. Mark Liao, G.J. Yu, L.H. Chen, Fast face detection via morphology-based pre-processing, Pattern Recognition 1999, to appear.

[7] H.Y. Mark Liao, C.C. Han, G.J. Yu, H.R. Tyan, M.C. Chen, L.H. Chen, Face recognition using a face-only database: a new approach, Proceedings of the third Asian Conference on Computer Vision, Hong Kong, Lecture Notes in Computer Science, Vol. 1352, 1998, pp. 742–749.
[8] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 696–710.
[9] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.
[10] K. Liu, Y. Cheng, J. Yang, Algebraic feature extraction for image recognition based on an optimal discriminant criterion, Pattern Recognition 26 (6) (1993) 903–911.
[11] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, N. Otsu, Face recognition system using local autocorrelations and multiscale integration, IEEE Trans. Pattern Anal. Mach. Intell. 18 (10) (1996) 1024–1028.
[12] D. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.
[13] P.N. Belhumeur, J.P. Hespanha, D.J. Kiregman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.
[15] Q. Tian, M. Barbero, Z.H. Gu, S.H. Lee, Image classification by the Foley–Sammon transform, Opt. Eng. 25 (7) (1986) 834–840.
[16] Zi-Quan Hong, Jing-Yu Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, Pattern Recognition 24 (4) (1991) 317–324.
[17] Y.Q. Cheng, Y.M. Zhuang, J.Y. Yang, Optimal fisher discriminant analysis using the rank decomposition, Pattern Recognition 25 (1) (1992) 101–111.
[18] K. Liu, Y. Cheng, J. Yang, A generalized optimal set of discriminant vectors, Pattern Recognition 25 (7) (1992) 731–739.
[19] K. Liu, Y.Q. Cheng, J.Y. Yang, X. Liu, An efficient algorithm for Foley–Sammon optimal set of discriminant vectors by algebraic method, Int. J. Pattern Recog. Artif. Intell. 6 (5) (1992) 817–829.
[20] D.H. Foley, J.W. Sammon, An optimal set of discriminant vectors, IEEE Trans. Comput. 24 (1975) 281–289.
[21] L.F. Chen, H.Y.M. Liao, C.C. Han, J.C. Lin, Why a statistics-based face recognition system should base its recognition on the pure face portion: a probabilistic decision-based proof, Proceedings of the 1998 Symposium on Image, Speech, Signal Processing and Robotics, The Chinese University of Hong Kong, September 3–4, 1998 (invited), pp. 225–230.
[22] A. Fisher, The Mathematical Theory of Probabilities, Macmillan, New York, 1923.
[23] G.W. Stewart, Introduction to Matrix Computations, Academic Press, New York, 1973.
[24] B. Noble, J.W. Daniel, Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[25] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Addison-Wesley, Reading, MA, 1992.
[26] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

About the Author—LI-FEN CHEN received the B.S. degree in computer science from the National Chiao Tung University, Hsing-Chu, Taiwan, in 1993, and she is now a Ph.D. student in the department of computer and information science at National Chiao Tung University from 1993. Her research interests include image processing, pattern recognition, computer vision, and wavelets.

About the Author—MARK LIAO received his B.S. degree in physics from the National Tsing-Hua University, Hsin-Chu, Taiwan, in 1981, and the M.S. and Ph.D. degrees in electrical engineering from the North-western University in 1985 and 1990, respectively. He was a research associate in the Computer Vision and Image Processing Laboratory at the Northwestern University during 1990–1991. In July 1991, he joined the Institute of Information Science, Academia Sinica, Taiwan, as an assistant research fellow. He was promoted to associate research fellow and then research fellow in 1995 and 1998, respectively. Currently, he is the deputy director of the same institute. Dr. Liao's current research interests are in computer vision, multimedia signal processing, wavelet-based image analysis, content-based image retrieval, and image watermaking. He was the recipient of the Young Investigators' award of Academia Sinica in 1998; the best paper award of the Image Processing and Pattern Recognition Society of Taiwan in 1998; and the paper award of the above society in 1996. Dr. Liao served as the program chair of the International Symposium on Multimedia Information Processing (ISMIP), 1997. He also served on the program committees of the International Symposium on Artificial Neural Networks, 1994–1995; the 1996 International symposium on Multi-technology Information Processing; and the 1998 International Conference on Tools for AI. Dr. Liao is an Associate Editor of the IEEE Transactions on Multimedia (1998–2001) and the Journal of Information Science and Engineering. He is a member of the IEEE Computer Society and the International Neural Network Society (INNS).

About the Author—JA-CHEN LIN was born in 1955 in Taiwan, Republic of China. He received his B.S. degree in computer science in 1977 and M.S. degree in applied mathematics in 1979, both from the National Chiao Tung University, Taiwan. In 1988 he received his Ph.D. degree in mathematics from Purdue University, USA. In 1981–1982, he was an instructor at the National Chiao Tung University. From 1984 to 1988, he was a graduate instructor at Purdue University. He joined the Department of Computer and Information Science at National Chiao Tung University in August 1988, and is currently a professor there. His recent research interests include pattern recognition and image processing. Dr. Lin is a member of the Phi-Tau-Phi Scholastic Honor Society.

About the Author—MING-TAT KO received a B.S. and an M.S. in mathematics from the National Taiwan University in 1979 and 1982, respectively. He received a Ph.D. in computer science from the National Tsing Hua University in 1988. Since then he joined the Institute of Information Science as an associate research fellow. Dr. Ko's major research interest includes the design and analysis of algorithms, computational geometry, graph algorithms, real-time systems and computer graphics.

About the Author—GWO-JONG YU was born in Keelung, Taiwan in 1967. He received the B.S. degree in Information Computer Engineering from the Chung-Yuan Christian University, Chung-Li, Taiwan in 1989. He is currently working toward the Ph.D. degree in Computer Science. His research interests include face recognition, statistical pattern recognition and neural networks.