Neurocomputing 73 (2010) 2168–2176

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An efficient method for computing orthogonal discriminant vectors

Jinghua Wang [b], Yong Xu [a,*], David Zhang [b], Jane You [b]

[a] Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, Guangdong 518055, People's Republic of China
[b] Biometrics Research Center, Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 22 September 2009
Received in revised form
16 January 2010
Accepted 4 February 2010
Communicated by D. Tao
Available online 6 March 2010

Keywords:
Fisher discriminant analysis
Pattern recognition
Orthogonal discriminant vectors
Dimension reduction

ABSTRACT

We propose a linear discriminant analysis method. In this method, every discriminant vector, except for the first one, is worked out by maximizing a Fisher criterion defined in a transformed space which is the null space of the previously obtained discriminant vectors. All of these discriminant vectors are used for dimension reduction. We also propose two algorithms to implement the model. Based on the algorithms, we prove that the discriminant vectors will be orthogonal if the within-class scatter matrix is not singular. The experimental results show that the proposed method is effective and efficient.

1. Introduction

Among dimension reduction methods, the principal component analysis (PCA) [1] is a well-known and widely used method. The goal of PCA is to produce a set of low-dimensional samples, which are the best linear approximations of the original dataset [2]. Although PCA can generate the best approximations of the original samples, PCA does not take into account the separability between samples of different classes when extracting features. It is recognized that PCA may fail in capturing much useful discriminant information [3–5].

Linear discriminant analysis (LDA) is another effective dimension reduction method. It has been successfully used in many classification problems, such as face recognition [6–8] and multimedia information retrieval. Differing from PCA, LDA searches for the projecting axes, on which the data points of different classes are far from each other while data points from the same classes are clustered together. Zhang et al. [35] proposed a systematic framework, named patch alignment, for understanding the common properties and intrinsic differences between PCA and LDA. Among all the LDA methods, the Fisher discriminant analysis [9] is the center of attention. Based on the original Fisher discriminant analysis, Wilks proposed a method that can work out the $L-1$ discriminant vectors in an L-class problem [10]. This is usually referred to as the classical linear discriminant analysis (CLDA). It has been proven that the Fisher discriminant analysis is equivalent to the maximum-likelihood (ML) parameter estimates of a Gaussian model, if all class discrimination information resides in a d-dimensional subspace of the original n-dimensional feature space and that the within-class covariance is equal for all classes [11].

We briefly present the development of the discriminant analysis as follows. Mathematically, dimension reduction can be viewed as an expression of the original samples through the use of a new coordinate system with the discriminant vectors being the coordinate axes. We can consider that CLDA uses a non-orthogonal coordinate system to express the original samples, since the discrimination vectors generated are not orthogonal. Usually orthogonal rather than non-orthogonal coordinate systems are preferred. This is largely because no redundant information exists among the resulting data components if the original data are projected onto the coordinate axes of an orthogonal coordinate system. Foley and Sammon [12] also proposed an algorithm for the two-class problem, known as the Foley–Sammon linear discriminant analysis (FSLDA), which requires the discriminant vectors to satisfy the orthogonality constraint. Okada and Tomita [13] presented an optimal orthogonal system for the discrimination analysis to obtain more discriminant vectors. In a transformed space, Duchene and Leclercq [14] proposed an orthogonal discriminant analysis. Both in [13,14], the authors claimed that an orthogonal set of vectors were more powerful than the classical discriminant vectors in terms of both discriminant ratio and mean error probability, and they performed experiments extensively to support their claims. Moreover, Liu et al. [15] presented additional comprehensive solutions for the discriminant analysis, this is also considered to be the open solution for the problem that maximizing the Fisher

* Corresponding author. Tel.: +86 755 26032458; fax: +86 755 26032461.
E-mail address: laterfall2@yahoo.com.cn (Y. Xu).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.02.009

criterion under the constraint that $W^TW=I$ where each column of the matrix W is a discriminant vector. All of the methods in [12–15] are very time consuming to obtain the orthogonal discriminant vectors for dimension reduction. Using the constraint of statistical uncorrelation, Jin et al. [16,17] proposed an uncorrelated linear discriminant analysis (ULDA). Both Xu et al. [18] and Yang et al. [19] analyzed the ULDA and pointed out that uncorrelated discriminant vectors may be superior for dimension reduction. Xu et al. [20] proposed a LDA method in which a mixed criterion is used to conduct the process of solving the discriminant vectors. We refer to this method as a mixed LDA (MLDA), for the simple reason that it has a mixed criterion.

It is also noted that LDA often suffers from a "small sample size" (SSS) problem, particularly when the number of samples available is much smaller than the dimensionality of the sample space. Many algorithms, such as pseudoinverse LDA [21], regularized LDA [22], LDA/GSVD [23], two-stage LDA [24] and the methods in [25–27] have been designed to deal with the SSS problem. Zhang et al. [41,42] designed two schemes, discriminative locality alignment (DLA) [41] and local coordinates alignment (LCA) [42], which can avoid the SSS problem. Studies demonstrated that merging of class can reduce the classification accuracy in LDA [36,37]. To improve the performance of LDA, Lotlikar and Kothari [36] developed the fractionalstep FLDA (FS-FLDA) and Loog et al. [37] proposed the approximate pairwise accuracy criterion (aPAC). Tao et al. [38] studied the geometric mean for subspace selection and proposed three criteria for subspace selection. Bian and Tao [39] replaced the geometric mean by harmonic mean in subspace selection to further reduce the class separation problem. Lu et al. [40] improved the method in [36] for the high-dimension problems.

In this paper, we propose a novel LDA method, in which every single discriminant vector is worked out by maximizing a Fisher criterion. Thus, if k discriminant vectors are needed, there should be k different Fisher criteria, all of which except for the first one are defined in new spaces different from the original sample space and the new spaces are different from each other. The Fisher criterion for the $(k+1)$th discriminant vector is defined in the null space of the first k discriminant vectors. The experimental results show that this method is effective and efficient. We note that before solving discriminant vectors, many of the previous LDA methods [12–18] perform a number of matrix operations in advance. These include matrix inverse and multiplication operations. However, for our method, the main computational burden before solving discriminant vectors is incurred by a procedure of updating the samples and the scatter matrices using given formulas in this paper.

The remainder of the paper is organized as follows. The next section briefly reviews FSLDA, ULDA and MLDA. Section 3 formulates the proposed novel LDA method and presents two algorithms to implement it. Section 4 describes the experimental results. Section 5 offers the conclusions.

2. Related work

Let $\omega_1, \omega_2, \ldots, \omega_c$ be c known pattern classes in N dimensional space. The number of samples in the ith class is denoted by l_i, and the total number of samples is $L=\sum_{i=1}^{c} l_i$. We use $x_j^i (1 \le i \le c, 1 \le j \le l_i)$ to denote the jth sample in the ith class; we can also use $x_i (1 \le i \le L)$ to represent the ith sample of the sample set. The prior probability of the ith class is $p_i (0 < p_i < 1, 1 \le i \le c)$. Since there are c classes in all, the formula $\sum_{i=1}^{c} p_i = 1$ holds.

The object of LDA can be stated as follows: calculating the rectangular matrix $W \in R^{N \times n}$ that maximizes the Fisher criterion

$$J(W) = \frac{|W^T S_b W|}{|W^T S_w W|} \tag{1}$$

where S_b and S_w are the between-class scatter matrix and within-class scatter matrix, respectively. These two matrices can be expressed in terms of the following equations:

$$S_b = \sum_{i=1}^{c} p_i (m_i - m)(m_i - m)^T \tag{2}$$

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{l_i} (x_j^i - m_i)(x_j^i - m_i)^T \tag{3}$$

where $m_i = 1/l_i \sum_{j=1}^{l_i} x_j^i$, $m = \sum_{k=1}^{L} x_k = \sum_{i=1}^{c} p_i m_i$ are the mean of the samples in the ith class and that of all the samples, respectively.

It has been proven that the matrix W should be composed of the eigenvectors corresponding to leading eigenvalues of the following generalized eigenequation:

$$S_b \alpha_i = \lambda_i S_w \alpha_i \tag{4}$$

If S_w is nonsingular, the k discriminant vectors are the k eigenvectors corresponding to the k largest eigenvalues of the matrix $S_w^{-1} S_b$. As $rank(S_b) \le c-1$ (c is the number of the sample classes), there are at most $c-1$ effective discriminant vectors [10,28,29]. While the above method calculates discriminant vectors in a simple way, it also has the characteristic where the sample features obtained may contain redundant information.

2.1. Foley–Sammon linear discriminant analysis (FSLDA) [12]

FSLDA can be briefly described as follows. It has the same first discriminant vector as CLDA. That is, its first discriminant vector is the vector that maximizes the Fisher criterion Eq. (1), i.e., the eigenvector that corresponds to the largest eigenvalue of the generalized eigenequation Eq (4). Suppose j discriminant vectors $\alpha_1, \alpha_2, \ldots, \alpha_j$ have been obtained, then the next discriminant vector can be worked out by maximizing the Fisher criterion with the following orthogonal constraints:

$$\alpha_{j+1}^T \alpha_k = 0 \quad (k = 1, 2, \ldots, j) \tag{5}$$

It has been proven that the $(j+1)$th discriminant vector is the eigenvector corresponding to the largest eigenvalue of the generalized eigenequation [17]

$$M S_b \alpha = \lambda S_w \alpha \tag{6}$$

where $M = I - D(D^T S_w^{-1} D)^{-1} D^T S_w^{-1}$, $D = [\alpha_1 \alpha_2, \cdots, \alpha_j]$. I is the identity matrix sharing the size of S_w.

2.2. Uncorrelated linear discriminant analysis (ULDA) [30]

We will first introduce the definition of the correlation coefficient between two feature components before presenting ULDA. For a pattern x, the feature component corresponding to the discriminant vector α_i is $y_i = \alpha_i^T x$, and the feature component corresponding to α_j is $y_j = \alpha_j^T x$. The covariance between y_i and y_j is defined as [19]

$$\mathrm{cov}(y_i, y_j) = E[(y_i - E(y_i))(y_j - E(y_j))] \tag{7}$$

and the correlation coefficient between them is defined using the following formula:

$$\rho(y_i, y_j) = \frac{\mathrm{cov}(y_i, y_j)}{\sqrt{\mathrm{cov}(y_i, y_i)}\sqrt{\mathrm{cov}(y_j, y_j)}} = \frac{\alpha_i^T S_t \alpha_j}{\sqrt{\alpha_i^T S_t \alpha_i}\sqrt{\alpha_j^T S_t \alpha_j}} \tag{8}$$

where $S_t = S_b + S_w$ is the total scatter matrix. This correlation coefficient measures the ways that these two feature components are correlated with each other in the new sample space.

If the feature components corresponding to different discriminant vectors are uncorrelated, then the correlation coefficient between them will be zero, which means that

$$\rho(y_i, y_j) = 0 \Leftrightarrow \mathrm{cov}(y_i, y_j) = 0 \Leftrightarrow \alpha_i^T S_t \alpha_j = 0 \tag{9}$$

The above uncorrelated constraints are employed in ULDA to generate effective projecting axes. After obtaining the first j discriminant vectors α_1, α_2 ,..., α_j, one can determine the $(j+1)$th discriminant vector by maximizing the Fisher criterion with the following uncorrelated constraints

$$\alpha_{j+1}^T S_t \alpha_i = 0 \quad (i = 1, 2, \ldots, j) \tag{10}$$

Studies [30] showed that the jth uncorrelated discriminant vector is the eigenvector corresponding to the maximal eigenvalue of the following generalized eigenequation

$$U_j S_b \alpha_j = \lambda_j S_w \alpha_j \tag{11}$$

where

$U_1 = I_N$,
$U_j = I_N - S_t D_j^T (D_j S_t S_w^{-1} S_t D_j^T)^{-1} D_j S_t S_w^{-1} (j > 1)$
$D_j = [\alpha_1 \alpha_2 \cdots \alpha_{j-1}] (j > 1)$
$I_N = diag\{1, 1, \ldots, 1\}$.

It has been proven that if: (1) $rank(S_b) = c-1$, where c is the number of class; (2) S_w is nonsingular; and (3) all the $c-1$ nonzero eigenvalues of the matrix $S_w^{-1} S_b$ are different from each other, then the $c-1$ uncorrelated optimal discriminant vectors are the same as the classical discriminant vectors [16]. It has been shown that ULDA may outperform FSLDA in some problems [30]. Xu et al. [18] also proved that the Fisher criterion ratio of each orthogonal discriminant vector is not less than that of the corresponding uncorrelated discriminant vector.

2.3. Linear discriminant analysis with a mixed criterion (MLDA) [20]

As both the maximal Fisher criterion and minimal correlation between extracted feature components are important, MLDA solves discriminant vectors based on a new criterion that is a combination of these two criteria [20].

If $j (j \geq 1)$ discriminant vectors α_1, α_2, ..., α_j are obtained, the criterion for determining the next one is

$$P(\alpha) = \gamma_0 J'(\alpha) - \sum_{i=1}^{j} \gamma_i \rho^2(\alpha, \alpha_i) \tag{12}$$

where $0 \leq \gamma_i \leq 1 (i = 0, 1, \ldots, j)$ are weighting coefficients and $\sum_{k=0}^{j} \gamma_k = 1$, $\rho(\alpha, \alpha_i)$ is the correlation coefficient defined above, and

$$J'(\alpha) = \frac{\alpha^T S_w \alpha}{\alpha^T S_t \alpha} \tag{13}$$

is the modified Fisher criterion. As $S_t = S_b + S_w$

$$J'(\alpha) = \frac{\alpha^T S_w \alpha}{\alpha^T S_t \alpha} = \frac{1}{\frac{\alpha^T S_w \alpha + \alpha^T S_b \alpha}{\alpha^T S_w \alpha}} = \frac{1}{1 + J(\alpha)} \tag{14}$$

which means that $J(\alpha)$ and $J'(\alpha)$ reach their maximum values at the same α.

Under the constraint $\alpha^T S_t \alpha = 1$, the $(j+1)$th discriminant vector of this model is the eigenvector corresponding to the maximal eigenvalue of the following generalized eigenequation [20]:

$$M\alpha = \lambda S_t \alpha \tag{15}$$

where

$$M = \gamma_0 S_b - \sum_{i=1}^{j} \gamma_i \frac{S_t(\alpha_i \alpha_i^T) S_t}{\alpha_i^T S_t \alpha_i}$$

As the criterion is composed by two components, the discriminant vector is usually not the vector that maximizes the Fisher criterion and may correlate with other discriminant vectors. However, the Fisher criterion will not be too small and the vector will not be highly correlated with others.

2.4. Discussion

As evident from Eqs. (6), (11) and (15), FSLDA, ULDA and MLDA have a common weakness where the algorithms are not efficient. For a single discriminant vector, FSLDA and ULDA have to compute three inverse matrices and many times of matrix multiplication before solving a generalized eigenequation. For the jth discriminant vector, MLDA has to perform $6^*(j-1)$ times (unfavorably increases with j) the matrix operation before solving a generalized eigenequation. In fact, MLDA makes a compromise in computing the discriminant vector, so the resulting vectors are also usually not optimal.

We can also learn from the above methods that different criteria for choosing the jth $(j > 1)$ discriminant vector have their specific concerns, but the first discriminant vector is obtained by the same way. It means that if only one discriminant vector is required, it is the very eigenvector corresponding to the maximal eigenvalue of the same generalized eigenequation.

3. A novel method

In this section, we will first propose a novel method that does not suffer from the same weaknesses of CLDA, FSLDA, ULDA and MLDA. Then we will present an algorithm to implement the novel method. Moreover, we will show an alternative algorithm that is more efficient. We will also investigate the relationship between the novel method and the other methods. Finally, we will prove that the discriminant vectors generated from the novel method are orthogonal under some conditions.

In this novel method, we intend to calculate discriminant vectors one by one, and we will design two algorithms to implement it. The main idea of the first algorithm is as follows. First, we obtained a discriminant vector α_i by maximizing the Fisher criterion that is defined based on the currently available samples. Second, we updated every sample by subtracting its component in the direction of the obtained discriminant vector α_i. In the algorithm, we repeated the above two steps until all of the discriminant vectors were worked out. The output of the second algorithm was the same as the first one, but more efficient.

3.1. Algorithms of the novel method

In this subsection, we first introduce the procedure of updating a sample based on an obtained discriminant vector before presenting the algorithms. For the original sample x, the feature component corresponding to α_i is $y_i = \alpha_i^T x$. Then, y_i also measures the length of the component of x in the direction of α_i provided that α_i is a unit vector, i.e. $\|\alpha_i\| = 1$. We note that the extracted feature y_i has some influence in determining the coming discriminant vectors. To avoid such a detrimental influence, we removed the component of every sample in the direction of the obtained discriminant vector α_i before calculating the next discriminant vector. We can also say that the samples are subsequently, updated. We use the following formula to remove the component of x in the direction of α_i:

$$x_{new} = x - \frac{\alpha_i^T x}{\alpha_i^T \alpha_i} \alpha \tag{16}$$

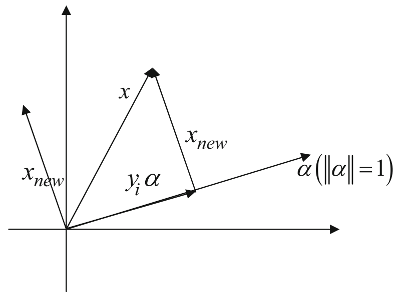

Fig. 1. Description of sample updating.

This is also shown in Fig. 1.

After updating every sample based on the discriminant vector(s) obtained in the previous step(s), we obtain a set of new samples. Then we can define a new Fisher criterion again and get another discriminant vector. Now, we present an algorithm to implement the novel method. The algorithm can be described as follows:

Step 1. Compute the sample-mean for each class using the formula $m_i = 1/l_i \sum_{j=1}^{l_i} x_j^i$, and the sample-mean of all samples using $m = \sum_{k=1}^{L} x_k = \sum_{i=1}^{c} p_i m_i$.

Step 2. Compute the between-class and within-class scatter matrices using Eqs. (2) and (3).

Step 3. Work out the eigenvector corresponding to the maximal eigenvalue of the generalized eigenequation $S_b\alpha_i = \lambda_i S_w \alpha_i$, and denote it as α_i.

Step 4. Update all the samples using Eq. (16), i.e., remove the component of each sample in the direction of the discriminant vector obtained in the last step.

Step 5. Go to step 1. The procedure is not terminated until the $c-1$ discriminant vectors are worked out.

If α_i is a unit vector, the update formula can be expressed as follows:

$$x_{new} = x - \frac{\alpha_i^T x}{\alpha_i^T \alpha_i}\alpha_i = x - \alpha_i^T x \alpha_i \tag{17}$$

As the discriminant vector is the eigenvector of the generalized eigenequation, and

$$S_b\alpha_i = \lambda_i S_w \alpha_i \Leftrightarrow S_b \frac{\alpha_i}{\|\alpha_i\|} = \lambda_i S_w \frac{\alpha_i}{\|\alpha_i\|} \tag{18}$$

the discriminant vector can always be normalized. Hereafter, the discriminant vectors are considered to have been normalized, i.e., $\|\alpha\| = 1$.

In fact, it is not necessary to update all the samples before computing the class means and scatter matrices. If the sample-mean for all the samples in the current loop is m, and the discriminant vector obtained in this loop is α_i, then we can compute the sample-mean for all samples in the next loop using the following formula:

$$\begin{aligned} m_{new} &= \frac{1}{L}\sum_{i=1}^{L} x_{new} = \frac{1}{L}\sum_{i=1}^{L}(x - \alpha_i^T x \alpha_i) \\ &= \frac{1}{L}\sum_{i=1}^{L} x - \alpha_i^T\left(\frac{1}{L}\sum_{i=1}^{L} x\right)\alpha_i = m - \alpha_i^T m \alpha_i \end{aligned} \tag{19}$$

In the same way, the sample-mean for class i is

$$m_{i\_new} = m_i - \alpha_i^T m_i \alpha_i \tag{20}$$

Then we can compute the between-class matrix based on these newly obtained class means

$$\begin{aligned} S_{b\_new} &= \sum_{i=1}^{c} p_i (m_{i\_new} - m_{new})(m_{i\_new} - m_{new})^T \\ &= \sum_{i=1}^{c} p_i \left[(m_i - \alpha_i^T m_i \alpha_i) - (m - \alpha_i^T m \alpha_i)\right]\left[(m_i - \alpha_i^T m_i \alpha_i) - (m - \alpha_i^T m \alpha_i)\right]^T \\ &= \sum_{i=1}^{c} p_i [(m_i - m) - \alpha_i^T (m_i - m)\alpha_i][(m_i - m) - \alpha_i^T (m_i - m)\alpha_i]^T \\ &= \sum_{i=1}^{c} p_i (m_i - m)(m_i - m)^T + \sum_{i=1}^{c} p_i (\alpha_i^T (m_i - m))^2 \alpha_i \alpha_i^T \\ &\quad - \sum_{i=1}^{c} p_i \alpha_i^T (m_i - m)(\alpha_i (m_i - m)^T + (m_i - m)\alpha_i^T) \\ &= S_b - \sum_{i=1}^{c} p_i \alpha_i^T (m_i - m)(\alpha_i (m_i - m)^T + (m_i - m)\alpha_i^T) \\ &\quad + \sum_{i=1}^{c} p_i (\alpha_i^T (m_i - m))^2 \alpha_i \alpha_i^T = S_b - \sum_{i=1}^{c} p_i \omega_i (\alpha_i (m_i - m)^T \\ &\quad + (m_i - m)\alpha_i^T) + \sum_{i=1}^{c} p_i \omega_i \omega_i \alpha_i \alpha_i^T \end{aligned} \tag{21}$$

where $\omega_i = \alpha_i^T (m_i - m)$.

Similarly,

$$S_{w\_new} = S_w - \sum_{i=1}^{c}\sum_{j=1}^{l_i} \omega_j^i (\alpha_i (x_j^i - m_i)^T + (x_j^i - m_i)\alpha_i^T) + \sum_{i=1}^{c}\sum_{j=1}^{l_i} \omega_j^i \omega_j^i \alpha_i \alpha_i^T \tag{22}$$

where $\omega_j^i = \alpha_i^T (x_j^i - m_i)$.

It is without doubt that these formulas can simplify the computation, and the algorithm to implement the novel LDA method can be revised as follows:

Step 1. Compute the sample-mean for each class using the formula $m_i = 1/l_i \sum_{j=1}^{l_i} x_j^i$, and the sample-mean of all samples using $m = \sum_{k=1}^{L} x_k = \sum_{i=1}^{c} p_i m_i$.

Step 2. Compute the between-class and within-class scatter matrices using the following formulas:

$$S_b = \sum_{i=1}^{c} p_i (m_i - m)(m_i - m)^T; S_w = \sum_{i=1}^{c}\sum_{j=1}^{l_i} (x_j^i - m_i)(x_j^i - m_i)^T.$$

Step 3. Work out the eigenvector corresponding to the maximum eigenvalue of the generalized eigenequation

$$S_b \alpha_i = \lambda_i S_w \alpha_i$$

and denote it as α_i where α_i is the discriminant vector obtained in this loop.

Step 4. Update the between-class scatter matrix using Eq. (21) and within-class scatter matrix using Eq. (22).

Step 5. Go to step 3, until the $c-1$ discriminant vectors are worked out.

3.2. Discussion

In this subsection, we will first analysis the time complexity of the LDA algorithms. Then we will prove that the discriminant vectors generated by the novel method are orthogonal if the within-class scatter matrix is nonsingular. After that, we will present the dimension reduction procedure using the obtained discriminant vectors.

To the best of our knowledge, in all the Fisher criterion based linear discriminant analysis methods, every discriminant vector is worked out by solving a generalized eigenequation, such as in FSLDA, ULDA and MLDA. The main difference among FSLDA, ULDA, MLDA and our method lies in the ways of obtaining the generalized eigenequation. The rest of this paragraph shows the time complexities of generating the generalized eigenequation in FSLDA, ULDA, MLDA and our method. To generate the generalized eigenequation, FSLDA and ULDA involve inverse matrix computation, the time cost of which is $O(N^3)$. In MLDA, no inverse matrix computing is needed, but much more matrix multiplication is necessary. In fact, the time cost of generating the generalized eigenequation in MLDA is $O(c^2N^2)$. Additionally, MLDA should select an appropriate value for the parameters. The method proposed in this paper takes only a time cost of $O(LN^2)$ to generate the generalized eigenequation and has no parameter.

It can be proven that the discriminant vectors obtained by the novel method are orthogonal if the within-class scatter matrices are nonsingular. We will go through the details in the coming paragraphs.

Hereafter, the discriminant vector obtained in the ith loop is denoted by $\alpha^{(i)}$. The between-class and within-class scatter matrices in the ith loop are denoted by $S_b^{(i)}$ and $S_w^{(i)}$, respectively. Here, $x_k^{j(i)}, m_k^{(i)}, m^i$ represent the counterparts of x_k^j, m_k, m in the ith loop, respectively.

Lemma 1. For any $i \geq 1$, $S_b^{(i+1)}\alpha^{(i)} = 0, \quad S_w^{(i+1)}\alpha^{(i)} = 0$.

Proof. After the ith loop, all samples are updated using Eq. (17). Then for any sample x, we have

$$(x_j^{k(i+1)})^T\alpha^{(i)} = (x_j^{k(i)} - \alpha^{(i)^T}x_j^{k(i)}\alpha^{(i)})^T\alpha^{(i)} = (x_j^{k(i)})^T\alpha^{(i)} - \alpha^{(i)^T}((x_j^{k(i)})^T\alpha^{(i)})\alpha^{(i)} = 0$$

As the sample-mean of the ith class is $m_j^{(i+1)} = 1/l_i\sum_{k=1}^{l_i} x_k^{j(i+1)}$, we have $(\alpha^{(i)})^T m_j^{(i+1)} = 0$. In the same way, $(\alpha^{(i)})^T m^{(i=1)} = 0$.

$$S_b^{(i+1)}\alpha^{(i)} = \sum_{j=1}^{c} p_j(m_j^{(i+1)} - m^{(i+1)})(m_j^{(i+1)} - m^{(i+1)})^T\alpha^{(i)}$$
$$= \sum_{j=1}^{c} p_j(m_j^{(i+1)} - m^{(i+1)})((m_j^{i+10)})^T\alpha^{(i)} - (m^{(i+1)})^T\alpha^{(i)}) = 0$$

We can also obtain, $S_w^{(i+1)}\alpha^{(i)} = 0$. □

Lemma 2. If $S_w^{(i+1)}$ is nonsingular, $(\alpha^{(i)})^T\alpha^{(i+1)} = 0$.

Proof. As $S_w{}^{(i+1)}$ is nonsingular and $S_w{}^{(i+1)}\alpha^{(i)} = 0$, we can get the following equation:

$$(S_w^{(i+1)})^{-1}\alpha^{(i)} = (S_w^{(i+1)})^{-1}(S_w^{(i+1)})^{-1}S_w^{(i+1)}\alpha^{(i)} = 0$$

In the $(i+1)$th loop

$$S_b^{(i+1)}\alpha^{(i+1)} = \lambda S_w^{(i+1)}\alpha^{(i+1)} \Leftrightarrow (S_w^{(i+1)})^{-1}S_b^{(i+1)}\alpha^{(i+1)} = \lambda\alpha^{(i+1)}(\lambda > 0)$$

From the definitions of $S_b^{(i)}$ and $S_w^{(i)}$, we know that they are symmetric matrices, and furthermore, $(S_w^{(i)})^{-1}$ is a symmetric matrix. As a result, we have

$$(S_w^{(i+1)})^{-1}\alpha^{(i)} = 0 \Leftrightarrow (\alpha^{(i)})^T(S_w^{(i+1)})^{-1} = 0$$

and

$$\lambda(\alpha^{(i)})^T\alpha^{(i+1)} = (\alpha^{(i)})^T(S_w^{(i+1)})^{-1}S_b^{(i+1)}\alpha^{(i+1)} = 0$$

As $\lambda \neq 0$, $(\alpha^{(i)})^T\alpha^{(i+1)} = 0$.

Although the matrix $S_w^{(i)}$ is usually singular, the regularization method can be used, i.e. the nonsingular matrix $S_w^{(i)} + \eta I$ (η is a positive number) can be used to replace $S_w^{(i)}$ if $|S_w^{(i)}| = 0$ [22]. □

Lemma 3. For any x, its corresponding sample in the ith ($i > 1$) loop is $x^{(i)}$. If the within-class scatter matrix is always nonsingular, then $(x^{(i)})^T\alpha^{(1)} = 0$ and $(\alpha^{(i)})^T\alpha^{(1)} = 0$.

Proof. If $i = 2$, $x^{(2)} = x - (\alpha^{(1)})^T x\alpha^{(1)}(\|\alpha^{(1)}\| = 1)$, it is easy to know that the equation $(x^{(2)})^T\alpha^{(1)} = 0$ holds. Lemma 2 tells us that $(\alpha^{(2)})^T\alpha^{(1)} = 0$ holds.

Suppose these two equations hold for all $1 < i \leq k$, then they also hold for $i = k+1$. This can be proven as follows:

As $x^{(k+1)} = x^{(k)} - (\alpha^{(k)})^T x^{(k)}\alpha^{(k)}$, we have

$$(x^{(k+1)})^T\alpha^{(1)} = (x^{(k)})^T\alpha^{(1)} - (\alpha^{(k)})^T(x^{(k)})^T\alpha^{(k)}\alpha^{(1)} = 0$$

Hence, for any sample in the $(k+1)$th loop, $(x^{(k+1)})^T\alpha^{(1)} = 0$ is certain. Then by Lemmas 1 and 2, $(\alpha^{(k+1)})^T\alpha^{(1)} = 0$. □

Theorem 1. For any $i,j(i \neq j)$, $\alpha_i^T\alpha_j = 0$.

Proof. Suppose $i > j$, then the samples in the jth loop can be considered as the original samples, and the $\alpha^{(j)}$ can be considered as the first discriminant vector. Then $\alpha^{(i)}$ can be considered as the $(i-j+1)$th discriminant vector. From Lemma 3, we know that $\alpha_i^T\alpha_j = 0$. This has demonstrated that the discriminant vectors generated by the novel method are proven to be orthogonal if the within-class scatter matrices are nonsingular. Thus, they form an orthogonal coordinate system. □

After generating all the discriminant vectors, we can obtain low dimensional data by projecting the original samples onto the discriminant vectors. The first discriminant vector $\alpha^{(1)}$ is worked out in the original sample space, so the first feature component of the original sample x is

$$y_1 = (\alpha^{(1)})^T x \tag{23}$$

We know from the algorithms that every training sample is updated before calculating a discriminant vector (except for the first one). The testing sample should also be updated in the same way in advance if the discriminant vector is employed as the projection axis. According to this, the second feature component is worked out by the following formula:

$$y_2 = (\alpha^{(2)})^T x^{(2)} = (\alpha^{(2)})^T(x - (\alpha^{(1)})^T x\alpha^{(1)}) = (\alpha^{(2)})^T(x - y_1\alpha^{(1)}) \tag{24}$$

If $x^{(1)} = x$, the generalized projection formulas are as follows:

$$\begin{cases} y_i = (\alpha^{(i)})^T x^{(i)} \\ x^{(i)} = x^{(i-1)} - (\alpha^{(i-1)})^T x^{(i)}\alpha^{(i-1)} \end{cases} \quad (i \geq 2) \tag{25}$$

Theorem 1 shows that the discriminant vectors are orthogonal if within-class scatter matrices are nonsingular, so

$$\begin{aligned} y_i &= (\alpha^{(i)})^T x^{(i)} = (\alpha^{(i)})^T(x^{(i-1)} - (\alpha^{(i-1)})^T x^{(i)}\alpha^{(i-1)}) \\ &= (\alpha^{(i)})^T x^{(i-1)} - ((\alpha^{(i-1)})^T x^{(i)})((\alpha^{(i)})^T\alpha^{(i-1)}) = (\alpha^{(i)})^T x^{(i-1)} = \cdots \\ &= (\alpha^{(i)})^T x^{(1)} = (\alpha^{(i)})^T x \end{aligned} \tag{26}$$

Then the procedure of the dimension reduction can be performed in the original sample space as follows:

$$\begin{cases} y = W^T x \\ W = [\alpha^{(1)} \cdots \alpha^{(d)}] \end{cases} \tag{27}$$

4. Experiments

In this section, we will perform experiments to test the proposed method. The first experiment is performed on a popular ORL face database, followed by the experiments on CENPARMI handwritten numeral database and several UCI datasets.

4.1. Face recognition

Face recognition can be defined as the identification of individuals using the images of their faces, assisted by a stored database of faces labeled with people's identities [31]. It is one of the most important examples of pattern recognition, where the patterns are images of faces.

Appearance-based approaches for face recognition has been extensively studied recently, in which a two dimensional image of size m by n is represented by a point in a $m \times n$ dimensional space. As $m \times n$ is usually very large, the SSS problem often occurs. In our experiments, we used Karhunen–Loeve (KL) expansion to overcome the SSS problem.

Any of the scatter matrices (between, within, and total class) can be chosen as the generalization matrix of KL expansion [32]. Here, the within-class scatter matrix is used. The within-class scatter matrix is very large. In fact, it is $m \times n$ by $m \times n$ according to its definition. Usually, it is computationally expensive to work out the eigenvalues and their corresponding eigenvectors of a large matrix. Fortunately for us, KL expansion used in our experiments can simplify the computation. We will clarify the procedure later.

The within-class scatter has the follow expression:

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{l_i}(x_j^i - m_i)(x_j^i - m_i)^T$$

Let $Z = [x_1^1 - m_1 \cdots x_{l_1}^1 - m_1 \cdots x_1^2 - m_2 \cdots x_{l_c}^c - m_c] \in \mathbb{R}^{N \times L}$, then $S_w = ZZ^T$. Generally speaking, $rank(S_w) = L-c$ provided that all of the samples are independent, and there should be $L-c$ nonzero eigenvectors of S_w. Let $S'_w = Z^TZ \in \mathbb{R}^{L \times L}$, then the size S'_w is much smaller, and it is relatively easy to get its $L-c$ nonzero eigenvalues λ_i and the corresponding eigenvectors β_i. By the singular decomposition theorem in algebraic theory, the $L-c$ eigenvalues of S'_w are the $L-c$ nonzero eigenvalues of S_w and the eigenvector α_i of S_w corresponding to the eigenvalue λ_i can be calculated using the following formula:

$$\alpha_i = \frac{1}{\sqrt{\lambda_i}} Z\beta_i \quad (28)$$

Therefore, a linear transformation from $\mathbb{R}^N$ to $\mathbb{R}^{L-c}$ can be defined by

$$Y = HX = [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_{L-c}]^T X \quad (29)$$

where each column of $X \in \mathbb{R}^{N \times L}$ is a training sample and $Y \in \mathbb{R}^{(L-c) \times L}$. Then, the new within-class scatter matrix $\overline{S_w}$ in the lower dimensional space $\mathbb{R}^{L-c}$ can be computed using the following formula:

$$\overline{S_w} = HS_wH^T \quad (30)$$

As $\alpha_i (1 \le i \le L-c)$ are the eigenvectors of S_w that corresponding to different nonzero eigenvalues, the new within-class scatter matrix is nonsingular. In other words, the SSS problem does not occur.

The ORL is one of the most popular face image databases [17]. This database contains ten face images each for 40 different people. In order to provide suitable research material, the images of this database were taken at different times, and in various lighting. To model the faces in daily life, the faces had different expressions (open/closed eyes, smiling/not smiling) and some of them were facilitated with details (glasses/no glasses). Fig. 2 depicts the ten face images of a subject in the ORL database.

Half of the images, five from each person, were used for training and the rest are used for testing in this experiment. The images were originally 92×112. We obtained smaller sized images as follows. We used a pixel to represent every 2×2 sub-image, which was the average value of the four pixels. As a result, the original image was resized to 46×56. In the same way, we worked out the size for images that were 23×28.

In our experiments, we used the following six steps to accomplish the task of face recognition to implement the method proposed in this paper:

Step 1. Calculate the generalization matrix S'_w of KL expansion together with the eigenvalues and their corresponding eigenvectors.

Step 2. Calculate the eigenvectors of the within-class matrix S_w.

Step 3. Perform the linear compression transformation using the eigenvectors worked out in the last step to make S_w nonsingular.

Step 4. Calculate the total, between class and within class scatter matrices in the new space.

Step 5. Work out all the discriminant vectors using the proposed algorithm.

Step 6. Perform another linear transformation using the discriminant vectors obtained in the last step and perform classification.

It is known that we need at least, the $c-1$ linear features, to form a sufficient statistic solution in a c-class classification problem. In our experiments, we calculated no more than 39 discriminant vectors in this 40-classes problem, and all of them were used for dimension reduction. We used the nearest neighbor classifier in step 6. The average classification accuracies of ten runnings (the training samples were randomly selected in each running) are presented in the following three figures (Figs. 3(a), (b), and (c)). The proposed method is compared with DLA [41], MLDA [20], LCA [42], ULDA [16] and FSLDA [13]. Fig. 3(a) shows

Fig. 2. Samples from the ORL database.

the results of the algorithms when taking the 92×112 images as inputs, Figs. 3(b) and (c) show the results when 46×56 and 23×28 images are used.

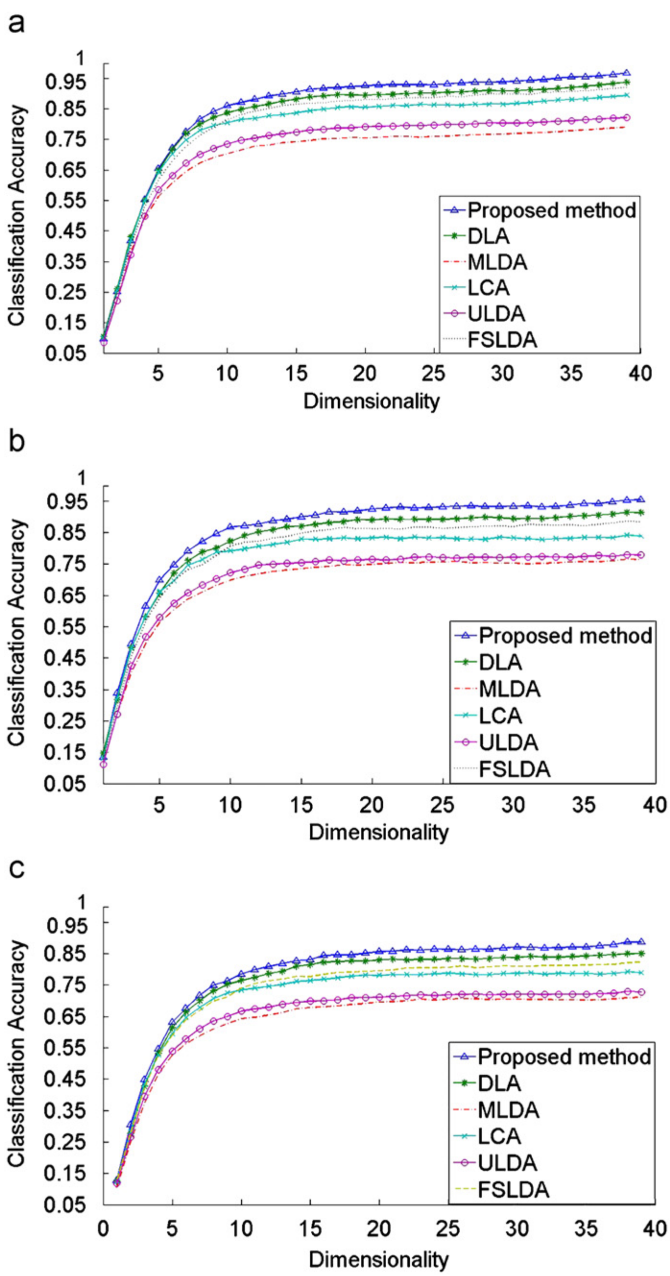

Fig. 3. Experimental results on ORL: (a) 92×112 images; (b) 46×56 images; (c) 23×28 images.

As evident from the figures, the proposed method performs better than the other five methods in terms of classification accuracy. Its running time is less than one-tenth of the FSLDA which also uses the orthogonal discriminant vectors for dimension reduction in our experiments. Moreover, because the FSLDA has to perform matrix inversion and many times of matrix multiplication and the dimensionality of the matrices are high, the resultant vectors maybe not as precise as the results of our methods.

4.2. Experiments on other datasets

We first tested FSLDA, ULDA, MLDA, DLA, LCA and our method using the CENPARMI handwritten numeral database from the Concordia University. This database contains the samples of ten digits (from 0 to 9), with each having 600 samples. The 256-dimensional Gabor transformation feature [33] was used in the experiment. We then tested these different LDA methods using the UCI datasets [34]. It was noted that the UCI datasets were quite different in size, sample numbers and feature numbers [34]. For each dataset, we first randomly portioned the database into two halves. Then we tested different methods with the first half being the training set and the second half being the testing set. We repeated this testing ten times for each dataset. We calculated the average recognition rate of the ten tests and it is shown in Table 1.

Table 1 shows that our method produces a higher, at least not a lower recognition rate than the other methods. We should point out that the samples in the Zoo dataset are linearly separable after dimensional reduction and all of the methods obtain the same recognition rate. It means that under this circumstance, a set of linearly separable samples can be obtained after dimension reduction. The results show that our method can present these discriminant vectors. One the other hand, samples of different classes from the Lung Cancer dataset have bad linear separability and there is no method that produces a recognition rate higher than 75%. However, our method produces a much higher rate of 82.3%.

5. Conclusion

In this paper, we have proposed a novel linear discriminant analysis method. In the method, every discriminant vector is obtained through maximizing a Fisher criterion defined in the null space of the previously obtained discriminant vectors. We can see that the Fisher criteria for different discriminant vectors are defined in different spaces. After obtaining each discriminant vector, we update every sample using the given formula. Then we define a new Fisher criterion based on the new samples. We also present two efficient formulas to calculate the new scatter matrices. Then we prove that the discriminant vectors obtained by this method are orthogonal if the within-class matrix is not singular.

Table 1
The average recognition rate of four LDA methods.

	CENPARMI (%)	Car (%)	Iris (%)	Wine (%)	Lung cancer (%)	Glass (%)	E. coli (%)	Zoo (%)
FSLDA	80.9	98.0	97.3	97.7	70.5	88.7	80.0	100
LCA	78.3	97.9	96.8	93.5	71.1	87.5	76.2	100
ULDA	77.5	97.5	97.6	98.9	70.5	88.0	81.7	100
DLA	79.8	96.2	97.4	96.4	74.6	85.4	78.5	100
MLDA	80.1	96.6	97.3	94.4	62.5	85.3	71.1	100
Our method	84.6	98.7	98.6	97.9	82.3	89.9	81.4	100

The proposed method mainly has the following advantages as a dimension reduction method:

- Before calculating a discriminant vector, the samples are updated by subtracting its component in the direction of the previous discriminant vector. Then the obtained discriminant vectors have no detrimental influence in determining the coming discriminant vectors. The Fisher criterion is redefined based on the updated samples.
- The discriminant vectors are proven to be orthogonal if the within-class scatter matrix is nonsingular. These discriminant vectors form an orthogonal coordinate system to express the original samples, so that no redundant information exists among the resulting data components.
- The formulas for updating the scatter matrices are employed to simplify computation and there is no inverse matrix computation before solving the discriminant vector. Hence, the novel method is much more efficient than other methods [12,13,16,17,20]. Since the proposed method needs fewer matrix computations than other methods, it also causes fewer truncation errors.
- The proposed method is free from parameters.

The experimental results on the ORL dataset, CENPARMI and seven UCI datasets provide sound proof that the proposed method is an efficient and effective LDA method.

Acknowledgements

The authors are most grateful for the advice and comments from the anonymous reviewers for the revision of the manuscript. The funding support from Hong Kong Government under its GRF scheme with Grants PolyU 5322/07E and PolyU 5341/08E and the research support from the Hong Kong Polytechnic University are greatly appreciated.

This article is partly supported by the National Nature Science Committee of China under Grant nos. 60602038, 60632050, 60702076 and 863, program project under Grant nos. 2006AA01Z193, 2007AA01Z195, the Natural Science Foundation of Guangdong Province under Grant 06300862 and GRF grants from Hong Kong Government.

We wish to thank K. Liu and C.Y. Suen at the Concordia University for their support with the CENPARMI handwritten numeral database, and the AT&T Laboratories Cambridge for the ORL database. We also want to thank Dong Xu, Dongxiu Sun and Linlin Chen for their help.

References

[1] Y. Xu, D. Zhang, J.-Y. Yang, A feature extraction method for use with bimodal biometrics, Pattern Recognition 43 (2010) 1106–1115.
[2] X. Sheng, D.Q. Dai, Improved discriminant analysis for high-dimensional data and its application to face recognition, Pattern Recognition 40 (2007) 1570–1578.
[3] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative common vectors for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (1) (2005) 4–13.
[4] D.Q. Dai, P.C. Yuen, Regularized discriminant analysis and its application to face recognition, Pattern Recognition 36 (3) (2003) 845–847.
[5] X. Wang, X. Tang, Dual-space linear discriminant analysis for face recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2 (2004) 564–569.
[6] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2) (1997) 153–158.
[7] H.M. Lee, C.M. Chen, J.M. Chen, Y.L. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Transactions on Systems, Man, and Cybernetics B 31 (3) (2001) 426–432.
[8] N.R. Pal, V.K. Eluri, Two efficient connectionist schemes for structure preserving dimensionality reduction, IEEE Transactions on Neural Networks 9 (6) (1998) 1142–1154.
[9] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 178–188.
[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Second Ed., China Machine Press, Beijing, 2004.
[11] N.A. Campbell, Canonical variate analysis—a general model formulation, Australian Journal of Statistics 26 (1984) 86–96.
[12] D.H. Foley, J.W. Sammon, An optimal set of discriminant vectors, IEEE Transactions on Computers 24 (3) (1975) 281–289.
[13] T. Okada, S. Tomita, An optimal orthonormal system for discriminant analysis, Pattern Recognition 18 (2) (1985) 139–144.
[14] J. Duchene, S. Leclercq, An optimal transformation for discriminant and principal component analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (6) (1988) 978–983.
[15] K. Liu, Y.Q. Cheng, J.Y. Yang, Algebraic feature extraction for image recognition based on an optimal discrimination criterion, Pattern Recognition 26 (6) (1993) 903–911.
[16] Z. Jin, J. Yang, Z. Tang, Z. Hu, A theorem on the uncorrelated optimal discrimination vectors, Pattern Recognition 34 (10) (2001) 2041–2047.
[17] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face recognition based on the uncorrelated discriminant transformation, Pattern Recognition 34 (2001) 1405–1416.
[18] Y. Xu, J.Y. Yang, Z. Jin, Theory analysis on FSLDA and ULDA, Pattern Recognition 36 (2003) 3031–3033.
[19] J. Yang, J. Yang, D. Zhang, What's wrong with Fisher criterion? Pattern Recognition 35 (11) (2002) 2665–2668.
[20] Y. Xu, J.Y. Yang, Z. Jin, A novel method for Fisher discriminant analysis, J. Pattern Recognition 37 (2004) 381–384.
[21] R.P. Duin, Small sample size generalization, Proceedings of Ninth Scandinavian Conference on Image Analysis (1995) 957–964.
[22] J.H. Friedman, Regularized discriminant analysis, Journal of the American Statistical Association 84 (405) (1989) 165–175.
[23] P. Howland, H. Park, Generalizing discriminant analysis using the generalized singular value decomposition, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (8) (2004) 995–1006.
[24] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[25] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (10) (2000) 2385–2404.
[26] S. Mika, G. Rätsch, J. Weston, B.S. chölkopf, K.R. Müller, Fisher discriminant analysis with kernels, in: Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing IX, 1999, pp. 41–48.
[27] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, K.R. Müller, Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 623–628.
[28] H. Gao, J.W. Davis, Why direct LDA is not equivalent to LDA, Pattern Recognition 39 (2006) 1002–1006.
[29] S. Zhang, T. Sim, Discriminant subspace analysis: a Fukunaga–Koontz approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1732–1745.
[30] Z. Jin, Z. Lou, J.Y. Yang, An optimal discriminant plan with uncorrelated features, Pattern Recognition and Artificial Intelligence 12 (3) (1999) 334–339 (in Chinese).
[31] C. Hakan, M. Neamtu, Discriminative common vectors for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (1) (2005) 4–13.
[32] Z. Jin, Z.S. Hu, J.Y. Yang, A face recognition method based on the BP neural network, Journal of Computer Research and Development 36 (3) (1999) 274–277 (in Chinese).
[33] H. Yoshihiko, Recognition of handwritten numerals using Gabor features, in: Proceedings of the 13th ICPR, 250–253, 1996.
[34] A. Asuncion, D.J. Newman, UCI Machine Learning Repository [〈http://www.ics.uci.edu/∼mlearn/MLRepository.html〉]. Irvine, CA: University of California, School of Information and Computer Science, 2007.
[35] T. Zhang, D. Tao, X. Li, J. Yang, Patch Alignment for Dimensionality Reduction, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1299–1313.
[36] R. Lotlikar, R. Kothari, Fractional-step dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (6) (2000) 623–627.
[37] M. Loog, R.P.W. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise Fisher criteria, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (7) (July 2001) 762–766.
[38] D. Tao, X. Li, X. Wu, S.J. Maybank, Geometric mean for subspace selection, IEEE Transaction on Pattern Analysis and Machine Intelligence 31 (2) (2009) 260–274.
[39] W. Bian, D. Tao. Harmonic Mean for Subspace Selection, in: Proceedings of the 19th International Conference on Pattern Recognition, 2008, pp. 1–4.
[40] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using LDA based algorithms, IEEE Transactions on Neural Networks 14 (1) (2003) 195–200.
[41] T. Zhang, D. Tao, J. Yang, Discriminative locality alignment, in: Proceedings of the 10th European Conference on Computer Vision: Part I, Marseille, France, 2008, pp. 725–738.

[42] T. Zhang, X. Li, D. Tao, J. Yang, Local coordinates alignment and its linearization, International Journal of Pattern Recognition and Artificial Intelligence 4984 (2008) 643–652.

Jinghua Wang received the B.S. degree in Computer Science from Shandong University in 2005. He received his M.S. degree from Harbin Institute of Technology in 2009. He is currently a Ph.D. candidate in the Department of Computing, The Hong Kong Polytechnic University. His current research interests are in the areas of pattern recognition and image processing.

Yong Xu was born in Sichuan, China, in 1972. He received his B.S. degree, M.S. degree in 1994 and 1997 respectively. He received the Ph.D. degree in Pattern recognition and Intelligence System at NUST(China) in 2005. Now he works at Shenzhen graduate school, Harbin Institute of Technology. His current interests include feature extraction, biometric, face recognition, machine learning and image processing.

David Zhang graduated in Computer Science from Peking University in 1974. In 1983 he received his MSc in Computer Science and Engineering from (HIT) and then in 1985 his PhD from the same institution. In 1994 he received his second PhD in Electrical and Computer Engineering from the University of Waterloo, Ontario, Canada. Professor Zhang is currently at the Hong Kong Polytechnic University where he is the Founding Director of the Biometrics Technology Center a body supported by the Hong Kong SAR Government. He also serves as Adjunct Professor in Tsinghua University, Shanghai Jiao Tong University, Harbin Institute of Technology, and the University of Waterloo. Professor Zhang's research interests include automated biometrics-based authentication, pattern recognition, and biometric technology and systems. Professor Zhang holds a number of patents in both the USA and China and is a current Croucher Senior Research Fellow.

Jane You received her B.Eng. degree in electronic engineering from Xi'an Jiaotong University, P.R. China, in 1986, and obtained her Ph.D. degree in Computer Science from La Trobe University, Australia, in 1992. Currently Dr. You is an associate professor in Department of Computing, the Hong Kong Polytechnic University. Her research interests include medical imaging, visual information retrieval, pattern recognition, multimedia systems, biometrics computing and data mining.