Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 1570–1578
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Improved discriminate analysis for high-dimensional data and its application to face recognition

Xiao-Sheng Zhuang, Dao-Qing Dai*

Center for Computer Vision and Department of Mathematics, Faculty of Mathematics and Computing, Sun Yat-Sen (Zhongshan) University, Guangzhou 510275, China

Received 17 October 2005; received in revised form 16 April 2006; accepted 20 November 2006

Abstract

Many pattern recognition applications involve the treatment of high-dimensional data and the small sample size problem. Principal component analysis (PCA) is a common used dimension reduction technique. Linear discriminate analysis (LDA) is often employed for classification. PCA plus LDA is a famous framework for discriminant analysis in high-dimensional space and singular cases. In this paper, we examine the theory of this framework and find out that even if there is no small sample size problem the PCA dimension reduction cannot guarantee the subsequent successful application of LDA. We thus develop an improved discriminate analysis method by introducing an inverse Fisher criterion and adding a constrain in PCA procedure so that the singularity phenomenon will not occur. Experiment results on face recognition suggest that this new approach works well and can be applied even when the number of training samples is one per class.

Keywords: Linear discriminant analysis; Principal component analysis; Small sample size problem; Feature extraction; Face recognition

1. Introduction

Linear discriminate analysis (LDA) is a useful tool for pattern classification. Although successful in many cases, many LDA-based algorithms suffer from the so-called "small sample size problem" (SSS) [1] which exists in high-dimensional pattern recognition tasks, where the number of available samples is smaller than the dimensionality of the samples. An active field where such problem appears is image retrieval/classification. In particular, face recognition (FR) [2,3] technique has found a wide range of applications. As a result, numerous FR algorithms have been proposed, and theories related to these fields have been studied. Among various solutions to this problem, the most successful are those appearance-based approaches, such as Eigenfaces and Fisherfaces [4–7], are built on these techniques or their variants. Since SSS problems are common, it is necessary to develop new and more effective algorithms to deal with them. A number of regularization techniques that might alleviate this problem have been suggested. Mika et al. [8,9] used the technique of making the inner product matrix nonsingular by adding a scalar matrix. Baudat and Anouar [10] employed an orthogonal decomposition technique to avoid the singularity by removing the zero eigenvalues. In Refs. [11–13] the technique of regularization was used. A well-known approach, called Fisher discriminant analysis (FDA), to avoid the SSS problem was proposed by Belhumeur et al. [4]. This method consists of two steps. The first step is the use of principal component analysis (PCA) for dimensionality reduction. The second step is the application of LDA for the transformed data. The basic idea is that after the PCA step the within-class scatter matrix for the transformed data is not singular. Although the effectiveness of this framework in face recognition is obvious, see Refs. [3,4,14,15], and the theoretical foundation for this framework has also been laid [16,17], yet in this paper we find out that the PCA step cannot guarantee the successful application of subsequent LDA, the transformed within-class scatter matrix might still be singular.

On the other hand, many researchers have been dedicated to searching for more effective discriminant subspaces [16–23]. A significant result is the finding that there exists crucial discriminative information in the null space of the within-class

* Corresponding author. Tel.: +86 20 8411 0141; fax: +86 20 8403 7978.
E-mail address: stsddq@mail.sysu.edu.cn (D.-Q. Dai).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.11.015

scatter matrix. This kind of discriminative information is called irregular discriminant information, in contrast with regular discriminant information outside of the null space [24].

Unfortunately, in order to proceed LDA after PCA, many of the above methods discard the discriminant information contained in the null space of the within-class scatter matrix, yet this discriminant information is very effective for the SSS problem. Chen et al. [19] emphasized the irregular information and proposed a more effective way to extract it, but overlooked the regular information. Yu and Yang [16] took two kinds of discriminatory information into account and suggested extracting them within the range space of the between-class scatter matrix. Since the dimension of the range space is up to $K-1$, Yu et al.'s algorithm, direct LDA (DLDA), is computationally more efficient for SSS problems in that the computational complexity is reduced to be $\mathcal{O}(K^3)$.

This paper is the full version of Ref. [25]. Motivated by the success and power of the two-phase framework (PCA plus LDA) in pattern regression and classification tasks, considering the importance of the irregular information in the null space of the within-class scatter matrix, and in view of the limitation of the PCA step, we propose a new framework for the SSS problem. The algorithm of our new method modifies the procedure of PCA and derives the regular and irregular information from the within-class scatter matrix by a new criterion, which is called inverse Fisher discriminant criterion.

The rest of this paper is organized as follows. Since our method is built on PCA and LDA, in Section 2, we start the analysis by briefly reviewing the two latter methods. We point out the deficiency of the PCA plus LDA method through an example. Following that, the proposed framework is introduced and analyzed in Section 3. The relationship of the two different frameworks is also discussed. Algorithm of the new framework and the computational complexity of the algorithm will be considered, too. In Section 4, experiments with face image data are presented to demonstrate the effectiveness of the new method. Conclusions are summarized in Section 5.

2. The PCA plus LDA approach and its deficiency

In this section, we will outline the schemes of PCA procedure and LDA procedure briefly. These two procedures provide us a solid theoretical foundation for the new algorithm that will be presented in Section 3 and it is the fundamentals from which our new framework can be derived. After that, we will make some comments about the PCA plus LDA schemes and give an example to demonstrate that the LDA may fail after applying PCA to lower the dimension of the feature space.

For convenience, we consider the face recognition problem (K-class problem) as follows. Suppose there are K classes, labelled as $G_1, G_2, \ldots, G_K$. For face recognition, we randomly select n_i samples from each class $G_i, i = 1, 2, \ldots, K$, for training:

$$\begin{aligned} G_1: &\quad X_1^{(1)} \quad X_2^{(1)} \quad \cdots \quad X_{n_1}^{(1)}, \\ G_2: &\quad X_1^{(2)} \quad X_2^{(2)} \quad \cdots \quad X_{n_2}^{(2)}, \\ &\quad \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \\ G_K: &\quad X_1^{(K)} \quad X_2^{(K)} \quad \cdots \quad X_{n_K}^{(K)}. \end{aligned}$$

In face recognition applications, $X_i^{(j)}$'s are $d \times 1$ vectors representing image of size ($d = m \times n$).

2.1. The PCA procedure

Since d is very large in most cases, it is difficult to deal with the sample matrix directly. Therefore, it is quite necessary to lower the dimension of the image space first. Many methods on this problem have been proposed, among which PCA is one of the most well-known method. PCA, also known as Karhunen–Loeve (K–L) method, is a technique now commonly used for dimensionality reduction. Simply, its object is to provide a sequence of best linear approximations in low feature space to the original data set in high-dimensional space and at the same time keep the elements in the sequence uncorrelated to each other.

More precisely, rearrange the N ($N = \sum_{i=1}^{K} n_i$) samples as a $d \times N$ matrix X, where

$$\begin{aligned} X &= (X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, X_2^{(2)}, \ldots, \\ &\qquad X_{n_2}^{(2)}, \ldots, X_1^{(K)}, X_2^{(K)}, \ldots, X_{n_K}^{(K)}) \\ &= (X_1, X_2, \ldots, X_N) \end{aligned}$$

is a $d \times N$ matrix. Set $\mu = (1/N)\sum_{i=1}^{N} X_i$ and $\tilde{X} = (X_1 - \mu, X_2 - \mu, \ldots, X_N - \mu)$, then $\tilde{X}$ is the centered matrix of X and we can define the total scatter matrix S_t as

$$S_t = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)(X_i - \mu)^{\mathrm{T}} = \frac{1}{N}\tilde{X}\tilde{X}^{\mathrm{T}}. \tag{1}$$

The goal of PCA is to find out a linear transformation or projection matrix $W_{PCA} \in \mathbb{R}^{d \times d'}$ that maps the original d-dimensional image space into an d'-dimensional feature space ($d' < d$) and maximize the determinant of the total scatter of the projected samples, i.e.,

$$\begin{aligned} W_{PCA} &= \arg\max_{W \in \mathbb{R}^{d \times d'}} |W^{\mathrm{T}} S_t W| \\ &= [w_1, w_2, \ldots, w_{d'}] \\ &\text{s.t. } \|w_i\| = 1, \quad i = 1, 2, \ldots, d', \end{aligned} \tag{2}$$

where $\{w_i | i = 1, 2, \ldots, d'\}$ is the set of d'-dimensional eigenvectors of S_t corresponding to the d' largest eigenvalues, which can be obtained by single-value decomposition (SVD). The new feature vectors $Y_i \in \mathbb{R}^{d'}$ are defined by $Y_i = W_{PCA}^{\mathrm{T}} \cdot X_i, i = 1, 2, \ldots, N$.

2.2. The LDA procedure

LDA is another technique that has been successfully used for many classification problems, such as voice recognition, face recognition, and multimedia information retrieval. In statistical theory, LDA arises in those classification problems assuming that the densities of all classes are multivariate Gaussian with a common covariance matrix. In such assumption, it is very easy to deduce the decision boundaries of each class and use them for classification. For recognition, the aim of LDA is also to find a projection matrix as in PCA that maximizes the so-called *Fisher criterion*. Before we introduce it, we must define two matrices first.

Again, $X_i^{(j)}$'s ($i=1, 2, \ldots, n_j$, $j=1, 2, \ldots, K$) are samples from classes G_j, $j=1, 2, \ldots, K$. Set $\mu_j=(1/n_j)\sum_{i=1}^{n_j} X_i^{(j)}$, $j=1, 2, \ldots, K$, and μ is defined as previous. Then μ_j is the mean value of class G_j and μ is the mean value of all samples. Let the between-class scatter matrix be defined as

$$S_b = \frac{1}{N}\sum_{j=1}^{K} n_j(\mu_j-\mu)(\mu_j-\mu)^{\mathrm{T}} \tag{3}$$

and the within-class scatter matrix be defined as

$$S_w = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{n_j}(X_i^{(j)}-\mu_j)(X_i^{(j)}-\mu_j)^{\mathrm{T}}. \tag{4}$$

It is easy to verify that $S_t = S_b + S_w$. Now, the projection matrix $W_{LDA} \in \mathbb{R}^{d\times d'}$ of LDA is chosen as a matrix with orthonormal columns maximizing the following quotient, called *Fisher criterion*:

$$W_{LDA} = \arg\max_{W\in\mathbb{R}^{d\times d'}} \frac{|W^{\mathrm{T}}S_bW|}{|W^{\mathrm{T}}S_wW|} = [w_1, w_2, \ldots, w_{d'}], \tag{5}$$

where $\{w_i|i=1, 2, \ldots, d'\}$ is the set of generalized eigenvectors of S_b and S_w corresponding to the d' largest generalized eigenvalues $\{\lambda_i|i=1, 2, \ldots, d'\}$, i.e.,

$$S_bw_i = \lambda_i S_w w_i, \quad i=1, 2, \ldots, d'.$$

When the inverse of S_w exists, the generalized eigenvectors can be obtained by eigenvalue decomposition of $S_w^{-1}S_b$. The new feature vectors $Y_i \in \mathbb{R}^{d'}$ are defined by $Y_i = W_{LDA}^{\mathrm{T}} \cdot X_i$, $i=1, 2, \ldots, N$.

2.3. The deficiency of PCA plus LDA approach

For PCA, it will compute a vector that has the largest variance, while for LDA it will compute a vector which best discriminates between two classes. Consequently, in the case when the dimension is not very large and the sample size is not relatively small, LDA will usually outperform PCA for classification tasks [5]. However, in real-world applications, especially in face/image recognition, because of the high dimensionality, LDA suffers from two aspects of difficulties: the singularity of the within-class scatter matrix S_w and the computational difficulty in the high-dimensional feature space.

The so-called PCA plus LDA approach [4] is a very popular technique which intends to overcome such circumstances. The theoretical foundation for the reason why this two-phase framework can perform quite well has been also proposed by Yang et al. [17]. This approach consists of two steps: in the first step, it applies PCA to lower the dimensionality from d to d' and get the projection matrix W_{PCA}. In the second step, it applies LDA to find out the feature representation in the lower dimension feature space $\mathbb{R}^{d'}$ and obtain the transformation matrix W_{LDA}. Thus, the transformation matrix of the PCA plus LDA approach is given by

$$W_{opt}^{\mathrm{T}} = W_{LDA}^{\mathrm{T}} \cdot W_{PCA}^{\mathrm{T}}, \tag{6}$$

where W_{PCA} is the result of optimization problem (2), and

$$W_{LDA} = \arg\max_{W} \frac{|W^{\mathrm{T}}W_{PCA}^{\mathrm{T}}S_bW_{PCA}W|}{|W^{\mathrm{T}}W_{PCA}^{\mathrm{T}}S_wW_{PCA}W|} = \arg\max_{W} \frac{|W^{\mathrm{T}}S_b'W|}{|W^{\mathrm{T}}S_w'W|}. \tag{7}$$

As is well known, the rank of the within-class scatter matrix $S_w \in \mathbb{R}^{d\times d}$ satisfies $rank(S_w) \leqslant \min\{d, N-K\}$. When small sample size problem occurs, i.e., $N < d+K$, the within-class scatter matrix S_w is singular, hence the optimization problem (5) is not solvable. In order to perform LDA (7) after PCA procedure (2), in PCA step, d' must be an integer no large than $N-K-c$, where c is a positive integer generally equal to 1. In most cases, the within-class scatter matrix S_w' would not be singular after PCA procedure. However, it is possible that the within-class scatter matrix S_w' might be still singular even after the PCA step reducing the dimension d to $d'=N-K-c$. The following simple example demonstrates the possibility of such case.

Suppose $d=2$, $K=2$, $N=4\kappa$, $n_1=n_2=2\kappa$, where κ is a positive integer. We consider a two-class problem in $\mathbb{R}^2$. Let the data be given by

$$X = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 & -1 & -1 & \cdots & -1 & -1 \\ \sigma_1 & -\sigma_1 & \cdots & \sigma_\kappa & -\sigma_\kappa & \sigma_1 & -\sigma_1 & \cdots & \sigma_\kappa & -\sigma_\kappa \end{pmatrix},$$

the first 2κ vectors (columns) belong to class 1 and the last 2κ vectors belong to class 2, σ_i is a constant such that $|\sigma_i|<1$, $i=1, \ldots, \kappa$.

By simple calculation, we have for the total scatter matrix S_t, the within-class scatter matrix S_w and the between-class scatter matrix S_b

$$S_t = \begin{pmatrix} 1 & 0 \\ 0 & \sum_{i=1}^{\kappa}\sigma_i^2/\kappa \end{pmatrix}, \quad S_w = \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^{\kappa}\sigma_i^2/\kappa \end{pmatrix}$$

$$\text{and} \quad S_b = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

S_t is already diagonal, the two eigenvectors $e_1=(1, 0)^{\mathrm{T}}$, $e_2=(0, 1)^{\mathrm{T}}$ correspond to its two eigenvalues $\lambda_1=1$ and $\lambda_2=\sum_{i=1}^{\kappa}\sigma_i^2/\kappa$.

In the PCA plus LDA framework, the goal of PCA is to choose W_{PCA} such that the matrix

$$(W_{PCA})^{\mathrm{T}} S_w W_{PCA}$$

is not singular so that the subsequent LDA procedure can be applied. That is the eigen problem for the matrix

$$((W_{PCA})^{\mathrm{T}} S_w W_{PCA})^{-1}((W_{PCA})^{\mathrm{T}} S_b W_{PCA})$$

can be solved.

Since $|\sigma_i^2| < 1, i = 1, \ldots, \kappa$, we have $\lambda_1 > \lambda_2$, the first eigenvector e_1 will be used for projection:

$$W_{PCA} = e_1.$$

But the reduced within-class scatter matrix

$$(W_{PCA})^{\mathrm{T}} S_w W_{PCA} = e_1^{\mathrm{T}} S_w e_1 = 0$$

is still singular, which leads to the *failure of LDA procedure*. We notice that in this case *there is no SSS problem*.

On the other hand, by projecting X on e_1 the first class becomes $\{1, \ldots, 1\}$ and the second class becomes $\{-1, \ldots, -1\}$. It is clear that the two classes can be well separated.

3. The improved linear discriminate analysis

In this section, we will develop a new discriminant analysis algorithm built on a new criterion. We would like to introduce our new criterion for FDA first and then we will present a new framework based on PCA and LDA and show that how our modification of PCA can be applied to our new criterion. We will prove that the use of the new criterion after our modified PCA procedure is appropriate. After that, the new algorithm will be introduced. Computational considerations related to our method will also be discussed. Applications to face recognition will be introduced at the end of this section. For convenience, in this section, we still consider the K-class problem depicted in Section 2.

3.1. Fundamentals

Before presenting our method, some remarks on the PCA plus LDA approach should be made first.

- Some small principal components that might be essential for classification are thrown away after PCA step. Since in PCA step, it just chooses d' eigenvectors corresponding to the first d' largest eigenvalues of S_t. It is very likely that the remainder contains some potential and valuable discriminatory information for the next LDA step.
- LDA might fail even after PCA dimension reduction as mentioned in the previous section.
- The null space of the within-class scatter matrix S_w contains discriminative information for classification. For a projection direct β, if $\beta^{\mathrm{T}} S_w \beta = 0$ and $\beta^{\mathrm{T}} S_b \beta \neq 0$, obviously, the optimalization problem (5) is maximized. This kind of information is ignored in the PCA plus LDA approach.

In view of the above three remarks, we will try to take into account these factors and manage to find a new approach for classification.

On the one hand, let us focus on the Fisher criterion. Since the PCA plus LDA approach would run into a dilemma when the within-class scatter matrix S'_w is singular after PCA procedure, many variants of Fisher criterion (5) have been suggested [4,26] in order to proceed the approach. For instance, by simply replacing the within-class scatter matrix S_w in the denominator by the total scatter matrix S_t, one has

$$W_{LDA}^1 = \arg\max_{W \in \mathbb{R}^{d \times d'}} \frac{|W^{\mathrm{T}} S_b W|}{|W^{\mathrm{T}} S_t W|}$$

or one can just ignore the denominator to get

$$W_{LDA}^2 = \arg\max_{W \in \mathbb{R}^{d \times d'}} |W^{\mathrm{T}} S_b W| = [w_1, w_2, \ldots, w_{d'}]$$

$$\text{s.t.} \quad w_i^{\mathrm{T}} S_w w_i = 0, \quad \|w_i\| = 1, \quad i = 1, 2, \ldots, d'$$

or

$$W_{LDA}^3 = \arg\max_{W \in \mathbb{R}^{d \times d'}} |W^{\mathrm{T}} S_b W|$$

$$= [w_1, w_2, \ldots, w_{d'}]$$

$$\text{s.t.} \quad \|w_i\| = 1, \quad i = 1, 2, \ldots, d'.$$

Each of the above three criteria has its own virtues and shortcomings. We would not compare the differences and utilization of these criteria here. Instead, we propose our new criterion.

From the inverse relationship

$$\arg\max_{W \in \mathbb{R}^{d \times d'}} \frac{|W^{\mathrm{T}} S_b W|}{|W^{\mathrm{T}} S_w W|} \iff \arg\min_{W \in \mathbb{R}^{d \times d'}} \frac{|W^{\mathrm{T}} S_w W|}{|W^{\mathrm{T}} S_b W|},$$

we deduce, without considering the singularity, *the inverse Fisher criterion*, that is

$$W_{IFDA} = \arg\min_{W \in \mathbb{R}^{d \times d'}} \frac{|W^{\mathrm{T}} S_w W|}{|W^{\mathrm{T}} S_b W|} = [w_1, w_2, \ldots, w_{d'}]. \tag{8}$$

In contrast with LDA or FDA, we name the procedure using the above inverse Fisher criterion as *the inverse Fisher discriminant analysis (IFDA)*. Obviously, the Fisher criterion (5) and inverse Fisher criterion (8) are equivalent, provided that the within-class scatter matrix S_w and the between-class scatter matrix S_b are not singular. However, we notice that the rank of the between-class scatter matrix $S_b \in \mathbb{R}^{d \times d}$ satisfies $rank(S_b) \leqslant K - 1$. Thus, the difficulty of SSS problem still exists for this new criterion.

On the other hand, let us return to exploit the PCA. For the optimization problem (2), it gives optimal projection vectors that have the largest variance and PCA just selects d' eigenvectors corresponding to the first d' largest eigenvalues of S_t but ignores the smaller ones. If we want to take those eigenvectors into account, we should abandon or modify such criterion for vector selection. Moreover, it is very likely that vectors w_i, $i = 1, 2, \ldots, d'$, satisfying the inequality

$$w_i^{\mathrm{T}} S_b w_i > w_i^{\mathrm{T}} S_w w_i,$$

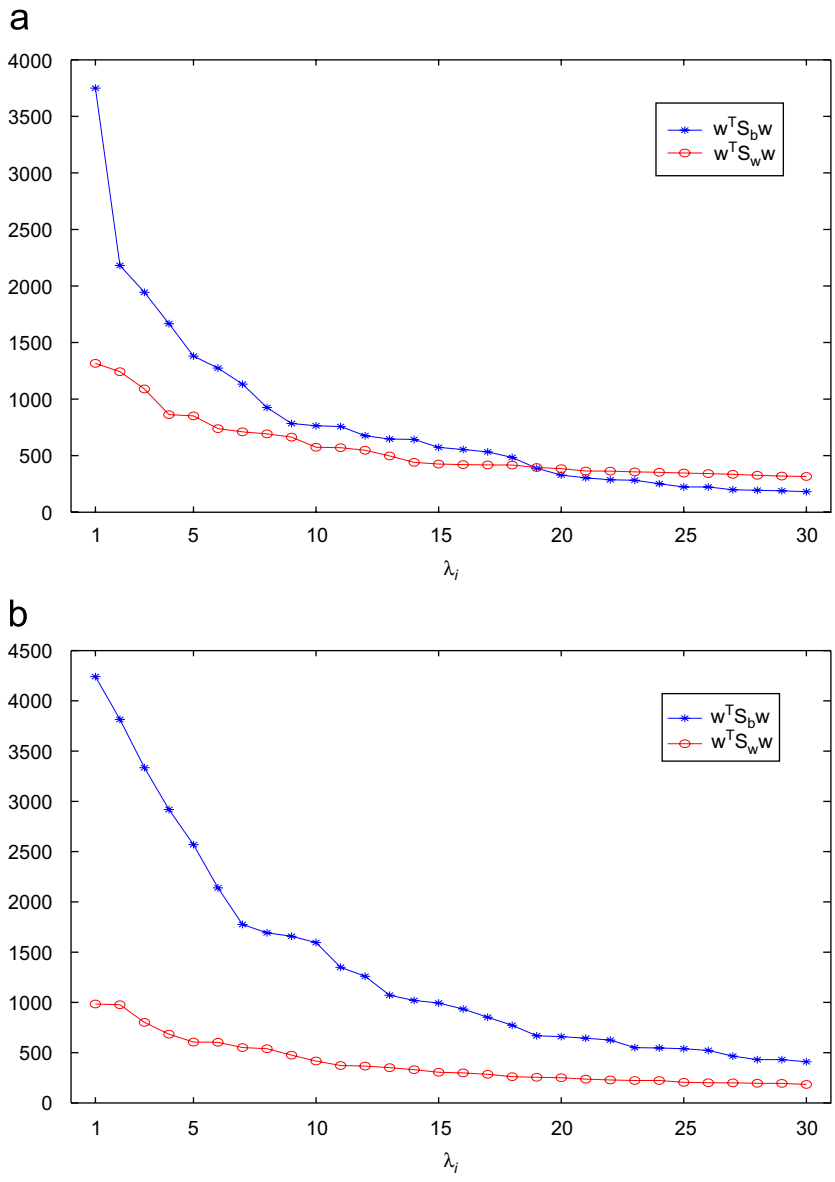

Fig. 1. Distribution of $w_i^{\mathrm{T}} S_b w_i$ and $w_i^{\mathrm{T}} S_w w_i$, w_i is the eigenvector of S_t with respect to the ith largest eigenvalue: (a) with the ORL database and (b) with the FERET database.

would be more effective for classification. In Fig. 1 the distributions of $w_i^{\mathrm{T}} S_b w_i$ and $w_i^{\mathrm{T}} S_w w_i$, where w_i is the ith largest normalized eigenvector for the ORL database, are plotted. Here we would like to present a new criterion by modifying Eq. (2) as follows:

$$W_{PCA\_S} = \arg\max_{W \in \mathbb{R}^{d \times d'}} |W^{\mathrm{T}} S_t W| = [w_1, w_2, \ldots, w_{d'}]$$

$$\text{s.t.} \quad w_i^{\mathrm{T}} S_b w_i > w_i^{\mathrm{T}} S_w w_i,$$

$$\|w_i\| = 1, \quad i = 1, 2, \ldots, d'. \tag{9}$$

we call this step *PCA with selection (PCA_S)*. The reduced matrix $S_b' = W_{PCA\_S}^{\mathrm{T}} S_b W_{PCA\_S}$ might still be singular. It is obvious that we should not work in the null space of the reduced within-covariance matrix S_b'. We further project S_b' onto its range space and denote this operation as $W_{proj} \in \mathbb{R}^{d' \times d''} (d'' \leqslant d')$.

Finally, we can introduce our new algorithm. First, we apply our modified PCA procedure to lower the dimension from d to d' and get a projection matrix $W_{PCA\_S} \in \mathbb{R}^{d \times d'}$. Moreover, we project onto the range space of the matrix S_b' and get a projection matrix $W_{proj} \in \mathbb{R}^{d' \times d''}$. Second, we use IFDA to find out the feature representation in the lower dimensionality feature space $\mathbb{R}^{d''}$ and obtain a transformation matrix W_{IFDA}. Consequently, we have the transformation matrix W_{opt} of our new approach as follows:

$$W_{opt}^{\mathrm{T}} = W_{IFDA}^{\mathrm{T}} \cdot W_{proj}^{\mathrm{T}} \cdot W_{PCA\_S}^{\mathrm{T}}, \tag{10}$$

where W_{PCA_S} is the result of optimization problem (9) and

$$W_{IFDA} = \arg\min_W \frac{|W^T W_{proj}^T W_{PCA\_S}^T S_w W_{PCA\_S} W_{proj} W|}{|W^T W_{proj}^T W_{PCA\_S}^T S_b W_{PCA\_S} W_{proj} W|}$$
$$= \arg\min_W \frac{|W^T W_{proj}^T S'_w W_{proj} W|}{|W^T W_{proj}^T S'_b W_{proj} W|}$$
$$= \arg\min_W \frac{|W^T S''_w W|}{|W^T S''_b W|}. \quad (11)$$

Before we go to the end of this part, we would like to make some comments on our new framework.

- Those eigenvectors with respect to the smaller eigenvalues of S_t are taken into account in our modified PCA step.
- Our inverse Fisher criterion can extract discriminant vectors in the null space of S_w rather than just throw them away.

3.2. The improved discriminate analysis (IDA)

In summary of the discussion so far, the algorithm for our new approach is given below:

- *Step 1* (*PCA*): For the K-class problem, the total scatter matrix $S_t \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix, and we have single value decomposition $S_t = U^T \Lambda U$, where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_g, 0, \ldots, 0)$, $g = rank(S_t)$, and $U = (u_1, u_2, \ldots, u_d)$ is a unitary matrix $U^T U = Id$ such that u_i is eigenvectors corresponding to λ_i for $i = 1, 2, \ldots, g$.
- *Step 2* (*Eigenvector selection*): Applying selection rule (for $i = 1, 2, \ldots, g$ if $u_i^T S_b u_i > u_i^T S_w u_i$ then u_i is selected) to the set of $\{u_1, u_2, \ldots, u_g\}$, we get $W_{PCA\_S} = [u_{i_1}, u_{i_2}, \ldots, u_{i_{d'}}]$, where $d' \leqslant \min\{g, K-1\}$.
- *Step 3* (*Dimension reduction*): We have the projection matrix $W_{PCA\_S} : \mathbb{R}^d \to \mathbb{R}^{d'}$. Applying it to the sampling matrix X, we calculate the between-class scatter matrix $S'_b = W_{PCA\_S}^T S_b W_{PCA\_S}$ in the reduced feature space $\mathbb{R}^{d'}$. Project onto the range of the matrix S'_b. $W_{proj} : \mathbb{R}^{d'} \to \mathbb{R}^{d''} (d'' \leqslant d')$. Calculate the between-class scatter matrix $S''_b = W_{proj}^T S'_b W_{proj}$ and the within-class scatter matrix $S''_w = W_{proj}^T W_{PCA\_S}^T S_w W_{PCA\_S} W_{proj}$ in the reduced feature space $\mathbb{R}^{d''}$. We get $Y = W_{proj}^T W_{PCA\_S}^T \cdot X = (Y_1, Y_2, \ldots, Y_N)$.
- *Step 4* (*IFDA*): The optimization problem (11) using *inverse Fisher criterion* is solved by $(S''_b)^{-1} S''_w v = \gamma v$ with eigenvalues $0 \leqslant \gamma_1 \leqslant \gamma_2 \leqslant \cdots \leqslant \gamma_q$, $q \leqslant d''$ and corresponding normalized eigenvectors $v_1, v_2, \ldots, v_q$. We have $W_{IFDA} = [v_1, v_2, \ldots, v_q]$.
- *Step 5* (*Feature representation*): Applying the inverse Fisher transform W_{IFDA} to new sample Y_i's, $i = 1, 2, \ldots, N$, we get the feature representation $Z = W_{opt}^T \cdot X = W_{IFDA}^T \cdot Y = W_{IFDA}^T \cdot W_{proj}^T \cdot W_{PCA\_S}^T \cdot X$.
- *Step 6* (*Decision*): For a new test sample $X_{new} \in \mathbb{R}^d$ and class G_j, many distance types can be applied to decide the distance $d(X_{new}, G_j)$ between X_{new} and class G_j, $j = 1, 2, \ldots, K$, such as Euclid distance l^2, l^1, cosine distance and Mahalanobis distance. The decision result is given by

$$X_{new} \in G_k = \arg\min_{G_k} d(X_{new}, G_j).$$

3.3. Computational complexity of IDA

For our algorithm, considering the complexity of eigenanalysis, there are three time-consuming steps: PCA, dimension reduction and IFDA. For PCA step, we have the method presented by Turk and Pentland [7] for eigenface problem, whose computational complexity is $\mathcal{O}(N^3)$, where N is the number of training samples. For the step of dimension reduction, Turk and Pentland's method can be applied to S'_b as well, thus, the computational complexity of this step is $\mathcal{O}(K^3)$. For IFDA step, since the rank of S''_b is up to $K-1$ and the dimension of the feature space is reduced to $d'' \leqslant K-1$, thus, similar to DLDA, the computational complexity of IFDA is also $\mathcal{O}(K^3)$. Therefore, the computational complexity of our algorithm is $\mathcal{O}(N^3 + 2K^3)$.

3.4. Application to face recognition

The IDA method can be used in pattern recognition problems with high-dimensional data, such as face recognition, gene classification [27,28], etc. When it is applied to face recognition, we call the columns of the transform W_{opt} defined in Eq. (10) the IDA face (IDAFace) and this new approach is named as IDAFace method.

4. Experiment results

In this section, experiments are designed to evaluate the performance of our new approach: IDAFace. Experiment for comparing the performance between FisherFace and IDAFace is also done.

Two standard databases from the Olivetti Research Laboratory (ORL) and the FERET are selected for evaluation. These databases could be utilized to test moderate variations in pose, illumination and facial expression. The Olivetti set contains 400 images of 40 persons. Each one has 10 images of size 92×112 with variations in pose, illumination and facial expression (see Fig. 2a). For the FERET set we use 432 images of 72 persons. Each person has six images whose resolution after cropping is also 92×112 (see Fig. 2b). Moreover, we combine the two to get a new larger set, the ORLFERET, which has 832 images of 112 persons. All the above database have been done with histogram equalization.

We implement our IDAFace algorithm and test its performance on the above three databases. In "Decision step", we use the l^2 metric as the distance measure. For the classifier we use the nearest neighbor rule with class mean of each class. The recognition rate is calculated as the ratio of number of successful recognition and the total number of test samples. The experiments are repeated 50 times on each database and average recognition rates for each database are reported.

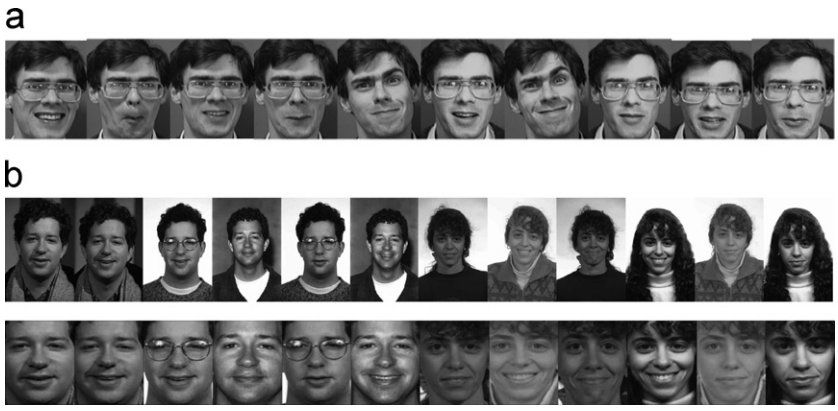

Fig. 2. (a) Example images of one subject from the ORL database; (b) example images of two subjects (the first row) and the cropped images (the second row) with the FERET database.

Table 1
Recognition rate of IDAFace method with ORL database

Number of training samples	1	2	3	4	5	6	7	8	9
Recognition rate (%)	63.3	73.9	84.3	89.7	92.4	93.9	95.5	96.0	95.9

Table 2
Recognition rate of IDAFace method with FERET database

Number of training samples	1	2	3	4	5
Recognition rate (%)	75.1	85.0	89.5	92.2	94.3

4.1. Recognition performance of the IDAFace method

We run our algorithm for the ORL database and the FERET database separately. Table 1 is the result of average recognition rate for the ORL database, and Table 2 is for the FERET database.

From Tables 1 and 2, we can see that the average recognition rates of our IDAFace method with the ORL database change from 63% to 96% when the number of training sample per class increases from 1 to 9. For the more challenging FERET database, IDAFace method has even better performance, it changes from 75% to 94% for training images from 1 to 5.

Fig. 3 shows the recognition rates from Ranks 1 to 10 for different training sample size with ORL in left and FERET in right. From Fig. 3, we can see that, when the training sample size is 5, the recognition rates of Rank 5 for both databases are nearly 99%. These results indicate the effectiveness of our new IDAFace method in real-world applications.

4.2. Comparison between IDAFace method and FisherFace method

As we know, LDA is based on an assumption that all classes are multivariate Gaussian with a common covariance matrix. For ORL database or FERET database, the assumption is reasonable since a great deal of experiments on these two database using FisherFace algorithm have substantiated the efficiency of this two-phase algorithm. However, when each class has

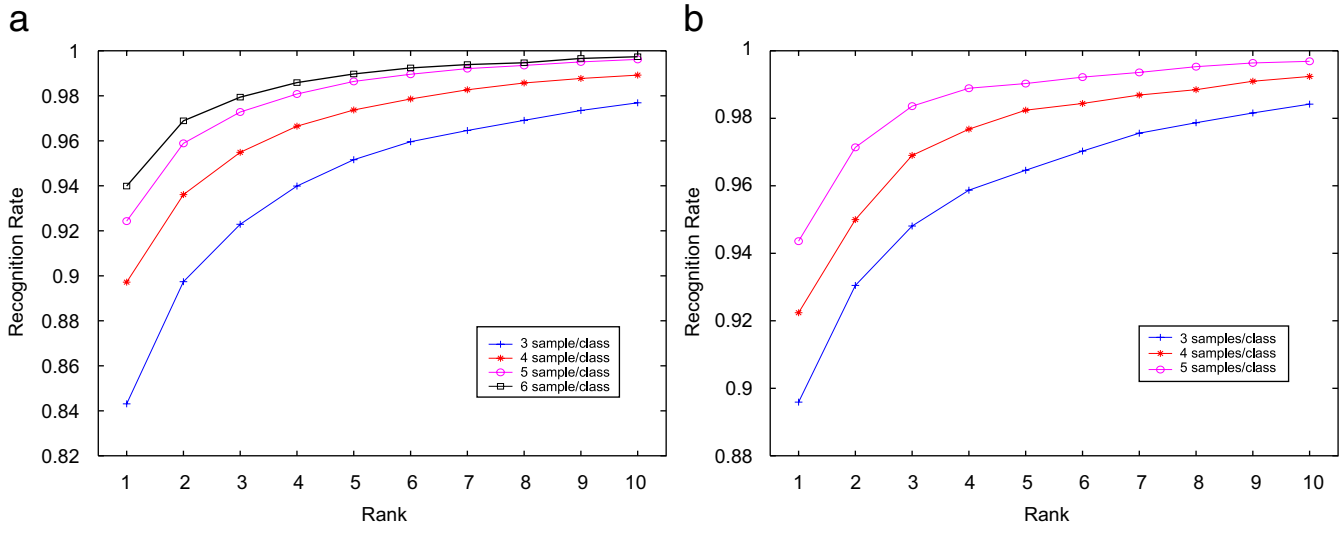

Fig. 3. Recognition rates from Ranks 1 to 10 for different training sample per class with ORL database (a) and FERET database (b).

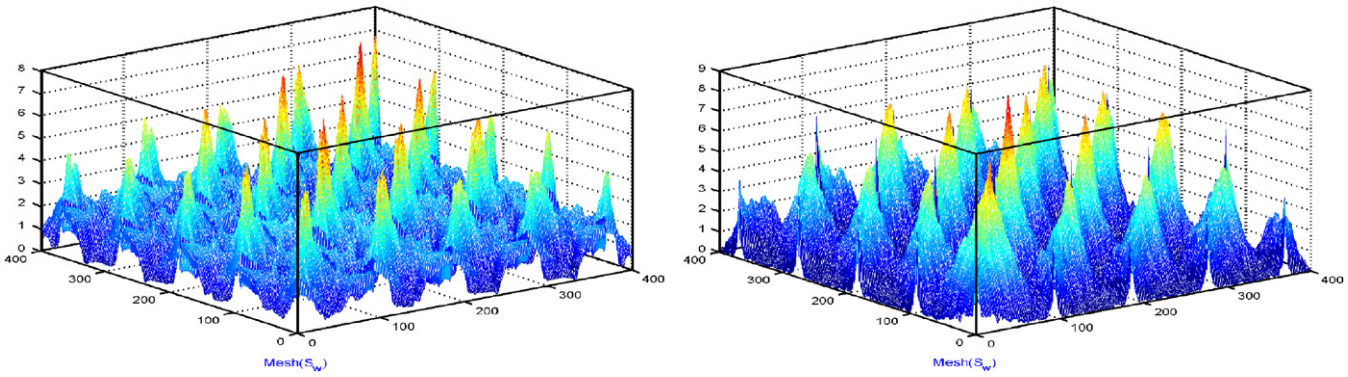

Fig. 4. The 3-D mesh surface of the within-class scatter matrix S_w with ORL database (left) and FERET database (right).

Table 3
Recognition rates of IDAFace method and FisherFace method with ORLFERET database

Number of training samples	1	2	3	4	5
Recognition rate of FisherFace (%)	N/A	74.9	83.8	86.3	87.5
Recognition rate of IDAFace (%)	66.7	75.1	85.3	89.6	92.5

different covariance matrix, this algorithm might not work very well. Fig. 4 shows parts of 3-D mesh surface of the within-class scatter matrix S_w for ORL database (left) and FERET database (right). From it we can see that the variations of the ORL database and the FERET database are different. Therefore, the combination of the two databases would result in a bigger database having different covariance matrix for different class.

In this experiment, we implement the PCA plus LDA algorithm, test the performance of our IDAFace method with the combined ORLREFRET database and compare it with the result of FisherFace method. The experiment is repeated 50 times for each number of training sample per class running through 1–5. Each time, the training samples selected from the ORLFERET database are all the same for both IDAFace method and FisherFace method. The result is shown in Table 3.

From Table 3, for our new IDAFace method, when the number of training sample per class increases from 1 to 5, the recognition rate is from 66.7% to 92.5%, while for FisherFace method, it is up to 87.5%. We notice that the IDAFace method works even when the number of training images per class is one. Moreover, from Fig. 5 we can see that IDAFace outperforms FisherFace for every number of training sample for each class, taken 5 for example, the average recognition rates are 92.5% for IDAFace, while for FisherFace it is only 87.6%. This experiment suggests that our IDAFace method can work well even in the case that the covariance matrices for different classes are not all the same.

5. Conclusion

In this paper, we proposed a new discriminant analysis framework for high-dimensional data: PCA with selection plus IFDA. Based on this framework, we present a new algorithm for recognition tasks. The algorithm applied to face recognition is implemented and experiments are also carried out to evaluate this method. Comparison is made with the PCA plus LDA approach. According to our theory and experiment results, a number of conclusions can be drawn as follows.

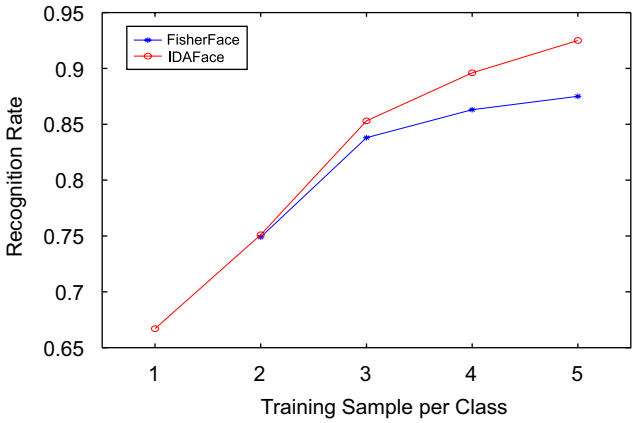

Fig. 5. Comparison between FisherFace and IDAFace.

Firstly, the projected between-class scatter matrix S_b'' in IFDA procedure will never be singular after our PCA procedure with selection, thereby guaranteeing the successful application of IFDA after PCA_S, while this is not true for PCA plus LDA approach.

Secondly, IDAFace can outperform FisherFace when dealing with the situation that not all the covariance matrices are the same for every class.

Acknowledgments

This project is supported in part by NSF of China (Grant nos: 60575004, 10231040), NSF of GuangDong, Grants from the Ministry of Education of China (Grant no.: NCET-04-0791) and Grants from Sun Yat-Sen University.

References

[1] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 252–264.

[2] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (1) (2004) 4–20.
[3] W. Zhao, R. Chellappa, P.J. Phillips, et al., Face recognition: a literature survey, ACM Comput. Surv. 35 (4) (2003) 399–459.
[4] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 711–720.
[5] A.M. Martínez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 233–288.
[6] M. Turk, A random walk through eigenspace, IEICE Trans. Inform. Syst. E84-D (12) (2001) 1586–1695.
[7] M.A. Turk, A.P. Pentland, Eigenfaces for recognition, J. Cognitive. Neurosci. 3 (1) (1991) 71–86.
[8] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Müller, Fisher discriminant analysis with kernels, Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing IX, August 1999, pp. 41–48.
[9] 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 Trans Pattern Anal. Mach. Intell. 25 (5) (2003) 623–628.
[10] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.
[11] D.Q. Dai, P.C. Yuen, Regularized discriminant analysis and its applications to face recognition, Pattern Recognition 36 (3) (2003) 845–847.
[12] D.Q. Dai, P.C. Yuen, A wavelet-based two-parameter regularization discriminant analysis for face recognition, Proceeding of the Fourth International Conference on Audio and Video Based Personal Authentication, Lecture Notes in Computer Science, vol. 2688 (2003) 137–144.
[13] I. Pima, M. Aladjem, Regularized discriminant analysis for face recognition, Pattern Recognition 37 (9) (2004) 1945–1948.
[14] C.J. Liu, H. Wechsler, A shape- and texture-based enhanced Fisher classifier for face recognition, IEEE Trans. Image Process. 10 (4) (2001) 598–608.
[15] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.
[16] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.
[17] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space?, Pattern Recognition 36 (2) (2003) 563–566.
[18] 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 Recognition Artif. Intell. 6 (5) (1992) 817–829.
[19] L.F. Chen, H.Y.M. Liao, J.C. Lin, M.D. Kao, G.J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.
[20] V. Perlibakas, Face recognition using principal component analysis and wavelet packet decomposition, Informatica 15 (2) (2004) 243–250.
[21] J. Yang, J.Y. Yang, Optimal FLD algorithm for facial feature extraction, Proceedings of SPIE on Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision, October 2001, pp. 438–444.
[22] J.P. Ye, R. Janardan, C.H. Park, H. Park, An optimization criterion for generalized discriminant analysis on undersampled problems, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 982–994.
[23] B. Zhang, H. Zhang, S. Sam Ge, Face recognition by applying wavelet subband representation and kernel associative memory, IEEE Trans. Neural Networks 15 (1) (2004) 166–177.
[24] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 230–244.
[25] X.S. Zhuang, D.Q. Dai, Inverse Fisher discriminate criteria for small sample size problem and its application to face recognition, Pattern Recognition 38 (11) (2005) 2192–2194.
[26] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1990.
[27] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (2002) 77–87.
[28] J.W. Lee, J.B. Lee, M. Park, S.H. Song, An extensive comparison of recent classification tools applied to microarray data, Comput. Stat. Data Anal. 48 (2005) 869–885.

About the Author—XIAO-SHENG ZHUANG received his bachelor degree and master degree in mathematics from Sun Yat-Sen (Zhongshan) University, China, in 2003 and 2005, respectively. He is now in University of Alberta, Canada, working for his Ph.D. degree. His current research interest includes wavelet analysis, face recognition and detection.

About the Author—DAO-QING DAI received his Ph.D. in mathematics from Wuhan University, China, in 1990. He is currently a professor and associate dean of the Faculty of Mathematics and Computing, Sun Yat-Sen (Zhongshan) University. He visited Free University, Berlin, as an Alexander von Humboldt research fellow from 1998 to 1999. He got the "outstanding research achievements in mathematics" award from ISAAC (International Society for Analysis, Applications and Computation) in 1999 at Fukuoka, Japan. He served as programm chair of Sinobiometrics'2004 and programm committee members for several international conferences. His current research interest includes image processing, wavelet analysis and human face recognition.