A Multiple Maximum Scatter Difference Discriminant Criterion for Facial Feature Extraction

Fengxi Song, David Zhang, *Senior Member, IEEE*, Dayong Mei, and Zhongwei Guo

***Abstract*—Maximum scatter difference (MSD) discriminant criterion was a recently presented binary discriminant criterion for pattern classification that utilizes the generalized scatter difference rather than the generalized Rayleigh quotient as a class separability measure, thereby avoiding the singularity problem when addressing small-sample-size problems. MSD classifiers based on this criterion have been quite effective on face-recognition tasks, but as they are binary classifiers, they are not as efficient on large-scale classification tasks. To address the problem, this paper generalizes the classification-oriented binary criterion to its multiple counterpart—multiple MSD (MMSD) discriminant criterion for facial feature extraction. The MMSD feature-extraction method, which is based on this novel discriminant criterion, is a new subspace-based feature-extraction method. Unlike most other subspace-based feature-extraction methods, the MMSD computes its discriminant vectors from both the range of the between-class scatter matrix and the null space of the within-class scatter matrix. The MMSD is theoretically elegant and easy to calculate. Extensive experimental studies conducted on the benchmark database, FERET, show that the MMSD outperforms state-of-the-art facial feature-extraction methods such as null space method, direct linear discriminant analysis (LDA), eigenface, Fisherface, and complete LDA.**

***Index Terms*—Face recognition, feature extraction, linear discriminant criterion.**

I. Introduction

THE MAXIMUM scatter difference (MSD) discriminant criterion [1] was a recently presented binary discriminant criterion for pattern classification. Because the MSD utilizes the generalized scatter difference rather than the generalized Rayleigh quotient as a class separability measure, it avoids the singularity problem when addressing the small-sample-size problems that trouble the Fisher discriminant criterion. Furthermore, studies have demonstrated that MSD classifiers that are based on this discriminant criterion have been quite effective on face-recognition tasks. The drawback of the MSD classifier is that, as a binary classifier, it cannot be applied directly to multiple-category classification tasks such as face recognition. This means that multiple-category classification tasks have to be divided into a series of binary classification problems using one of the three decomposition strategies: one versus rest, one versus one, or directed acyclic graph [2]. Experiments have shown that MSD classifiers are not very effective when using the first strategy, while using the latter two strategies requires the training of $l(l-1)/2$ MSD classifiers for an l-category classification task. The efficiency of such an approach will greatly be affected by any increase in the number of categories. Ultimately, then, like all binary classifiers, MSD classifiers are not suitable for large-scale pattern recognition problems.

To address the problem, this paper generalizes the classification-oriented binary criterion to its multiple counterpart—multiple MSD (MMSD) discriminant criterion for facial feature extraction. The MMSD feature-extraction method, which is based on this novel discriminant criterion, is a new subspace-based feature-extraction method. Unlike most conventional subspace-based feature-extraction methods that derive their discriminant vectors either in the range of the between-class scatter matrix or in the null space of the within-class scatter matrix, the MMSD computes its discriminant vectors in both subspaces. The MMSD is theoretically elegant and easy to calculate. Extensive experimental studies conducted on the benchmark database, FERET, show that the MMSD outperforms many state-of-the-art facial feature-extraction methods, including the null space method (NSM), direct linear discriminant analysis (D-LDA), eigenface, Fisherface, and complete LDA (C-LDA).

The remainder of this paper is organized as follows: In Section II, we describe the MMSD discriminant criterion and the MMSD-based facial feature-extraction algorithm. In Section III, we present some theoretical analyses of MMSD. In Section IV, we compare the performance of MMSD with that of several current facial feature-extraction approaches on the FERET database. Section V offers a brief conclusion.

II. Discriminant Criterion and Feature-Extraction Algorithm Based on MMSD

We first review the MSD discriminant criterion in Section II-A, then describe the MMSD discriminant criterion in Section II-B, and, finally, present the feature-extraction algorithm based on the MMSD in Section II-C.

Manuscript received November 15, 2006; revised February 9, 2007 and March 20, 2007. This work was supported in part by the National Science Foundation of China under Grant 60620160097. This paper was recommended by Associate Editor X. Wang.

F. Song is with the New Star Research Institute of Applied Technology, Hefei 230031, China, and also with Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: songfengxi@yahoo.com).

D. Zhang is with Hong Kong Polytechnic University, Kowloon, Hong Kong.

D. Mei is with the Academy of Military Science, Beijing 100080, China.

Z. Guo is with the New Star Research Institute of Applied Technology, Hefei 230031, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2007.906579

1083-4419/$25.00 © 2007 IEEE

A. MSD Discriminant Criterion

The optimization model corresponding to the MSD discriminant criterion [1] is as follows:

$$\max_{\mathbf{w}\neq 0} \frac{\mathbf{w}^{\mathrm{T}}(S_{\mathrm{b}} - c\cdot S_{\mathrm{w}})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}} \tag{1}$$

which is derived from the following multiobjective programming problem

$$\max \frac{\mathbf{w}^{\mathrm{T}}S_{\mathrm{b}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$$
$$\min \frac{\mathbf{w}^{\mathrm{T}}S_{\mathrm{w}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}} \tag{2}$$

by combining the two objectives using additive principle [3]. The parameter c in (1) is a nonnegative constant which balances the relative merits of maximizing the between-class scatter to the minimization of the within-class scatter, the between-class scatter matrix S_{b}, and the within-class scatter matrix S_{w}, both of which are defined as in [1].

It has been proven that the optimal projection direction determined by the MSD is the eigenvector of the matrix $(S_{\mathrm{b}} - c\cdot S_{\mathrm{w}})$ corresponding to the largest eigenvalue. Let $\mathbf{w}^*$ be the optimal projection direction determined by the MSD. The binary linear classifier based on the MSD is defined as follows:

$$f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{*\mathrm{T}}\mathbf{x} + w_0)\cdot \mathrm{sign}(\theta + w_0) \tag{3}$$

which assigns a label $\mathrm{sign}(\mathbf{w}^{*\mathrm{T}}\mathbf{x} + w_0)\cdot \mathrm{sign}(\theta + w_0)$ to a sample $\mathbf{x}$. Here, sign is the sign function, $f(\mathbf{x}) = 1$ means that $\mathbf{x}$ belongs to the first class, and $f(\mathbf{x}) = -1$ means that $\mathbf{x}$ belongs to the second class. The bias w_0 can simply be computed by

$$w_0 = -\mathbf{w}^{*\mathrm{T}}\mathbf{m} \tag{4}$$

where $\mathbf{m}$ is the mean training sample.

Theoretical analysis demonstrates that the MSD classifier has closed relations with other binary classifiers. For example, if the within-class scatter matrix S_{w} is singular, the asymptotic form of an MSD classifier is the large margin linear projection classifier [4] when the parameter c approaches infinity. In addition, if the matrix S_{w} is nonsingular, the MSD classifier is the Fisher classifier [5] when the parameter c is the unique solution to the following equation:

$$\max_{\mathbf{w}\neq 0} \frac{\mathbf{w}^{\mathrm{T}}(S_{\mathrm{b}} - c\cdot S_{\mathrm{w}})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}} = 0. \tag{5}$$

The experimental studies demonstrated that, as a binary classifier, the MSD is quite effective on small-sample-size problems such as appearance-based face recognition. Furthermore, its performance is quite robust on the parameter c. However, like all binary classifiers, the efficiency of the MSD declines greatly with the increase of the number of classes. Therefore, there is a need to extend this classification-oriented binary criterion to its multiple counterpart—MMSD discriminant criterion for feature extraction in large-scale recognition tasks.

B. MMSD Discriminant Criterion

The goal of discriminant criteria for feature extraction is to seek r discriminant vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ such that training samples from a high-dimensional input space are farthest apart after they are projected on these vectors. In order to eliminate the influences of the lengths of the discriminant vectors and linear dependences between these vectors, we usually require them to be orthonormal. That is, the discriminant vectors should satisfy the constraints $\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = \delta_{ij}, i, j = 1, 2, \ldots, r$.

Let $W = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r] \in \boldsymbol{R}^{d\times r}$ be the discriminant matrix. The projection of a sample $\mathbf{x}$ on the discriminant vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ is $W^{\mathrm{T}}\mathbf{x}$. The between- and within-class scatter matrices of the projected training samples $W^{\mathrm{T}}\mathbf{x}_1, W^{\mathrm{T}}\mathbf{x}_2, \ldots, W^{\mathrm{T}}\mathbf{x}_N$ are $\widetilde{S}_{\mathrm{B}} = W^{\mathrm{T}}S_{\mathrm{B}}W$and $\widetilde{S}_{\mathrm{W}} = W^{\mathrm{T}}S_{\mathrm{W}}W$, respectively.

The trace of the between-class scatter matrix of the projected training samples $\mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{B}}W)$ reflects the scatter of the projected training samples between categories, whereas the trace of the within-class scatter matrix of the projected training samples $\mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{W}}W)$ reflects the scatter of the projected training samples within each category. The larger the $\mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{B}}W)$, the more separable the projected data; the smaller the $\mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{W}}W)$, the more separable the projected data. Thus, we wish to achieve two distinct objectives

$$\max \mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{B}}W) \tag{6}$$
$$\min \mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{W}}W) \tag{7}$$

satisfying the orthonormal constraints

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = \delta_{ij}, \qquad i, j = 1, 2, \ldots, r. \tag{8}$$

The problem of seeking a set of discriminant vectors is then translated into a problem of solving a multiobjective programming model which is defined by (6)–(8). It is well known that a multiobjective programming model cannot be solved directly. It has to be converted first into a single-objective programming model. There are two main ways to convert a multiobjective programming model into a single-objective model: The first is the goal programming approach in which one of the objectives is optimized while the remaining objectives are converted into constraints. The second is the combining objective approach in which all objectives are combined into one scalar objective [3]. When using a combining objective approach, there are two major rules: the multiplicative and the additive. By utilizing the multiplicative rule to combine objectives (6) and (7), we gain a discriminant criterion that is similar to the generalized Fisher discriminant criterion [6]. If we use the additive rule to combine the two objectives, we obtain the MMSD discriminant criterion. The single-objective optimization model corresponding to the MMSD discriminant criterion is as follows:

$$\max_{\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j=\delta_{ij}, i,j=1,2,\ldots,r} \mathrm{tr}\left\{[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r]^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}) \times [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r]\right\}. \tag{9}$$

It can further be transformed into

$$\max_{\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j=\delta_{ij},\, i,j=1,2,\ldots,r} \sum_{i=1}^{r} \mathbf{w}_i^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})\mathbf{w}_i. \tag{10}$$

Here, $S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}$ is a generalized scatter difference matrix, and c is a nonnegative parameter. Obviously, (10) is an extension of (1).

Based on [6, Th. 4], we can conclude that the orthonormal eigenvectors $\varphi_1, \varphi_2, \ldots, \varphi_r$ of the matrix $S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}$, corresponding to the r largest eigenvalues, make an optimal solution of (10). We call these orthonormal eigenvectors as MMSD discriminant vectors.

It should be pointed out that a similar discriminant criterion—Differential Scatter Discriminant Criterion has been investigated in [7] and extended to General Tensor Discriminant Analysis [8] and Tensor Minimax Probability Machines [9].

C. MMSD-Based Feature-Extraction Algorithm

It is expensive, both in time and memory, to directly perform eigendecomposition on the matrix $S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}$ when the dimensionality of the input space d is large enough. It is the key of the MMSD-based feature-extraction algorithm how to quickly decompose the matrix $S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}$. Fortunately, however, with the following lemma and theorem, we can always compute the eigenvectors of $S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}$ by performing eigendecomposition on a much smaller matrix whose dimension is equal to or less than $(N-1)\times(N-1)$.

Lemma 1: Suppose S_{B} and S_{W} to be the between- and within-class scatter matrices, respectively. Let P be the matrix of all unit eigenvectors of the total scatter matrix $S_{\mathrm{T}}(= S_{\mathrm{B}} + S_{\mathrm{W}})$ corresponding to nonzero eigenvalues. Then, the following conditions follow.

1) $PP^{\mathrm{T}}S_{\mathrm{B}} = S_{\mathrm{B}}$.
2) $PP^{\mathrm{T}}S_{\mathrm{W}} = S_{\mathrm{W}}$.

Proof: Let $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_s]$ be the matrix of all unit eigenvectors of the total scatter matrix corresponding to zero eigenvalues. It is obvious that $V = [P\ Q]$ is a unitary matrix.

Since $\mathbf{q}_i (i = 1, 2, \ldots, s)$ is an eigenvector of S_{T} corresponding to zero eigenvalue, it follows that $S_{\mathrm{T}}\mathbf{q}_i = \mathbf{0}$. Thus, we have $\mathbf{q}_i^{\mathrm{T}}S_{\mathrm{B}}\mathbf{q}_i + \mathbf{q}_i^{\mathrm{T}}S_{\mathrm{W}}\mathbf{q}_i = \mathbf{0}$. Considering the fact that S_{B} and S_{W} are both semipositive matrices, it can be concluded that $\mathbf{q}_i^{\mathrm{T}}S_{\mathrm{B}}\mathbf{q}_i = \mathbf{0}$ and $\mathbf{q}_i^{\mathrm{T}}S_{\mathrm{W}}\mathbf{q}_i = \mathbf{0}$. By using [10, Th. 2], it follows that $\mathbf{q}_i^{\mathrm{T}}S_{\mathrm{B}} = \mathbf{0}$ and $\mathbf{q}_i^{\mathrm{T}}S_{\mathrm{W}} = \mathbf{0}$. As a consequence, we have

$$\begin{aligned} S_{\mathrm{B}} &= VV^{\mathrm{T}}S_{\mathrm{B}} \\ &= PP^{\mathrm{T}}S_{\mathrm{B}} + QQ^{\mathrm{T}}S_{\mathrm{B}} \\ &= PP^{\mathrm{T}}S_{\mathrm{B}} + Q\begin{pmatrix} \mathbf{q}_1^{\mathrm{T}}S_{\mathrm{B}} \\ \vdots \\ \mathbf{q}_s^{\mathrm{T}}S_{\mathrm{B}} \end{pmatrix} \\ &= PP^{\mathrm{T}}S_{\mathrm{B}} \end{aligned}$$

and

$$\begin{aligned} S_{\mathrm{W}} &= VV^{\mathrm{T}}S_{\mathrm{W}} \\ &= PP^{\mathrm{T}}S_{\mathrm{W}} + QQ^{\mathrm{T}}S_{\mathrm{W}} \\ &= PP^{\mathrm{T}}S_{\mathrm{W}} + Q\begin{pmatrix} \mathbf{q}_1^{\mathrm{T}}S_{\mathrm{W}} \\ \vdots \\ \mathbf{q}_s^{\mathrm{T}}S_{\mathrm{W}} \end{pmatrix} \\ &= PP^{\mathrm{T}}S_{\mathrm{W}}. \end{aligned}$$

Theorem 1: Suppose $P \in R^{d\times t}$ to be the matrix of all unit eigenvectors of the total scatter matrix S_{T} corresponding to nonzero eigenvalues and $\varphi \in R^{t\times 1}$ to be the eigenvector of the matrix $P^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P$ corresponding to the eigenvalue λ. Then, $P\varphi$ is the eigenvector of the matrix $S_{\mathrm{B}} - c\cdot S_{\mathrm{W}}$ corresponding to the eigenvalue λ.

Proof: Since φ is the eigenvector of the matrix $P^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P$ corresponding to the eigenvalue λ, we have

$$\begin{aligned} P^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P\varphi = \lambda\varphi \Rightarrow PP^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P\varphi \\ = \lambda P\varphi \Rightarrow (PP^{\mathrm{T}}S_{\mathrm{B}} - c\cdot PP^{\mathrm{T}}S_{\mathrm{W}})P\varphi \\ = \lambda P\varphi. \end{aligned}$$

From Lemma 1, it follows that $(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P\varphi = \lambda P\varphi$.

Thus, we complete the proof of the theorem.

According to Theorem 1, we can calculate the MMSD discriminant vectors in a high-dimensional input space in three major steps: First, we calculate the matrix P of all unit eigenvectors of the total scatter matrix corresponding to nonzero eigenvalues using singular value decomposition theorem as in [11]; second, we map the high-dimensional input space into the range of the total scatter matrix using $P^{\mathrm{T}} : R^d \to R^t$, $\mathbf{x} \mapsto P^{\mathrm{T}}\mathbf{x}$; third, we perform eigendecomposition on the matrix $P^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P$.

Algorithm 1 is a detailed description of the novel facial feature-extraction algorithm based on MMSD.

Algorithm 1: Facial feature-extraction algorithm based on MMSD

Input: Training samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, class labels of these samples $l(\mathbf{x}_1), l(\mathbf{x}_2), \ldots, l(\mathbf{x}_N)$, parameter value c, and the number of extracted features r

Output: The discriminant matrix of MMSD V

1) Compute the between-class scatter matrix S_{B}, the with-class scatter matrix S_{W}, and the total scatter matrix S_{T}.
2) Calculate the matrix P of all unit eigenvectors of S_{T} corresponding to nonzero eigenvalues using the singular value decomposition theorem as in [11].
3) Work out the matrix U of the first r unit eigenvectors of $P^{\mathrm{T}}(S_{\mathrm{B}} - c\cdot S_{\mathrm{W}})P$ corresponding to the largest eigenvalues.
4) Compute the discriminant matrix of MMSD using the formula $V = PU$.

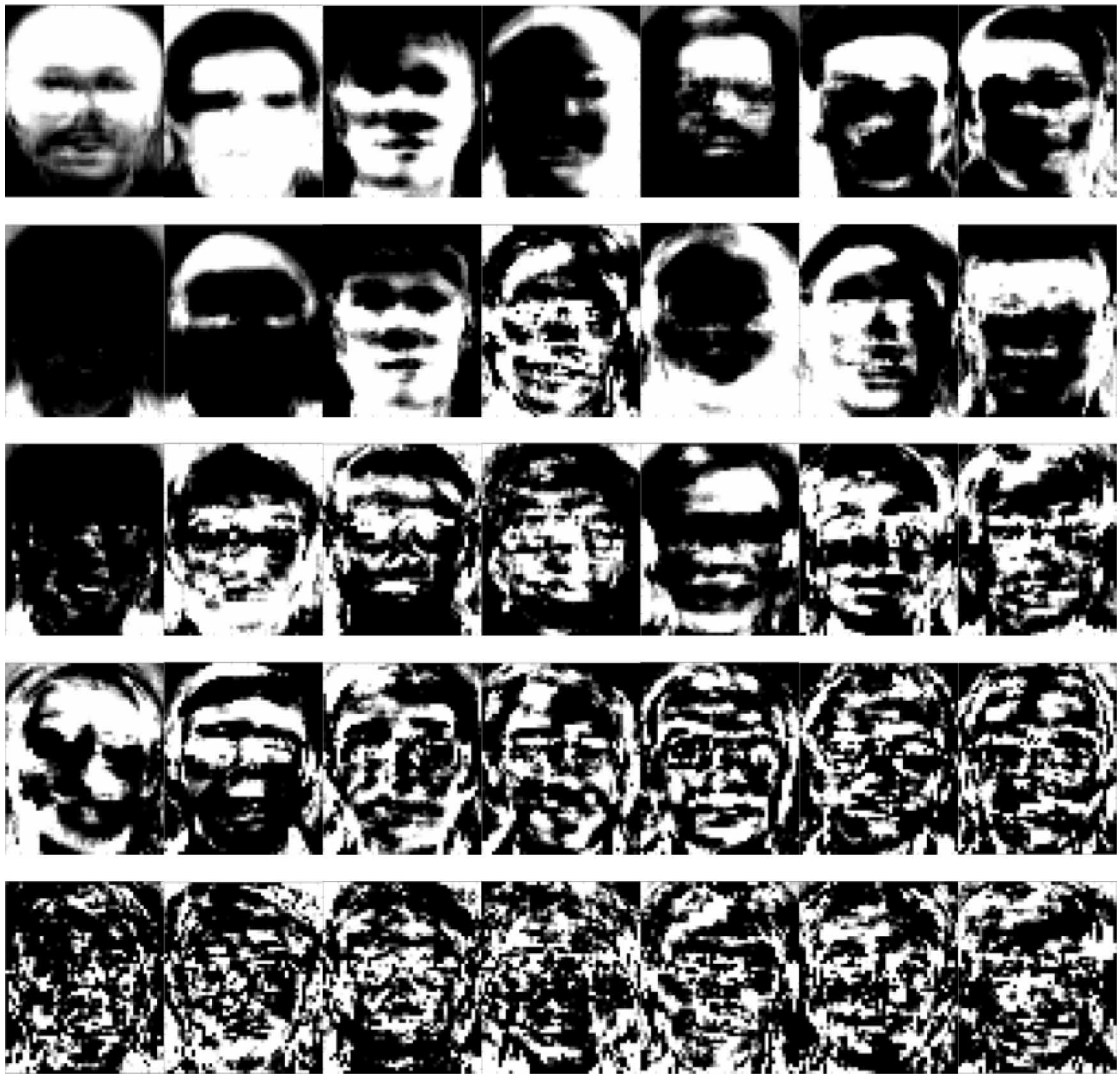


Fig. 1. First seven eigenfaces of MMSD when the parameter c assumes the values of -1, 1, 10, 100, and 1000.

III. Theoretical Analyses of MMSD Feature-Extraction Method

To further investigate the MMSD feature-extraction method, we try to reveal its relations to other feature-extraction approaches. Section III-A reveals the relation between the MMSD and principal component analysis (PCA) [12]. Section III-B reveals the relation between the MMSD and Karhunen–Loève (K–L) expansion. Section III-C reveals the relation between the MMSD and NSM [13]. In addition, Section III-D discusses the physical meaning of the parameter of MMSD.

A. Connection to PCA

Although MMSD is a supervised feature-extraction method, it is closely related to a well-known unsupervised feature-extraction approach PCA [12]. When the parameter c assumes the value of -1, the generalized scatter difference matrix is $S_{\mathrm{B}} - (-1) \cdot S_{\mathrm{W}} = S_{\mathrm{B}} + S_{\mathrm{W}} = S_{\mathrm{T}}$, i.e., the total scatter matrix. This implies that, when $c = -1$, the MMSD discriminant vectors are, in fact, principal component directions.

In facial feature extraction, we can obtain eigenfaces of MMSD by reverting the MMSD discriminant vectors to images. Fig. 1 displays the first seven eigenfaces of the MMSD when the parameter c assumes the values of -1, 1, 10, 100, and 1000. These eigenfaces are calculated on the training set, which consists of the first five images of each individual from the Olivetti Research Laboratory (ORL) face-image database [14]. As shown earlier, the eigenfaces of MMSD are actually Eigenfaces [15] when the parameter c assumes the value of -1, and as shown in Fig. 1, the details in the MMSD eigenfaces increase with the value of the parameter c.

B. Connection to K–L Expansion Based on the Between-Class Scatter Matrix

When the parameter c assumes the value of zero, MMSD degenerates into a feature-extraction method which derives its discriminant vectors in the range of the between-class scatter matrix. In fact, the MMSD is equivalent to the K–L expansion whose generation matrix is the between-class scatter matrix S_{B}. Here, the discriminant vectors of MMSD are orthonormal eigenvectors of the matrix S_{B} corresponding to the r largest eigenvalues. It implies that the discriminant vectors of MMSD are solely dependent on the objective (6) in this case.

C. Asymptotic Property of MMSD

The asymptotic property of MMSD is revealed by the following theorem.

Theorem 2: If S_{W} is singular, the discriminant vectors of MMSD are approaching the discriminant vectors of NSM when the value of the parameter c is approaching infinity.

Proof: We only prove that the first discriminant vector of MMSD is approaching the first discriminant vector of NSM when the value of the parameter c is approaching infinity. Other proofs are similar.

Let $\mathbf{w}_{\mathrm{b}}$ and $\mathbf{w}_c$ be the unit eigenvectors of matrices S_{B}, $(S_{\mathrm{B}} - c \cdot S_{\mathrm{W}})$ corresponding to the largest eigenvalues λ_{b}, λ_c, respectively.

Since S_{w} is a singular matrix, there exists a nonzero unit vector $\mathbf{w}_0$ such that $S_{\mathrm{W}}\mathbf{w}_0 = 0$. Considering the fact that S_{B} is a semipositive matrix, we have

$$\begin{aligned}\lambda_c &= \max_{\|\mathbf{w}\|=1} \mathbf{w}^{\mathrm{T}}(S_{\mathrm{b}} - c \cdot S_{\mathrm{w}})\mathbf{w} \\ &\geq \mathbf{w}_0^{\mathrm{T}} S_{\mathrm{b}} \mathbf{w}_0 - c \cdot \mathbf{w}_0^{\mathrm{T}} S_{\mathrm{w}} \mathbf{w}_0 \\ &= \mathbf{w}_0^{\mathrm{T}} S_{\mathrm{b}} \mathbf{w}_0 \geq 0. \end{aligned} \tag{11}$$

From the meaning of λ_c and $\mathbf{w}_c$, the following equation is always true for any positive real number c:

$$(S_{\mathrm{B}} - c \cdot S_{\mathrm{W}})\mathbf{w}_c = \lambda_c \mathbf{w}_c. \tag{12}$$

By combining equality (12), inequality (11), and the meaning of λ_{b}, we obtain

$$\begin{aligned}\mathbf{w}_c^{\mathrm{T}} S_{\mathrm{W}} \mathbf{w}_c &= \frac{1}{c}\left(\mathbf{w}_c^{\mathrm{T}} S_{\mathrm{B}} \mathbf{w}_c - \lambda_c\right) \\ &\leq \frac{1}{c}\mathbf{w}_c^{\mathrm{T}} S_{\mathrm{B}} \mathbf{w}_c \\ &\leq \frac{1}{c}\lambda_{\mathrm{b}}. \end{aligned} \tag{13}$$

Since the matrix S_{W} is also semipositive, the following inequality holds:

$$\mathbf{w}_c^{\mathrm{T}} S_{\mathrm{W}} \mathbf{w}_c \geq 0. \tag{14}$$

By combining inequalities (13) and (14), we can conclude that

$$\lim_{c \to \infty} \mathbf{w}_c^{\mathrm{T}} S_{\mathrm{W}} \mathbf{w}_c = 0. \tag{15}$$

Thus, we complete the proof of the theorem.

Fig. 2. Sample images from the subset of the FERET.

From the theorem, it is easy to understand that NSM is, in fact, an asymptotic form of MMSD.

D. Physical Meaning of the Parameter

From the discussions in Sections III-B and III-C, we find that, like C-LDA [16], MMSD derives its discriminant vectors both in the range of the between-class scatter matrix and in the null space of the within-class scatter matrix. However, MMSD is much more flexible than C-LDA. The parameter c can be used to adjust the balance between the two subspaces. When $c = 0$, the discriminant vectors of the MMSD are solely from the range of the between-class scatter matrix. With the increase of the value of c from zero to infinite, the discriminant vectors of MMSD are more and more from the null space of the within-class scatter matrix. When the parameter c is approaching infinity, the discriminant vectors of MMSD are solely from the null space of the within-class scatter matrix.

IV. Experimental Results

The proposed facial feature-extraction method was mainly evaluated on the benchmark face-image database FERET. The FERET face-image database is a result of the FERET program, which was sponsored by the U.S. Department of Defense through the DARPA program [17], [18]. It has become a standard database for the evaluation of state-of-the-art face-recognition techniques.

MMSD was evaluated on a subset of the FERET database. This subset includes 1400 images of 200 individuals with seven images of each individual and is composed of the images whose names are marked with two character strings: “ba,” “bj,” “bk,” “be,” “bf,” “bd,” and “bg.” The facial portion of each original image was automatically cropped based on the location of the eyes, and the cropped image was resized to 80 × 80 pixels and preprocessed by histogram equalization, as in [19]. Sample images from the subset are shown in Fig. 2.

In all the experiments, we used the nearest neighbor (NN) classifier with Euclidean distance.

A. Effectiveness of MMSD Over the Number of Extracted Features and the Value of the Parameter

Three images of each individual were randomly chosen for training, while the remaining four images were used for testing. Thus, the training sample set size was 600, and the testing sample set size was 800. In this way, we ran the system ten times and obtained ten different training and testing sample sets.

Fig. 3 demonstrates the average recognition rates of MMSD over various numbers of extracted features and different values of the parameter c.

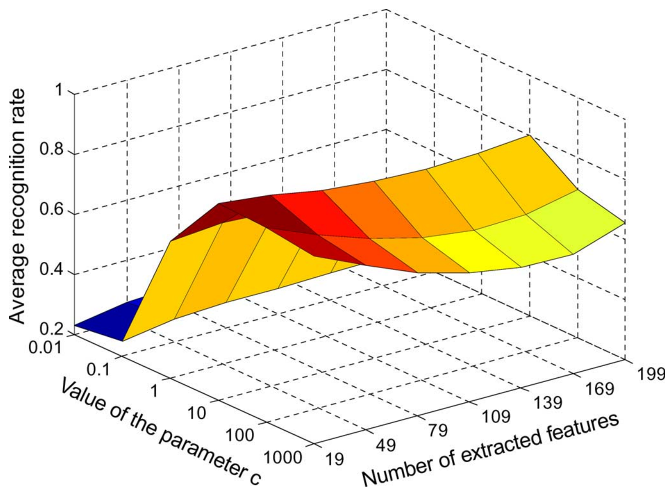


Fig. 3. Average recognition rate of MMSD over the number of extracted features and the value of the parameter c.

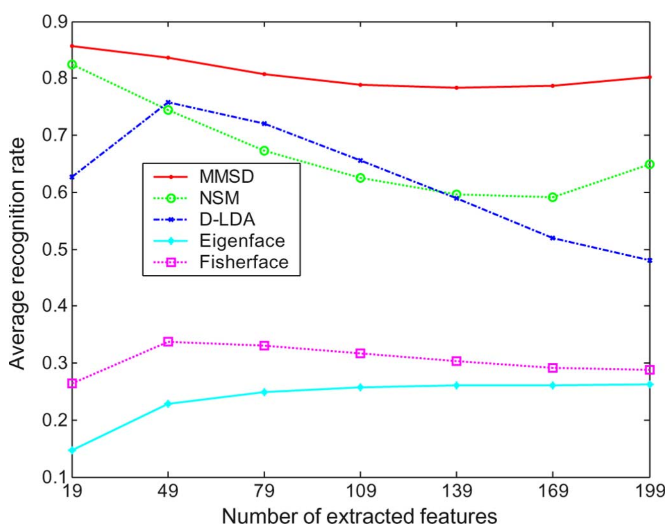


Fig. 4. Average recognition rate of various methods versus the number of extracted features.

In Fig. 3, we find two facts: First, the effectiveness of the MMSD is quite robust on the number of extracted features; second, the effectiveness of the MMSD is sensitive to the value of the parameter c.

B. Comparison of the Effectiveness of MMSD With the Effectiveness of State-of-the-Art Facial Feature-Extraction Methods With Varying Number of Extracted Features

The experimental design is the same as in Section IV-A. To make NSM applicable, the PCA is first used to compress a high-dimensional image space into the range of the total scatter matrix. Fig. 4 displays curves of the average recognition rates of MMSD ($c = 10$), NSM, D-LDA, eigenface, and Fisherface [20].

In Fig. 4, we find that the MMSD is much more effective than the other four facial feature-extraction methods, and Fisherface achieves its maximum value at 49.

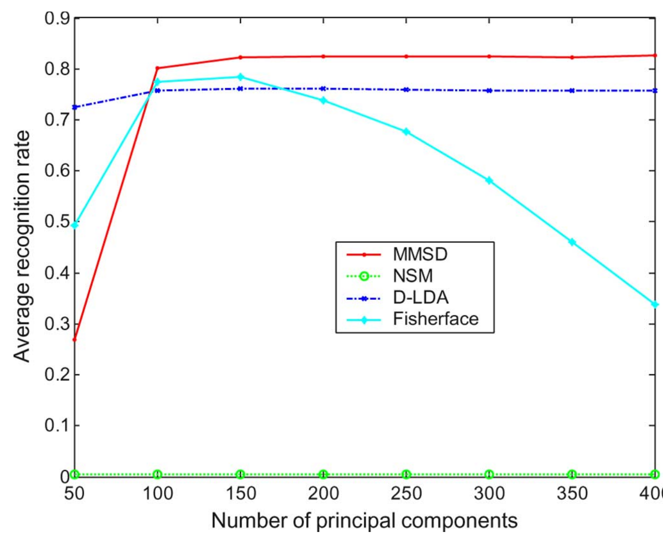


Fig. 5. Average recognition rate of various methods versus the number of principal components.

C. Further Comparison of MMSD With Fisherface and Other Methods

According to the study in [19], the effectiveness of Fisherface is heavily dependent on the number of principal components used in the PCA stage. In this section, we compare the effectiveness of the MMSD with that of NSM, D-LDA, and Fisherface when the number of principal components used in the PCA stage varies from 50 to 400. The experimental design is the same as in Section IV-A. The number of extracted features for each method is 49. The value of the parameter c of MMSD is ten.

Fig. 5 displays curves of the average recognition rates of various feature-extraction methods over varying number of principal components. We can see that the MMSD outperforms the other three feature-extraction methods when there are more than 100 principal components.

D. Comparison of the Effectiveness of MMSD With the Effectiveness of State-of-the-Art Feature-Extraction Methods Over Varying Number of Training Samples per Individual

According to the study in [21], the effectiveness of D-LDA is heavily dependent on the number of training samples per individual. In this section, we compare the effectiveness of MMSD with that of NSM, D-LDA, eigenface, Fisherface, and C-LDA when the number of training samples per individual varies from two to six. The experiment consisted of five tests of seven runs each. In each run of the ith test, $(i + 1)$ images of each individual were used for training, and the remaining $(6 - i)$ images were used for testing. Images of each individual numbered 1 to $(i + 1)$, 2 to $(i + 2), \ldots, 7$ to i were used as training samples in the first, second,..., seventh run, respectively. The numbers of the extracted features for MMSD, NSM, D-LDA, Fisherface, eigenface, and C-LDA in the ith test are 199, 199, 199, 199, $200i + 199$, and 398, respectively.

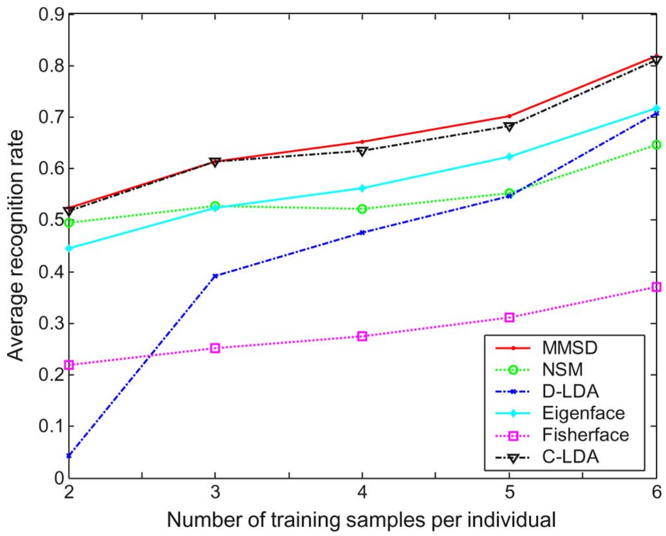


Fig. 6. Average recognition rates of various feature-extraction methods versus the numbers of training samples per individual.

TABLE I
MEANS AND STANDARD DEVIATIONS OF RECOGNITION RATES OF MMSD AND MSD

Parameter c		10	100	1000
MMSD	Mean	0.9605	0.9625	0.9630
	Std	0.0140	0.0153	0.0140
MSD	Mean	0.9505	0.9525	0.9515
	Std	0.0192	0.0186	0.0197

Fig. 6 displays curves of the average recognition rates of various feature-extraction methods over varying number of training samples per individual. MMSD is again the most effective of the facial feature-extraction approaches that were tested.

E. Comparison of the Effectiveness of MMSD With the Effectiveness of MSD

The comparison of the effectiveness of MMSD with the effectiveness of MSD is conducted on a small database ORL. In ORL dataset, there are ten different images for each of 40 individuals. All images are grayscale and normalized with a resolution of 112×92. In each of the ten runs, we use five images of each person for training and the remaining five for testing. The images of each person numbered 1 to 5, 2 to 6, …, 10 to 4 are used as training samples for the first, second,…, and the tenth run, respectively. In the experiment, the number of features extracted by MMSD is 39, and the classifier used in combination with MMSD is the NN.

Table I lists the means and standard deviations of the recognition rates of MMSD and MSD using various values of the parameter c.

Experimental results indicate that the MMSD is more effective than the MSD on the ORL face-image database.

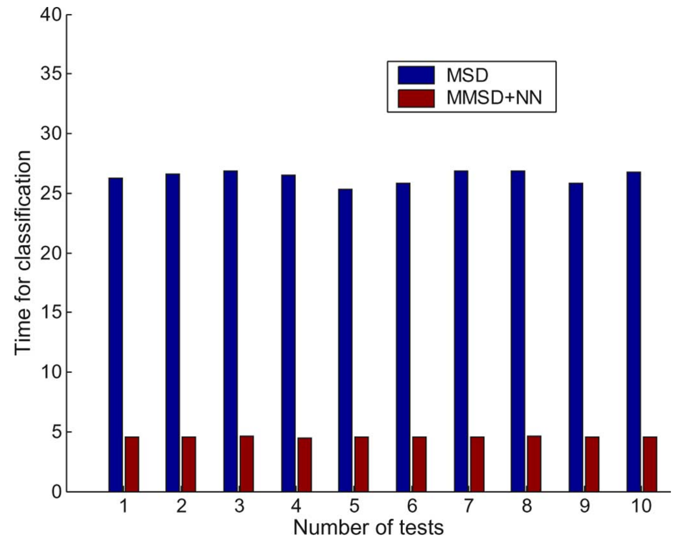


Fig. 7. Comparison of the efficiencies of MSD classifiers with MMSD + NN classifiers on the ORL face-image database.

F. Comparison of the Efficiency of MMSD With the Efficiency of MSD on the ORL Database

The chief motivation for extending the MSD to MMSD is to promote efficiency. In this section, we compare the efficiency of an MSD classifier with that of the MMSD in combination with an NN classifier on the ORL database.

The experimental design is the same as in Section IV-E.

Fig. 7 displays the time (in seconds) taken by the MSD classifier and MMSD in combination with an NN classifier for pattern classification in each run on the ORL database. The MMSD, in combination with the NN classifier, is much faster than the MSD classifier.

G. Comparison of the Efficiency of MMSD With the Efficiency of Other Facial Feature-Extraction Methods on the FERET Database

In this section, we compare the efficiency of MMSD with that of NSM, D-LDA, Fisherface, and C-LDA on the FERET database. The experimental design is the same as in Section IV-D.

Fig. 8 displays the average time (in seconds) taken by various methods used for feature extraction in each test on the FERET database. Although slower than D-LDA and Fisherface, MMSD is faster than NSM and C-LDA.

V. CONCLUSION

In this paper, we present a novel subspace-based facial feature-extraction method based on the MMSD discriminant criterion. Theoretical analysis and experimental studies demonstrate that the MMSD feature-extraction method has many advantages over conventional facial feature-extraction approaches. First, it is elegant in theory, avoiding the singularity problem by using a novel class separability measure. Second, it is easy to compute. Although slower than D-LDA and Fisherface, it is faster than NSM, C-LDA, and MSD.

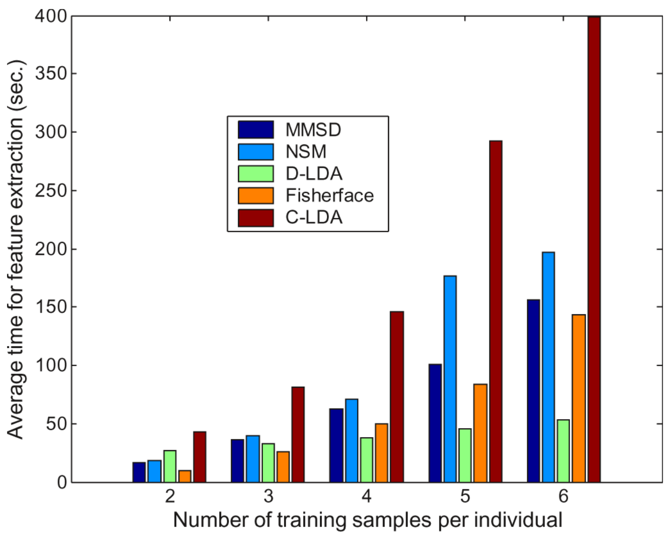


Fig. 8. Comparison of the efficiencies of MMSD with those of NSM, D-LDA, Fisherface, and C-LDA on the FERET database.

Third, it is very effective in face recognition. Since it derives discriminant vectors both in the range of the between-class scatter matrix and in the null space of the within-class scatter matrix, it outperforms the NSM, D-LDA, eigenface, Fisherface, and C-LDA.

REFERENCES

[1] F. X. Song, D. Zhang, Q. L. Chen, and J. Z. Wang, "Face recognition based on a novel linear discriminant criterion," *Pattern Anal. Appl.*, vol. 10, no. 3, pp. 165–174, Aug. 2007.

[2] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[3] H. Eschenauer, J. Koski, and A. Osyczka, *Multicriteria Design Optimization*. Berlin, Germany: Springer-Verlag, 1990.

[4] F. X. Song, J. Y. Yang, and S. H. Liu, "Large margin linear projection and face recognition," *Pattern Recognit.*, vol. 37, no. 9, pp. 1953–1955, Sep. 2004.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[6] Y. F. Guo, S. J. Li, J. Y. Yang, T. T. Shu, and L. D. Wu, "A generalized Foley–Sammon transform based on generalized fisher discriminant criterion and its application to face recognition," *Pattern Recognit. Lett.*, vol. 24, no. 1–3, pp. 147–158, Jan. 2003.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston, MA: Academic, 1990.

[8] D. C. Tao, X. L. Li, S. J. Maybank, and X. D. Wu, "Human carrying status in visual surveillance," in *Proc. CVPR*, 2006, vol. 2, pp. 1670–1677.

[9] D. C. Tao, X. L. Li, W. M. Hu, S. J. Maybank, and X. D. Wu, "Supervised tensor learning," in *Proc. ICDM*, 2005, pp. 450–457.

[10] K. Liu, Y. Q. Cheng, and J. Y. Yang, "A generalized optimal set of discriminant vectors," *Pattern Recognit.*, vol. 25, no. 7, pp. 731–739, Jul. 1992.

[11] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application to face recognition," *Pattern Recognit.*, vol. 34, no. 10, pp. 2067–2070, Oct. 2001.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Beijing, China: Tsinghua Univ. Press, 2001.

[13] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, Oct. 2000.

[14] F. X. Song, D. Zhang, and J. Y. Yang, "A novel dimensionality-reduction approach for face recognition," *Neurocomputing*, vol. 69, no. 13–15, pp. 1683–1687, Aug. 2006.

[15] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 72–86, 1991.

[16] J. Yang, J. Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognit.*, vol. 36, no. 2, pp. 563–566, Feb. 2003.

[17] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[18] P. J. Phillips. (2004). *The facial recognition technology (FERET) database*. [Online]. Available: http://www.itl.nist.gov/iad/humanid/feret/feret_master.html

[19] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[20] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

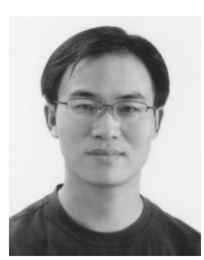
[21] J. W. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, Jan. 2005.

Fengxi Song was born in Anhui, China, on February 29, 1964. He received the B.S. degree in mathematics from Anhui University, Hefei, China, in 1984, the M.S. degree in applied mathematics from the Changsha Institute of Technology, Changsha, China, in 1987, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 2004.

Currently, he is a Professor with the New Star Research Institute of Applied Technology, Hefei, and a Postdoctoral Research Fellow with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His research interests include computer vision, machine learning, and automatic text categorization.

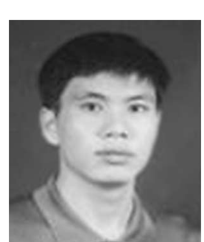
David Zhang (SM'95) received the degree in computer science from Peking University, Beijing, China, in 1974, the M.S. and Ph.D. degrees in computer science and engineering from the Harbin Institute of Technology, Harbin, China, in 1983 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

He is currently a Professor with the Hong Kong Polytechnic University, Kowloon, Hong Kong. He is the author of more than 140 journal papers, 20 book chapters, and ten books. He is the holder of several patents in both the U.S. and China. His research interests include automated biometrics-based authentication, pattern recognition, and biometric technology and systems.



Dayong Mei received the B.S. and M.S. degrees from the New Star Research Institute of Applied Technology, Hefei, China, in 2000 and 2003, respectively.

Currently, he is an Assistant Researcher with the Academy of Military Science, Beijing, China. His current research interests are in the areas of text categorization and virtual reality.



Zhongwei Guo received the M.S. degree in decision science from the New Star Research Institute of Applied Technology, Hefei, China, in 1999, and the Ph.D. degree in system engineering from Nanjing University of Science and Technology, Nanjing, China, in 2003.

He is currently an Associate Professor with the New Star Research Institute of Applied Technology. His career has focused on natural language processing and object recognition.