Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 2437–2448
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

A new method of feature fusion and its application in image recognition

Quan-Sen Sun[a,b,*], Sheng-Gen Zeng[a], Yan Liu[a], Pheng-Ann Heng[c], De-Shen Xia[a]

[a]*Department of Computer Science, Nanjing University of Science & Technology, Nanjing 210094, People's Republic of China*
[b]*Department of Mathematics, Jinan University, Jinan 250022, People's Republic of China*
[c]*Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong*

Received 11 March 2004; received in revised form 17 December 2004; accepted 17 December 2004

Abstract

A new method of feature extraction, based on feature fusion, is proposed in this paper according to the idea of canonical correlation analysis (CCA). At first, the theory framework of CCA used in pattern recognition and its reasonable description are discussed. The process can be explained as follows: extract two groups of feature vectors with the same pattern; establish the correlation criterion function between the two groups of feature vectors; and extract their canonical correlation features to form effective discriminant vector for recognition. Then, the problem of canonical projection vectors is solved when two total scatter matrixes are singular, such that it fits for the case of high-dimensional space and small sample size, in this sense, the applicable range of CCA is extended. At last, the inherent essence of this method used in recognition is analyzed further in theory. Experimental results on Concordia University CENPARMI database of handwritten Arabic numerals and Yale face database show that recognition rate is far higher than that of the algorithm adopting single feature or the existing fusion algorithm.

Keywords: Canonical correlation analysis (CCA); Feature fusion; Feature extraction; Handwritten character recognition; Face recognition

1. Introduction

Information fusion technology is one of the emerging technologies of data processing. Among the three levels of information fusion (pixel level, feature level, and decision level), the decision level fusion, delegated by multi-classifier combination, has been one of the hot research fields on pattern recognition, and has achieved successful application in the aspects of handwritten character and face recognition [1–3]. Although the research of the feature level fusion starts not as early as other fusion methods, it got delightful development [4,5].

The advantage of the feature level fusion is obvious. Different feature vectors extracted from the same pattern always reflects the different characteristic of patterns. By optimizing and combining these different features, it not only keeps the effective discriminant information of multi-feature, but also eliminates the redundant information to certain degree. This is especially important to classification and recognition.

There are the two existing feature fusion methods. One is to group two sets of feature vectors into one union-vector [4], and then to extract features in the higher-dimension real vector space. Another one is to combine two sets of feature vectors by a complex vector [5,6], and then to extract features in the complex vector space. Both feature fusion methods can increase the recognition rate, the feature fusion

* Corresponding author. Department of Computer Science, Nanjing University of Science & Technology, Nanjing 210094, People's Republic of China. Tel./fax: +86 5312927158.
E-mail address: qssun@beelink.com (Q.-S. Sun).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.12.013

method based on the union-vector is referred as *serial feature fusion* and the one based on the complex vector is called *parallel feature fusion* [6].

Canonical correlation analysis (CCA) is one of the statistical methods dealing with the mutual relationships between two random vectors, and it has the same importance as principal component analysis (PCA) and linear discriminant analysis (LDA) in multivariate statistical analysis. It is one of the valuable multi-data processing methods [7,8]. In recent years, CCA has been applied to several fields such as signal processing, computer vision, neural network and speech recognition [9–13].

In this paper, a new method of feature fusion is proposed adopting the idea of CCA. First, we discuss the framework of CCA used in pattern recognition. That is to extract two groups of feature vectors with the same sample, then to establish the correlation criterion function between the two groups of feature vectors, to extract their canonical correlation features according to this criterion, and to form effective discriminant vectors for recognition. Then, the problem of canonical projected vectors is solved when two total scatter matrixes are singular, such that it fits for the case of high-dimensional space and small sample size, in this sense, the applicable range of CCA is extended. At last, the inherent essence of this method used in recognition is analyzed further in theory. This method uses correlation features between two groups of feature vectors as effective discriminant information, so it not only is suitable for information fusion, but also eliminates the redundant information within the features. This is a new way to classification and recognition.

Experimental results on Concordia University CENPARMI database of handwritten Arabic numerals and Yale standard face database show that recognition rate is far higher than that of adopting a single feature or the fusion algorithm existed, and this algorithm is efficient and robust. At the same time, the results of simulation show that this algorithm cannot only realize the compression of primitive feature dimensions, but also be of good classification performance, reflecting the essential feature of the images.

The rest of this paper is organized as follows. In Section 2, the theory and method of CCA used in feature fusion are presented. In Section 3, the problem of combined feature extraction is solved under the case of the high-dimensional space and small sample size. In Section 4, the proposed feature fusion method has been tested on a big sample database and small sample database and compared to other methods. In Section 5, we discuss the inherent essence this method in theory and explain why it is effective in recognition. Finally, conclusions are drawn in Section 5.

2. The theory and method of feature fusion

2.1. The basic idea of CCA

In multivariate statistical analysis, the correlation problem of two random vectors often needs to be studied, that is to convert the correlation research of two random vectors into that of a few pairs of variables, which are uncorrelated. H. Hotelling developed this idea in 1936 [14].

Concretely, considering two zero-mean random vectors X and Y, CCA finds a pair of directions α and β that maximize the correlation between the projections $a_1 = \alpha^{\mathrm{T}} X$ and $b_1 = \beta^{\mathrm{T}} Y$. The projections a_1 and b_1 are called the first pair of canonical variates. Then finding the second pair of canonical variates a_2 and b_2, which is uncorrelated with canonical variates a_1 and b_1 each other and also maximize the correlation between them. Just do like this until all the correlation features of X and Y are extracted. In order to study the correlation of X and Y, we only need analyze the correlation of a few pairs of canonical variates.

2.2. The theory and algorithm of combine feature extraction

Suppose $\omega_1, \omega_2, \ldots, \omega_c$ are c known pattern classes. Let $\Omega = \{\xi \mid \xi \in \mathbb{R}^N\}$ is a training sample space. Given $A = \{x \mid x \in \mathbb{R}^p\}$ and $B = \{y \mid y \in \mathbb{R}^q\}$, where x and y are the two feature vectors of the same sample ξ extracted by different means respectively. We will discuss the feature fusion in the transformed training sample feature space A and B.

Suppose that A and B are regarded as two random vector spaces. Our idea is to extract the canonical correlation features between x and y based on the idea of CCA described in Section 2.1, we denote them as $\alpha_1^{\mathrm{T}} x$ and $\beta_1^{\mathrm{T}} y$ (the first pair), $\alpha_2^{\mathrm{T}} x$ and $\beta_2^{\mathrm{T}} y$ (the second pair), …, $\alpha_d^{\mathrm{T}} x$ and $\beta_d^{\mathrm{T}} y$ (the dth pair). Given the following:

$$X^* = (\alpha_1^{\mathrm{T}} x, \alpha_2^{\mathrm{T}} x, \ldots, \alpha_d^{\mathrm{T}} x)^{\mathrm{T}} = (\alpha_1, \alpha_2, \ldots, \alpha_d)^{\mathrm{T}} x = W_x^{\mathrm{T}} x, \tag{1}$$

$$Y^* = (\beta_1^{\mathrm{T}} y, \beta_2^{\mathrm{T}} y, \ldots, \beta_d^{\mathrm{T}} y)^{\mathrm{T}} = (\beta_1, \beta_2, \ldots, \beta_d)^{\mathrm{T}} y = W_y^{\mathrm{T}} y. \tag{2}$$

Following two linear transformation (3) and (4):

$$Z_1 = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^{\mathrm{T}} x \\ W_y^{\mathrm{T}} y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{3}$$

$$Z_2 = X^* + Y^* = W_x^{\mathrm{T}} x + W_y^{\mathrm{T}} = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix} \tag{4}$$

as the combinatorial feature projected respectively, used for classification, while the transformation matrix are

$$W_1 = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix} \quad \text{and} \quad W_2 = \begin{pmatrix} W_x \\ W_y \end{pmatrix},$$

where $W_x = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, $W_y = (\beta_1, \beta_2, \ldots, \beta_d)$.

Definition 1. The directions α_i and β_i are call the ith pair of canonical projective vectors (CPV) of x and y, $\alpha_i^{\mathrm{T}} x$ and $\beta_i^{\mathrm{T}} y$ as their ith canonical correlation features. We also call W_1 and W_2 as the canonical projective matrix (CPM), Z_1 and Z_2 as the canonical correlation discriminant feature (CCDF),

and linear transformation (3) and (4) as the feature fusion strategy I (FFS I) and the feature fusion strategy II(FFS II), respectively.

Next, we will discuss how to obtain the value and the quality of CPV and CCDF.

Suppose that $S_{xx} \in \mathbb{R}^{p\times p}$ and $S_{yy} \in \mathbb{R}^{q\times q}$ denote the covariance matrices (their total scatter matrix are nS_{xx} and nS_{yy}, where n is number of sample) of A and B respectively, while $S_{xy} \in \mathbb{R}^{p\times q}$ denotes their between-set covariance matrix, then the covariance matrix of $\begin{pmatrix} x \\ y \end{pmatrix}$ can be denoted as

$$S = \begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(y,x) & Var(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix},$$

where S_{xx} and S_{yy} are positive definite, $S_{xy}^{\mathrm{T}} = S_{yx}$, and $r = rank(S_{xy})$.

In this Section, we will only discuss the case when S_{xx} and S_{yy} are non-singular. The case of singular, we will discuss in Section 3.

Considering the linear combination of x and y

$$\alpha^{\mathrm{T}}x = a_1x_1 + a_2x_2 + \cdots + a_px_p,$$

$$\beta^{\mathrm{T}}y = b_1y_1 + b_2y_2 + \cdots + b_qy_q,$$

where $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ are two arbitrary nonzero vectors. CCA finds pairs of directions α and β that maximize correlation between the canonical variates $\alpha^{\mathrm{T}}x$ and $\beta^{\mathrm{T}}y$. Since

$$Var(\alpha^{\mathrm{T}}x) = \alpha^{\mathrm{T}}Var(x)\alpha = \alpha^{\mathrm{T}}S_{xx}\alpha,$$

$$Var(\beta^{\mathrm{T}}y) = \beta^{\mathrm{T}}Var(y)\beta = \beta^{\mathrm{T}}S_{yy}\beta,$$

$$Cov(\alpha^{\mathrm{T}}x, \beta^{\mathrm{T}}y) = \alpha^{\mathrm{T}}Cov(x,y)\beta = \alpha^{\mathrm{T}}S_{xy}\beta.$$

We can give the criterion function as the following:

$$J(\alpha,\beta) = \frac{\alpha^{\mathrm{T}}S_{xy}\beta}{(\alpha^{\mathrm{T}}S_{xx}\alpha\beta^{\mathrm{T}}S_{yy}\beta)^{1/2}}. \tag{5}$$

Obviously, the criterion function (5) has the following characteristics:

- $J(k\alpha,\ l\beta) = J(\alpha,\ \beta)$, $\forall k,\ l \in \mathbb{R}$.
- The extremum of $J(\alpha,\beta)$ is nothing to do with length of α and β, but has something to do with their direction.

According the above characteristic, we can think that

$$\alpha^{\mathrm{T}}S_{xx}\alpha = \beta^{\mathrm{T}}S_{yy}\beta = 1. \tag{6}$$

Now the question is transformed to the solving of CPV α and β in constraint (6), with the extremum of criterion (5).

Suppose that $(k-1)$ pair of CPVs $\{\alpha_1;\beta_1\}, \{\alpha_2;\beta_2\}, \ldots, \{\alpha_{k-1};\beta_{k-1}\}$ are obtained, then the kth can be done by solving the following optimization problem:

$$\text{Mode1}\begin{cases} \max J(\alpha,\beta), \\ \alpha^{\mathrm{T}}S_{xx}\alpha = \beta^{\mathrm{T}}S_{yy}\beta = 1, \\ \alpha_i^{\mathrm{T}}S_{xx}\alpha = \beta_i^{\mathrm{T}}S_{yy}\beta = 0 \quad (i = 1,2,\ldots,k-1), \\ \alpha \in \mathbb{R}^p, \quad \beta \in \mathbb{R}^q. \end{cases} \tag{7}$$

We will discuss the optimal solution of *model 1* as follows.

Using Lagrange multiplier method to transform Eq. (5). Let

$$L(\alpha,\beta) = \alpha^{\mathrm{T}}S_{xy}\beta - \frac{\lambda_1}{2}(\alpha^{\mathrm{T}}S_{xx}\alpha - 1) - \frac{\lambda_2}{2}(\beta^{\mathrm{T}}S_{yy}\beta - 1),$$

where λ_1 and λ_2 are Lagrange multipliers.

Setting the partial derivatives of $L(\alpha,\beta)$ with respect to α and β to zero

$$\frac{\partial L}{\partial \alpha} = S_{xy}\beta - \lambda_1 S_{xx}\alpha = 0, \tag{8}$$

$$\frac{\partial L}{\partial \beta} = S_{yx}\alpha - \lambda_2 S_{yy}\beta = 0. \tag{9}$$

Multiplying both side of Eqs. (8) and (9) by α^{T} and β^{T} respectively, considering the constraint (6), we obtain

$$\alpha^{\mathrm{T}}S_{xy}\beta = \lambda_1\alpha^{\mathrm{T}}S_{xx}\alpha = \lambda_1,$$

$$\beta^{\mathrm{T}}S_{yx}\alpha = \lambda_2\beta^{\mathrm{T}}S_{yy}\beta = \lambda_2.$$

Since $S_{xy}^{\mathrm{T}} = S_{yx}$, so $\lambda_1 = \lambda_1^{\mathrm{T}} = (\alpha^{\mathrm{T}}S_{xy}\beta)^{\mathrm{T}} = \beta^{\mathrm{T}}S_{yx}\alpha = \lambda_2$. Let $\lambda_1 = \lambda_2 = \lambda$, then

$$J(\alpha,\beta) = \alpha^{\mathrm{T}}S_{xy}\beta = \beta^{\mathrm{T}}S_{yx}\alpha = \lambda. \tag{10}$$

This shows that the Lagrange multipliers λ_1 and λ_2 are equal to the correlation coefficient of $\alpha^{\mathrm{T}}x$ and $\beta^{\mathrm{T}}y$.

So Eqs. (8) and (9) can also be written as

$$S_{xy}\beta - \lambda S_{xx}\alpha = 0, \tag{11}$$

$$S_{yx}\alpha - \lambda S_{yy}\beta = 0. \tag{12}$$

Since S_{xx} and S_{yy} are both positive definite, from Eqs. (11) and (12), we obtain

$$S_{xy}S_{yy}^{-1}S_{yx}\alpha = \lambda^2 S_{xx}\alpha, \tag{13}$$

$$S_{yx}S_{xx}^{-1}S_{xy}\beta = \lambda^2 S_{yy}\beta. \tag{14}$$

Now the question has been converted to the solving of two generalized eigenproblem. Given $M_{xy} = S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}$ and $M_{yx} = S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$, then Eqs. (12) and (13) are changed to

$$M_{xy}\alpha = \lambda^2\alpha, \tag{15}$$

$$M_{yx}\beta = \lambda^2\beta. \tag{16}$$

Here is the theorem about the eigenvalue and eigenvector of M_{xy} and M_{yx} from [8].

Theorem 1. *M_{xy} and M_{yx} have the same nonzero eigenvalues, which satisfies $1 > \lambda_1^2 \geqslant \lambda_2^2 \geqslant \cdots \geqslant \lambda_r^2 > 0$, where $r = rank(S_{xy})$.*

In order to get the solution under the constraint (7), suppose that

$$G_1 = S_{xx}^{-1/2} S_{xy} S_{yy}^{-1} S_{yx} S_{xx}^{-1/2},$$
$$G_2 = S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} S_{xy} S_{yy}^{-1/2}. \quad (17)$$

Then we can obtain that M_{xy} and G_1 have the same nonzero eigenvalues according to the related theorem of matrix, so do M_{yx} and G_2. Such that the nonzero eigenvalues are all equal to $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2$. Let $H = S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2}$, then $G_1 = HH^{\mathrm{T}}, G_2 = H^{\mathrm{T}}H$. Using singular value decompose (*SVD*) theorem on matrix H, we obtain $H = \sum_{i=1}^{r} \lambda_i u_i v_i^{\mathrm{T}}$, where $\lambda_1^2 \geqslant \lambda_2^2 \geqslant \cdots \geqslant \lambda_r^2$ are entire nonzero eigenvalues of G_1 and G_2, u_i and $v_i (i = 1, 2, \ldots, r)$ are the orthogonal eigenvectors of G_1 and G_2 corresponding to the nonzero eigenvalue λ_i^2, respectively.

From above, we can infer to an important Theorem 2 as follows:

Theorem 2. *Given $\alpha_i = S_{xx}^{-1/2} u_i, \beta_i = S_{yy}^{-1/2} v_i, \; i = 1, 2, \ldots, r$. Then*

(1) *α_i and β_i are the eigenvectors of M_{xy} and M_{yx} corresponded to λ_i^2.*

(2)

$$\begin{cases} \alpha_i^{\mathrm{T}} S_{xx} \alpha_j = \beta_i^{\mathrm{T}} S_{yy} \beta_j = \delta_{ij}, \\ \alpha_i^{\mathrm{T}} S_{xy} \beta_j = \lambda_i \delta_{ij}, \end{cases} \quad (18)$$

where $\delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j, \end{cases} \quad i, j = 1, 2, \ldots, r.$

As the matter of fact, Theorem 2 gives the solution under the constraint (6) and (7), and the extremum of criterion (5), that is also intended to the optimized solution of model 1.

According to Eq. (10) and Theorem 2, we can get the corollary as follows:

Corollary 1. *Given all the eigenvalues of G_1 or M_{xy} as $\lambda_1^2 \geqslant \cdots \geqslant \lambda_r^2 > \lambda_{r+1}^2 = \cdots = \lambda_p^2 = 0$ ($p \leqslant q$). Then the criterion function* (5) *has $J(\alpha_i, \beta_i) = \lambda_i \; (i = 1, \ldots, p)$.*

Corollary 2. *According to the above feature fusion strategy II (FFS II), the extracted combination features are uncorrelated, and the combined projective vectors are M—orthonormal. Where*

$$M = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}.$$

Proof. Let $X^* = W_x^{\mathrm{T}} x = (x_1^*, x_2^*, \ldots, x_d^*)^{\mathrm{T}}$, $Y^* = W_y^{\mathrm{T}} x = (y_1^*, y_2^*, \ldots, y_d^*)^{\mathrm{T}}$.

From Theorem 2, we know, $\forall i \neq j$

$$\mathrm{cov}(x_i^*, x_j^*) = \alpha_i^{\mathrm{T}} S_{xx} \alpha_j = 0,$$
$$\mathrm{cov}(y_i^*, y_j^*) = \beta_i^{\mathrm{T}} S_{yy} \beta_j = 0,$$
$$\mathrm{cov}(x_i^*, y_j^*) = \alpha_i^{\mathrm{T}} S_{xy} \beta_j = 0.$$

In Eq. (4),

$$\mathrm{cov}(x_i^* + y_i^*, x_j^* + y_j^*) = \mathrm{cov}(x_i^*, x_j^*) + 2\mathrm{cov}(x_i^*, y_j^*) + \mathrm{cov}(y_i^*, y_j^*) = 0.$$

So, the components of feature vector Z_2 are uncorrelated. Sine

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = 2(1 + \lambda_i)\delta_{ij}$$

so $\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}$ and $\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix}$ is M—orthonormal, where $i, j = 1, 2, \ldots, r$. □

Theorem 3. *Under criterion* (5), *the number of the efficient CPV, satisfying the constraint* (6) *and* (7), *is r pairs at most, where $r = rank(S_{xy})$, and getting $d (\leqslant r)$ pairs CPV are compose of the eigenvectors corresponding to first d maximum eigenvalues of two eigenequation* (15) *and* (16) *that satisfy Eq.* (18).

Proof. From Theorem 1 and Corollary 1, we know

$$J(\alpha_i, \beta_i) = \lambda_i, i = 1, 2, \ldots, \min(p, q), \quad \text{where}$$
$$\lambda_1^2 \geqslant \cdots \geqslant \lambda_r^2 > \lambda_{r+1}^2 = \cdots = \lambda_p^2 = 0.$$

So, $J(\alpha_i, \beta_i) = 0 (j = r + 1, \ldots, \min(p, q))$. In the case, the efficient CPV can not be extracted. It means that the number of the efficient CPV is r pairs at most. Then from Theorem 2, we know that, $d (\leqslant r)$ pairs CPV can be composed of the eigenvectors corresponding to first d maximum eigenvalues of two eigenequations $M_{xy}\alpha = \lambda^2 \alpha$ and $M_{yx}\beta = \lambda^2 \beta$ respectively, and satisfying condition (18). □

2.3. The steps of this algorithm

Step 1: Extract sets of two different feature vectors with the same pattern sample to form the training sample spaces A and B transformed from the original pattern sample space Ω.

Step 2: Compute the covariance matrixes S_{xx} and S_{yy} of the samples in A and B, and their between-set covariance matrix S_{xy}.

Step 3: Compute G_1 and G_2 according to Eq. (17), then find their nonzero eigenvalues $\lambda_1^2 \geqslant \lambda_2^2 \geqslant \cdots \geqslant \lambda_r^2$, and corresponded to orthonormal eigenvectors u_i and $v_i (i = 1, 2, \ldots, r)$.

Step 4: Get the CPV α_i and $\beta_i \; (i = 1, 2, \ldots, r)$, according to Theorem 2, choose the first d pairs vector to make CPT.

Step 5: Use FFS I or FFS II to extract CCDF, which are used for classification.

From Eqs. (11) and (12), we know that we only need solve one among each pair of CPV, the other one can be solved by Eqs. (11) or (12). Generally, we can choose the low-order matrix G_1 or G_2 to find its eigenvalues and eigenvectors, such that the computational complexity can be lowered.

2.4. Pretreatment of feature fusion

Normally, when two groups of features of the same pattern are extracted, CCDF can be directly extracted according to above arithmetic. According to Theorem 3, we can conclude that there can be at most r pairs of effective CPV. Therefore, number of dimensions of extracted CCDF does not exceed r. When number of dimension p of two groups of features differs greatly from q, does extracted CCDF satisfy need for classification? Does loss of large amount of information occur in the group of feature that has higher number of dimension? In direct perception, such possibilities exist. Therefore, we can adopt the following method of pretreatment to reduce detrimental effect brought about by large difference in number of dimension of feature. For two groups of feature vectors $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, fusion is first carried out according to serial mode to form a union vector of $(p+q)$-dimension. Then data are redistributed to obtain two groups of new feature vectors $x^* \in \mathbb{R}^{\frac{p+q}{2}}$ and $y^* \in \mathbb{R}^{\frac{p+q}{2}}$ (or $x^* \in \mathbb{R}^{\frac{p+q+1}{2}}$ and $y^* \in \mathbb{R}^{\frac{p+q-1}{2}}$). Finally, CCDF between x^* and y^* is extracted using above arithmetic. Effect of this pretreatment will be demonstrated in later experiments.

3. The high-dimensional case and small sample size problem

3.1. Algorithm and theory

There are many typical small sample problems in the realm of pattern recognition. For example, we always need to handle the problem of high-dimensional case and small sample size in face recognition. Commonly the total scatter matrix of the sample is singular in these problems. Because the dimension of the image vector to be recognized is high, it is very difficult or even impossible to find enough training sample so as to ensure the reversibility of the total scatter matrix. In this case, the main problem discussed in this section will be that how to get the CPV in the feature space A and B as mentioned in Section 2.2.

As an important multi-data processing method, CCA finds its application in many realms. However, these applications are merely limited to the instance that the covariance matrix of the feature space which the two random vectors belonging to is nonsingular [8–15]. An approach to dealing with singular covariance matrices and to controlling complexity is to add a multiple of the identity matrix λI, $\lambda > 0$ to S_{xx} and S_{yy}; this operation simply shifts the eigenvalues by λ, and, thus, if λ is chosen large enough, will render both matrices positive definite [16]. It is another approach that general reversibility is used to solve the problem of the solution of the CPV [7]. As mentioned above, the solution is not an exact solution theoretically. Thus its application is limited.

Our idea is to convert the high-dimensional feature space of the primitive samples to a low-dimensional Euclid space, in the premise of not losing any valid information. While the covariance matrix is nonsingular in the low-dimensional Euclidean space, so we only need to extract the CPV in the space.

Supposing that at least one of S_{xx} and S_{yy} is singular, for instance, S_{xx} is singular while S_{yy} is nonsingular (when both S_{xx} and S_{yy} are singular, the method is similar), and we can know from the definition of the covariance matrix that S_{xx} is semi-positive definite while S_{yy} is positive definite. Then we give an eye to how to get the solution of the CPV.

Suppose $\gamma_1, \gamma_2, \ldots, \gamma_p$ are p orthonormal eigenvectors of S_{xx}, then $\mathbb{R}^p = span(\gamma_1, \gamma_2, \ldots, \gamma_p)$.

Definition 2. Given a subspace $\Phi_x = span\{\gamma_1, \gamma_2, \ldots, \gamma_m\}$, its orth-complement is $\Phi_x^{\perp} = span\{\gamma_{m+1}, \ldots, \gamma_p\}$, where $m = rank(S_{xx})$. $\gamma_1, \gamma_2, \ldots, \gamma_m$ are eigenvectors corresponding to the nonzero eigenvalues of S_{xx}.

We can know from the Definition 2 that $\mathbb{R}^p = \Phi_x + \Phi_x^{\perp}$, for $\forall \alpha \in \mathbb{R}^p$, α can be denoted as $\alpha = \alpha^* + \varphi$, where $\alpha^* \in \Phi_x$, $\varphi \in \Phi_x^{\perp}$, then define the mapping $f : \alpha \to \alpha^*$. Obviously, f is a linear transformation from $\mathbb{R}^p$ to Φ_x.

According to the relational theorem of Linear Algebra, we can easily conclude that

Lemma. *If S_{xx} is singular, then $\alpha^{\mathrm{T}} S_{xx} \alpha = 0$ if and only if $S_{xx}\alpha = 0$.*

By Lemma, we can easily find that $\Phi_x^{\perp}$ is the null space of S_{xx}.

Theorem 4. *With the above linear transformation $f : \alpha \to \alpha^*$, then $J(\alpha^*, \beta) = J(\alpha, \beta)$.*

Proof. Lemma and the definition of $\Phi_x^{\perp}$ show that

$$\alpha^{*\mathrm{T}} S_{xx} \varphi = \varphi^{\mathrm{T}} S_{xx} \varphi = 0. \tag{19}$$

So

$$\begin{aligned} \alpha^{\mathrm{T}} S_{xx} \alpha &= \alpha^{*\mathrm{T}} S_{xx} \alpha^* + 2\alpha^{*\mathrm{T}} S_{xx} \varphi + \varphi^{\mathrm{T}} S_{xx} \varphi \\ &= \alpha^{*\mathrm{T}} S_{xx} \alpha^*. \end{aligned} \tag{20}$$

Suppose that

$$X = (x_1 - \mu_x, x_2 - \mu_x, \ldots, x_n - \mu_x),$$

$$Y = (y_1 - \mu_y, y_2 - \mu_y, \ldots, y_n - \mu_y),$$

where μ_x and μ_y are the mean vector of samples of A and B respectively, n is the total number of samples, then

$$S_{xx} = \frac{1}{n} XX^{\mathrm{T}}, \quad S_{xy} = \frac{1}{n} XY^{\mathrm{T}}.$$

From Eq. (19), we can conclude that

$$\varphi^{\mathrm{T}} S_{xx} \varphi = \frac{1}{n} \varphi^{\mathrm{T}} XX^{\mathrm{T}} \varphi = \frac{1}{n} (X^{\mathrm{T}}\varphi)^{\mathrm{T}} (X^{\mathrm{T}}\varphi) = 0 \;\Rightarrow\; X^{\mathrm{T}}\varphi = 0. \tag{21}$$

So

$$\varphi^{\mathrm{T}} S_{xy} \beta = \frac{1}{n} \varphi^{\mathrm{T}} XY^{\mathrm{T}} \beta = \frac{1}{n} (X^{\mathrm{T}}\varphi)^{\mathrm{T}} (Y^{\mathrm{T}}\beta) = 0. \tag{22}$$

With Eqs. (20) and (22), criterion function (5) can be changed into

$$\begin{aligned} J(\alpha, \beta) &= \frac{\alpha^{\mathrm{T}} S_{xy} \beta}{(\alpha^{\mathrm{T}} S_{xx} \alpha \beta^{\mathrm{T}} S_{yy} \beta)^{1/2}} \\ &= \frac{\alpha^{*\mathrm{T}} S_{xy} \beta + \varphi^{\mathrm{T}} S_{xy} \beta}{(\alpha^{*\mathrm{T}} S_{xx} \alpha^* \beta^{\mathrm{T}} S_{yy} \beta)^{1/2}} \\ &= \frac{\alpha^{*\mathrm{T}} S_{xy} \beta}{(\alpha^{*\mathrm{T}} S_{xx} \alpha^* \beta^{\mathrm{T}} S_{yy} \beta)^{1/2}} \\ &= J(\alpha^*, \beta). \qquad \square \end{aligned}$$

Theorem 4 shows that all CPV can be derived from both subspaces Φ_x and $\mathbb{R}^q$ without any loss of effective information with respect to criterion function (5). So *Model* 1 equates

$$\text{Model 2} \quad \begin{cases} \max J(\alpha, \beta), \\ \alpha^{\mathrm{T}} S_{xx} \alpha = \beta^{\mathrm{T}} S_{yy} \beta = 1, \\ \alpha_i^{\mathrm{T}} S_{xx} \alpha = \beta_i^{\mathrm{T}} S_{yy} \beta = 0 \quad (i = 1, 2, \ldots, k-1), \\ \alpha \in \Phi_x, \quad \beta \in \mathbb{R}^q. \end{cases}$$

We discuss the solution of *Model* 2 as follows.

Definition 3. Shows that $\dim \Phi_x = m$, so $\Phi_x \cong \mathbb{R}^m$, the corresponding isomorphic mapping is defined as $g : \tilde{\alpha} \to \alpha$, namely $\alpha = P\tilde{\alpha}$, where $P = (\gamma_1, \gamma_2, \ldots, \gamma_m)$, $\tilde{\alpha} \in \mathbb{R}^m$, $\alpha \in \Phi_x$.

With this isomorphic mapping, the criterion function $J(\alpha, \beta)$ turns into

$$\begin{aligned} J(\alpha, \beta) &= \frac{\alpha^{\mathrm{T}} S_{xy} \beta}{(\alpha^{\mathrm{T}} S_{xx} \alpha \beta^{\mathrm{T}} S_{yy} \beta)^{1/2}} \\ &= \frac{\tilde{\alpha}^{\mathrm{T}} (P^{\mathrm{T}} S_{xy}) \beta}{[\tilde{\alpha}^{\mathrm{T}} (P^{\mathrm{T}} S_{xx} P) \tilde{\alpha} \beta^{\mathrm{T}} S_{yy} \beta]^{1/2}}. \end{aligned}$$

Define criterion function

$$\tilde{J}(\tilde{\alpha}, \beta) = \frac{\tilde{\alpha}^{\mathrm{T}} \tilde{S}_{xy} \beta}{(\tilde{\alpha}^{\mathrm{T}} \tilde{S}_{xx} \tilde{\alpha} \beta^{\mathrm{T}} S_{yy} \beta)^{1/2}}, \tag{23}$$

where $\tilde{S}_{xx} = P^{\mathrm{T}} S_{xx} P$, $\tilde{S}_{xy} = P^{\mathrm{T}} S_{xy}$.

We can easily proof that $\tilde{S}_{xx}$ is positive definite matrix of m order, and $\tilde{S}_{xy}^{\mathrm{T}} = S_{xy}^{\mathrm{T}} P = S_{yx} P = \tilde{S}_{yx}^{\mathrm{T}}$.

Theorem 5. *Under the isomorphic mapping* $\alpha = P\tilde{\alpha}$, (α_0, β_0) *is the critical point of the criterion* $J(\alpha, \beta)$ *if and only if* $(\tilde{\alpha}_0, \beta_0)$ *is one of the criterion* $\tilde{J}(\tilde{\alpha}, \beta)$, *where* $\alpha_0 = P\tilde{\alpha}_0$.

Theorem 6. *Suppose* $P = (\gamma_1, \gamma_2, \ldots, \gamma_m)$, $\alpha_i = P\tilde{\alpha}_i$, $\alpha_j = P\tilde{\alpha}_j$, *then* α_i *and* α_j *are* S_{xx}*—orthonormal if and only if* $\tilde{\alpha}_i$ *and* $\tilde{\alpha}_j$ *are* $\tilde{S}_{xx}$*—orthonormal.*

Proof. Since

$$\alpha_i^{\mathrm{T}} S_{xx} \alpha_j = \tilde{\alpha}_i^{\mathrm{T}} (P^{\mathrm{T}} S_{xx} P) \tilde{\alpha}_j = \tilde{\alpha}_i^{\mathrm{T}} \tilde{S}_{xx} \tilde{\alpha}_j.$$

So

$$\alpha_i^{\mathrm{T}} S_{xx} \alpha_j = \delta_{ij} \;\Leftrightarrow\; \tilde{\alpha}_i^{\mathrm{T}} \tilde{S}_{xx} \tilde{\alpha}_j = \delta_{ij},$$

where $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$ $\square$

Under the isomorphic mapping $\alpha = P\tilde{\alpha}$, *Model* 2 equates

$$\text{Model 3} \quad \begin{cases} \max \tilde{J}(\tilde{\alpha}, \beta), \\ \tilde{\alpha}^{\mathrm{T}} \tilde{S}_{xx} \tilde{\alpha} = \beta^{\mathrm{T}} S_{yy} \beta = 1, \\ \tilde{\alpha}_i^{\mathrm{T}} \tilde{S}_{xx} \tilde{\alpha} = \beta_i^{\mathrm{T}} S_{yy} \beta = 0 \quad (i = 1, 2, \ldots, k-1), \\ \alpha \in \mathbb{R}^m, \beta \in \mathbb{R}^q. \end{cases}$$

According to Theorems 5 and 6, we can easily obtain

Theorem 7. *Suppose* $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_d$; $\beta_1, \ldots, \beta_d$ $(d \leqslant rank(\tilde{S}_{xy}))$ *are the optimal solutions of Model* 3, *then* $\alpha_1 = P\tilde{\alpha}_1, \ldots, \alpha_d = P\tilde{\alpha}_d$; $\beta_1, \ldots, \beta_d$ *are the CPV.*

Thus, we have given out S_{xx} which is singular and S_{yy} which is nonsingular, and the solution of the CPV. To get orthonormal eigenvectors corresponding to the nonzero eigenvalue of S_{xx}, firstly, assume $P = (\gamma_1, \gamma_2, \ldots, \gamma_m)$, then to get the covariance matrix $\tilde{S}_{xx} = P^{\mathrm{T}} S_{xx} P$ and the between-set covariance matrix $\tilde{S}_{xy} = P^{\mathrm{T}} S_{xy}$ of the sample on the $\mathbb{R}^m$ and $\mathbb{R}^q$, finally to get the CPV and the CCDF according to the algorithm in Section 2.3.

The situation for both S_{xx} and S_{yy} are singular is similar here, so it is ignored.

3.2. Algorithm analysis

The CPV which is obtained from Theorem 7 can be transformed below in order to extract feature.

$$\text{FFS I:} \quad Z_1 = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix} = W_1^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix}.$$

$$\text{FFS II:} \quad Z_2 = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix} = W_2^{\mathrm{T}} \begin{pmatrix} x \\ y \end{pmatrix}.$$

$$W_x = (\alpha_1, \ldots, \alpha_d) = (P\tilde{\alpha}_1, \ldots, P\tilde{\alpha}_d) = P(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_d) = P\tilde{W}_x.$$

So

$$W_1 = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix} = \begin{pmatrix} P\tilde{W}_x & 0 \\ 0 & W_y \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & E \end{pmatrix} \begin{pmatrix} \tilde{W}_x & 0 \\ 0 & W_y \end{pmatrix},$$

$$W_2 = \begin{pmatrix} W_x \\ W_y \end{pmatrix} = \begin{pmatrix} P\tilde{W}_x \\ W_y \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & E \end{pmatrix} \begin{pmatrix} \tilde{W}_x \\ W_y \end{pmatrix},$$

where E is an identity matrix.

The above transformation can be decomposed into two transformations:

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} P^{\mathrm{T}} & 0 \\ 0 & E \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{24}$$

$$Z_1 = \begin{pmatrix} \tilde{W}_x^{\mathrm{T}} & 0 \\ 0 & W_y^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}, \tag{25}$$

$$Z_2 = (\tilde{W}_x^{\mathrm{T}} \quad W_y^{\mathrm{T}}) \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}. \tag{26}$$

In the transformation (24), $\tilde{x} = P^{\mathrm{T}}x$, $\tilde{y} = y$, feature space B has not happened to change, but feature space A has become a new m-dimensional feature subspace. Since the column vectors of P are eigenvectors corresponding to nonzero eigenvalues of S_{xx}, so the transformation is called K-L transformation and in the K-L transformation space the covariance of the training sample is $\tilde{S}_{xx} = P^{\mathrm{T}} S_{xx} P$. So it is obvious that the optimal solution $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_d$; $\beta_1, \ldots, \beta_d$ $(d < rank(\tilde{S}_{xy}))$ which is determined by *Model* 3 is the CPV based on criterion $\tilde{J}(\tilde{\alpha}, \beta)$ in the K-L transformation space $\mathbb{R}^m$ and feature space B.

From the above analysis we can induce the extraction process of CCDF under the condition of singularity (consider the common situation, where S_{xx} and S_{yy} are both singular):

- The dimension of high-dimensional primitive sample can be reduced to $rank(S_{xx})$ and $rank(S_{yy})$ by the K-L transformation.
- Use Eq. (25) or (26) to extract the CCDF in the transformed feature spaces.

Above theoretically discuss the problem of solving the CPV when both S_{xx} and S_{yy} are singular. In practice, especially aiming to the problem of pattern classification, sometimes in order to improve the speed of feature extraction, during the step of PCA (or K-L transformation) we shall select eigenvectors corresponding to comparatively larger nonzero eigenvalues of S_{xx} (or S_{yy}) to compose transformation matrix W_{PCA}. However some eigenvectors corresponding to too small nonzero eigenvalues will be given up. So the obtained CCDF will not influence the classification effects at all, because in pattern classification too small nonzero eigenvalue probably include more interferential information which will result in overfitting problem [17].

4. Experiments and analysis

4.1. Experiment on CENPARMI handwritten numerals database

The goal of this experiment is to test the validity of the algorithm with the big sample proposed in this paper. The Concordia University CENPARMI database of handwritten Arabic numerals, popular in the world, is adopted. In this database, there are 10 class, i.e. 10 digits (from 0 to 9), and 600 sample for each. The training samples and the testing samples are 4000 and 2000, respectively. Hu et al. [18] had done some preprocessing work and extracted four kinds of features as follows:

X^{G}: 256-dimensional Gabor transformation feature [19],
X^{L}: 21- dimensional Legendre moment feature [20],
X^{P}: 36-dimensional Seudo-Zernike moment feature [21],
X^{Z}: 30-dimensional Zernike moment feature [22].

4.1.1. Experiment on two features fusion

Combine any two features of the above four features in the original feature space, using the algorithm described in Section 2.3, we obtain the CPM W_1 and W_2, then extract the CCDF by FFS I and FFS II. The minimum-distance classifier is used for classification, and classification error rate is shown in Table 1.

Notice when obtaining the CPV, we should obtain the low-dimensional CPV at first, and then the high-dimensional one according to Eq. (11) or (12). For example, when we want to combine two features X^{G} and X^{P}, for getting the CPV with 36-dimension and 256-dimension, we only need to obtain the eigenvalue and the corresponding eigenvector from the 36-dimensional matrix $G_{X^{\mathrm{P}}}$.

To compare this algorithm with two existing feature level fusion methods [4,6], experimental results are given here. At first, each sample should be normalized. Then we group two sets of feature vectors into one union-vector according to the serial feature fusion described in Ref. [4], and combine two sets of feature vectors by a complex vector according to the parallel feature fusion described in Ref. [6]. Finally, we extract the combined features in combined space, namely the total scatter matrix of their training sample do K-L transformation. The classification is done by the minimum-distance classifier. The classification results are shown in Table 1.

In above four groups of features, difference in number of dimensions of different groups of feature vectors can be large. In order to analyze effectiveness of the pretreatment method described in Section 2.4, we have carried out comparison experiments for the arithmetic proposed in this paper. For example, for feature combination of groups of

Table 1
Classification error rates based on two features in the different feature fusion methods

Combine features	Serial fusion	Dimension	Parallel fusion	Dimension	FFS I	Dimension	FFS II	Dimension
X^G–X^L	0.1920	96	0.1925	48	0.1170	242	0.1290	85
X^G–X^P	0.2280	128	0.2285	80	0.2050	72	0.2181	31
X^G–X^Z	0.2295	116	0.2290	74	0.2230	60	0.2255	20
X^L–X^P	0.2400	102	0.2410	61	0.1810	72	0.2160	34
X^L–X^Z	0.2505	79	0.2505	59	0.2110	60	0.2312	30
X^P–X^Z	0.4760	51	0.4785	31	0.3215	56	0.3295	30

Table 2
Classification error rates based on two features by the pretreatment of Section 2.4

Combine features	FFS I	Dimension	FFS II	Dimension
$(X^G + X^L)/2$–$(X^G + X^L)/2$	0.1175	376	0.1415	172
$(X^G + X^P)/2$–$(X^G + X^P)/2$	0.1430	292	0.1765	138
$(X^G + X^Z)/2$–$(X^G + X^Z)/2$	0.1500	286	0.1800	143
$(X^L + X^P)/2$–$(X^L + X^P)/2$	0.1400	156	0.1955	78
$(X^L + X^Z)/2$–$(X^L + X^Z)/2$	0.1495	150	0.2100	73

Table 3
Classification error rates based on the single feature in Gabor feature and Legendre feature

Single feature	Ref. [23]			Primitive feature
	ULDA	FSLD	Yong Xu	
X^G	0.199	0.274	0.198	0.269
X^L	0.141	0.270	0.150	0.479

X^G and X^P, we first combine corresponding two groups of feature vectors according to serial mode to a feature vector of 286-dimension. Then, by taking half of each feature vector, two groups of 143-dimensional feature vectors are formed. On these two groups of feature vectors, the arithmetic described in this paper is then used to extract CCDF. For this combination mode, we briefly record as: $(X^G + X^P)/2 - (X^G + X^P)/2$. In the experiments, we adopt minimum-distance classifier and corresponding experiment results are shown in Table 2 .

The advantage of this feature fusion, we list experimental results on the same database, based on the single feature for recognition in recent years, in Table 3. Such as the ULDA and FSLDA algorithm proposed by Yong Xu et al. [23]. In addition, we also give the recognition results based on primitive features. The experiments are all done using the minimal-distance classifier.

From Tables 1–3, we can see that, the recognition error rate of our method (FFS I and FFS II) is lower than that of the serial and parallel fusion method. When combining the Gabor feature and Legendre feature, FFS I and FFS II perform better than the other methods which are based on single feature. This suggests that the algorithm proposed in this paper is an efficient feature fusion method. In addition, we can see from the Tables 1–2 that, combining different features will affect the recognition rate. So the feature selection before fusion is of most importance.

Table 4
Classification error rates based on multi-features in FFS I and FFS II

Combine features	FFS I	Dimension	FFS II	Dimension
X^G–$(X^L + X^P)$	0.0820	300	0.0925	151
X^G–$(X^L + X^Z)$	0.0810	292	0.0935	132
$(X^G + X^P)$–$(X^L + X^Z)$	0.0830	302	0.0920	118
$(X^G + X^Z)$–$(X^L + X^P)$	0.0790	304	0.0895	140

It can be seen from Tables 1–2 that before fusion of X^G or X^L with X^P or X^Z, pretreatment (Section 2.4) is first carried out, and then extraction of CCDF, improving classification results to different extents. Take X^G with X^Z as an example, under two feature fusion strategies (FFS I and FFS II), pretreated classification error rate was lowered by 7.3% and 4.55%, respectively, showing effectiveness of the pretreatment mode described in Section 2.4. Besides, we can also see that no obvious change occurs in classification results after same pretreatment of X^G and X^L. In our opinion, numbers of dimension of two groups of feature vectors differ not much, CCDF can be directly extracted; pretreatment in Section 2.4 is suitable only when these differ greatly.

4.1.2. Experiment on multi-features fusion

This experiment gives results of several groups of features participating in the fusion at the same time. For example, for fusion of three groups of features X^G, X^L and X^Z, one combination mode is: first combine X^L with X^Z per serial mode to form one group of feature vectors, and then combine with the other X^G; CCDF is extracted according to arithmetic of this paper (Section 2.3), for classification. This combination mode is abbreviated as X^G–$(X^L + X^Z)$. In Table 4, Figs. 1 and 2 provide results of experiment in which several groups of features participate in fusion in different

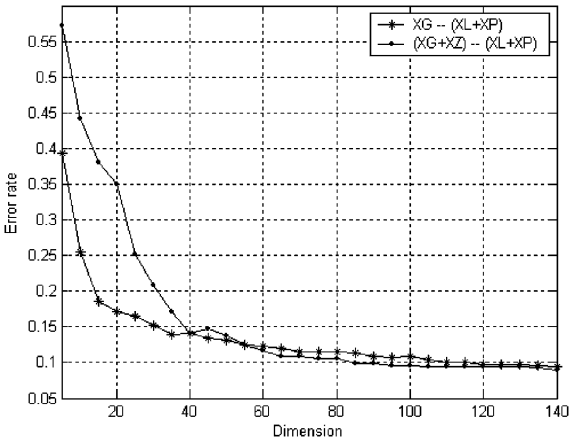

Fig. 1. Classification error rates of $X^G - (X^L + X^P)$ and $(X^G + X^Z) - (X^L + X^P)$ in FFS II.

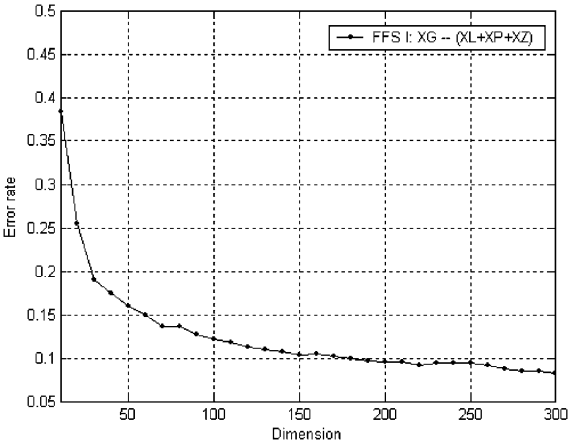

Fig. 3. Classification error rate of $X^G - (X^L + X^P + X^Z)$ in FFS I.

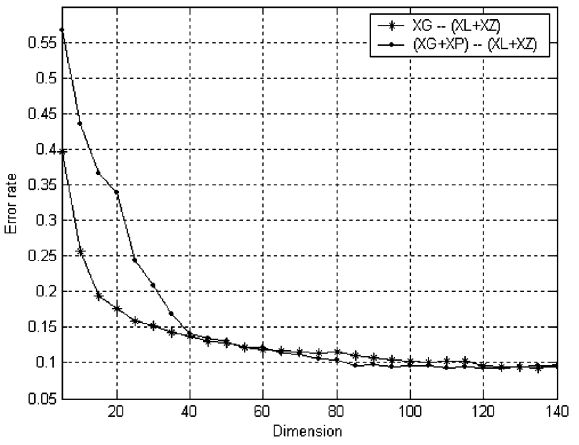

Fig. 2. Classification error rates of $X^G - (X^L + X^Z)$ and $(X^G + X^P) - (X^L + X^Z)$ in FFS II.

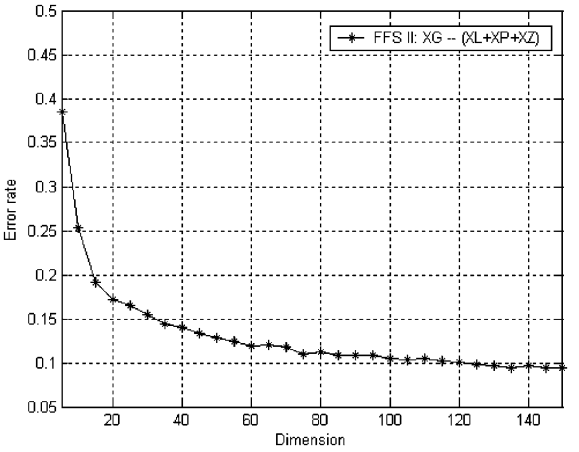

Fig. 4. Classification error rate of $X^G - (X^L + X^P + X^Z)$ in FFS II.

modes. All these experiments have been performed under minimum-distance classifier.

In all above experiments, when arithmetic of this paper was used to obtain CDV, the condition that are S_{xx} and S_{yy} both positive definite mentioned in Section 2.2 is satisfied, therefore, CCDF can be directly extracted using arithmetic procedure described in Section 2.3. Below we give other combination forms of four groups of features, e.g. $X^G - (X^L + X^P + X^Z)$, thereby singular S_{yy} arises, requiring arithmetic principle provided by Section 3. Assume notification of training sample space formed by X^G and $(X^L + X^P + X^Z)$ as $A = \{x \mid x \in \mathbb{R}^{256}\}$ and $B = \{y \mid y \in \mathbb{R}^{187}\}$, since $r = rank(S_{yy}) = 179$, we know that S_{yy} is singular. According to discussion in Section 3.2, we first use K-L transform to reduce 187-dimension of original feature to 179-dimension to form a new training sample space $\tilde{B} = \{\tilde{y} \mid \tilde{y} \in \mathbb{R}^{179}\}$, then the arithmetic of Section 2.3 is used to extract CCDF on training sample spaces A and $\tilde{B}$. Minimum-distance classifier is still used. Refer to Figs. 3and 4 for recognition results.

Results of above experiments show that after participation of several groups of features in fusion, due to increased information, extracted CCDF contains more effective discriminant information, so that recognition results are greatly improved. Under the two fusion strategies proposed in this paper, i.e. FFS I and FFS II, minimum-distance classifier is used and optimal recognition rate can reach 92% and 91%, respectively. In addition, classification error rate falls very quickly. Besides, it can be seen from experiment results that recognition result of FFS I is slightly better than that of FFS II. However, CCDF extracted by FFS II has advantage in number of dimension.

Fig. 5. Typical example with 11 face images for one person in the Yale database.

Fig. 6. A original image, low-frequency image and high-frequency image of double wavelet transform.

4.2. Experiment on Yale face image database

The Yale face database is adopted by this experiment. There are 15 persons who have 11 facial images respectively, so totally there are 165 images. The size of each image is 120 × 91 with 256 gray levels per pixel. These images are taken from different angle of view, having change of expression and illumination, and parts of the images are not integral. Fig. 5 shows a typical example of images for one person.

In this experiment, we use the first five images of each person for training and the remaining six for testing. Thus the total of training samples and testing sample are 75 and 90, respectively.

Firstly we perform double wavelet transform to original images using Daubechies orthonormal wavelet (Fig. 6 shows a original image, low-frequency image and high-frequency image of double wavelet transform), extract feature vectors of the low-frequency images to make the first feature space of samples $A = \{x \mid x \in \mathbb{R}^{690}\}$; then extract the singular value feature vectors of the original image to make the second feature space of samples $B = \{y \mid y \in \mathbb{R}^{91}\}$; finally we combine the two groups of feature.

The reason why we combine two groups of features above is that the low-frequency sub-images include more shape information, while the singular values that are extracted from the original images include more texture information. So the mutual complement of the two groups of features will be in favor of extracting effective CCDF.

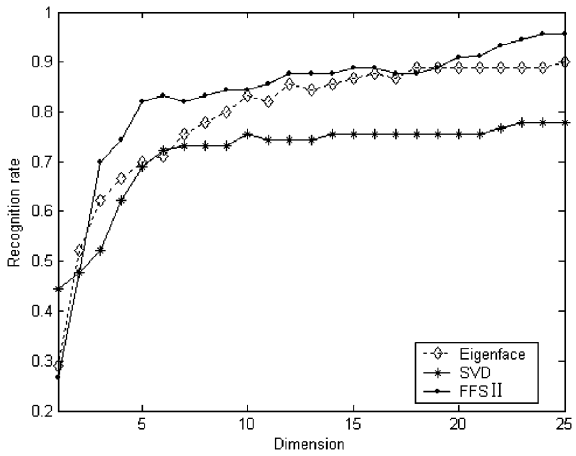

Fig. 7. Recognition rates of eigenface, SVD method and FFS II in nearest-neighbor classifier.

In this experiment, the ranks of S_{xx} and S_{yy} are computed first, and they are both equal to 74. Then the theory of SVD is used to computed the 30 orthonormal feature vectors which correspond to the nonzero feature values of S_{xx} and S_{yy}. Thus the transformation matrices P and Q are made up of those obtained feature vectors. Using K-L transform $\tilde{x} = P^{\mathrm{T}}x$ and $\tilde{y} = Q^{\mathrm{T}}y$, the low-frequency image vectors of $30 \times 23 = 690$dimension will be reduced to $m_1 = 30$ dimension and 91 dimension singular-value feature vectors of original image will be reduced to $m_2 = 30$ dimension. In the two feature sub-spaces after transformation, the two groups of features will be normalized, and in this paper the algorithm are used to obtain the CPM. Using FFS II to extract the CCDF and using the nearest-neighbor classifier and the recognition results is shown in Fig. 7.

In order to explain the validity of combined feature extraction, the Eigenface (PCA) [24] method and the SVD method [25] based on single feature are given out. Recognition result is also given respectively in Fig. 7 by taking the nearest-neighbor classifier.

Table 5
The recognition rates of Eigenface, SVD method, FFS I, and FFS II in different classifier

Classifier	Eigenface	SVD method	FFS I	FFS II
Minimal-distance	0.8556	0.5333	0.9444	0.9333
Nearest-neighbor	0.9111	0.7889	0.9667	0.9556

Moreover, for all the above methods, we also show the result of recognition in different classifier. The recognition result is shown in Table 5.

We can learn from the Fig. 7 and Table 5 that the classified result after adopting combined feature extraction is greatly improved than that of adopting a single feature extraction. Recognition rate by FFS I is above 96%, and it goes beyond that of PCA method 5%, that of SVD method 26%. So we can conclude that the combined CCDF has more powerful recognition ability, and it is an efficient combined feature extraction method. Besides, the results of experiment also indicate that the algorithm of this paper is not so sensitive to the illumination and expression comparing to other methods.

5. The analysis of the algorithm validity

In the theory of pattern recognition, the common principle of feature extraction is the smaller statistical correlation of selected features the better. The extracted features are better uncorrelated. A method of the uncorrelated optimal discriminant vectors is proposed by Zhong Jin et al. [26,27], and has been applied to the realms of face recognition and character recognition, which get a good performance. The essence of the theory is that the components of discriminant vectors are uncorrelated, and that projective vectors are orthonormal about the total scatter matrix.

From Theorem 2 and Corollary 2, the CPV are orthonormal about S_{xx}, S_{yy} and M, respectively, i.e. the components of CCDF are uncorrelated. So, this projective transform is optimal.

Commonly, the time taken greatly depends on the computational process of projective vectors, using algebraic method to extract discriminant feature. When the rank of matrix is very great, the computation of eigenvalues and eigenvectors would be time-consuming. In the same pattern, supposing that the feature vectors $x \in R^p$ and $y \in R^q$, projective vectors of Ref. [4] has been done in real vector space of $(p+q)$-dimension, and one of Refs. [5,6] has been done in complex vector space of $\max(p, q)$-dimension, but one of our methods has been only done in real vector space of $\min(p, q)$-dimension. When p and q are large, the advantage is obvious in the computational speed. For example, when we want to combine two features X^{G} and X^{Z} in Section 4.1, three kinds of combined methods choose the following dimension: 286, 256 and 30.

Furthermore, the linear discriminant analysis based on Fisher criterion(FLDA) is one of the best effective ways in feature extraction and recognition. The FLDA is also a special situation of CCA and may be solved by the theory, some algorithms based on Fisher criterion can be translated into the method presented in this paper (we should discuss in other paper). So applying CCA on pattern recognition is more general and has more potential development.

6. Conclusion

In this paper the idea of CCA is applied to feature fusion and image recognition for the first time. A new method feature fusion is proposed, which uses correlation feature of two groups of feature as effective discriminant information, so it is not only suitable for information fusion, but also eliminates the redundant information within the features. This offers a new way of applying two groups of feature fusion to discrimination.

The theory and method of using CCA in image recognition is discussed. The problem of the CPV can be solved when two total scatter matrixes are singular, such that it can be adapted to cases of high-dimensional space and small sample size, so the applicable range of CCA is extended in theory.

Moreover, comparison between the method proposed in this paper and two existing feature fusion methods is made on theory and the inherent essence of using this method in recognition is put forward. From the experimental results on big sample and small sample, extracting CCDF can realize the reduced primitive feature dimension, and is good at classifying performance reflecting the image's essential feature. The method in this paper is a good approach for information fusion on feature level. We should improve this method.

At last, it should be noticed that the conclusion obtained in Section 3, which is perfection and development of CCA, is completely suitable for other fields using CCA.

Acknowledgements

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region under Earmarked Research Grant (project no. CUHK4185/00E).

References

[1] Y.S. Huang, C.Y. Suen, Method of combining multiple experts for the recognition of unconstrained handwritten numerals, IEEE Trans. Pattern Anal. Mach. Intell. 7 (1) (1995) 90–94.
[2] A.S. Constantinidis, M.C. Fairhurst, A.F.R. Rahman, A new multi-expert decision combination algorithm and its application to the detection of circumscribed masses in digital mammograms, Pattern Recognition 34 (8) (2001) 1528–1537.
[3] X.-Y. Jin, D. Zhang, J.-Y. Yang, Face recognition based on a group decision-making combination approach, Pattern Recognition 36 (7) (2003) 1675–1678.

[4] 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.
[5] J. Yang, J.-Y. Yang, Generalized K-L transform based combined feature extraction, Pattern Recognition 35 (1) (2002) 295–297.
[6] J. Yang, J.Y. Yang, D. Zhang, J.F. Lu, Feature fusion: parallel strategy vs. serial strategy, Pattern Recognition 36 (6) (2003) 1369–1381.
[7] X.T. Zhang, K.T. Fang, Multivariate Statistical Introduction, Sciences Press, Beijing, 1999.
[8] W.S. Sun, L.X. Chen, Multivariate Statistical Analysis, Higher Education Press, Beijing, 1994.
[9] M. Borga, Learning multidimensional signal processing, Linköping studies in science and technology, Dissertations, vol.531, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 1998.
[10] S.-J. Yu, Direct blind channel equalization via the programmable canonical correlation analysis, Signal Process. 81 (2001) 1715–1724.
[11] C. Fyfe, P.L. Lai, Canonical correlation analysis neural networks, International Conference on Pattern Recognition, vol. 2, Barcelona, 2000, pp. 977–980.
[12] H.C. Choi, R.W. King, Speaker adaptation through spectral transformation for HMM based speech recognition, IEEE Int. Symp. Speech Image Process. Neural Networks 2 (1994) 686–689.
[13] D. Weenink, Canonical Correlation Analysis, Institute of Phonetic Sciences, University of Amsterdam, Proceedings, vol. 25, 2003, pp. 81–99.
[14] H. Hotelling, Relations between two sets of variates, Biometrika 8 (1936) 321–377.
[15] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 36 (9) (2003) 1961–1971.
[16] J. Friedman, Regularized discriminant analysis, J. Am. Statist. Assoc. 84 (405) (1989) 65–175.
[17] C.J. Liu, H. Wechsler, Robust coding schemes for indexing and retrieval from large face databases, IEEE Trans. Image Process. 9 (1) (2000) 132–137.
[18] Z.S. Hu, Z. Lou, J.Y. Yang, K. Liu, C.Y. Suen, Handwritten digit recognition basis on multi-classifier combination, Chinese J. Comput. 22 (4) (1999) 369–374.
[19] H. Yoshihiko, et al., Recognition of handwriting numerals using Gabor features, Proceedings of the Thirteenth ICPR, pp. 250–253.
[20] S.X. Liao, M. Pawlak, On image analysis by moments, IEEE Trans. Pattern Anal. Mach. Intell. 18 (3) (1996) 254–266.
[21] R.R. Bailey, S. Mandyam, Orthogonal moment feature for use with parametric and non-parametric classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 18 (4) (1996) 389–398.
[22] K. Alireza, H. Yawhua, Invariant image recognition by Zernike moments, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 489–497.
[23] Y. Xu, J.-Y. Yang, Z. Jin, A novel method for Fisher discriminant analysis, Pattern Recognition 37 (2) (2004) 381–384.
[24] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1991) 71–86.
[25] Z.Q. Hong, Algebraic feature extraction of image for recognition, Pattern Recognition 24 (3) (1991) 211–219.
[26] Z. Jin, J.-Y. Yang, Z.-M. Tang, Z.-S. Hu, A theorem on the uncorrelated optimal discriminant vectors, Pattern Recognition 34 (7) (2001) 2041–2047.
[27] Z. Jin, J.Y. Yang, Z.S. Hu, Face recognition based on uncorrelated discriminant transformation, Pattern Recognition 34 (7) (2001) 1405–1416.

About the Author—QUAN-SEN SUN is an associate professor in the Department of Mathematics at Jinan University. At the same time, he is working for his Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST). His current interests include pattern recognition, image processing, computer vision and data fusion.

About the Author—SHENG-GEN ZENG received his Ph.D. degree in computer specialty from Nanjing University of Science and Technology of China in 2004. His current interests include pattern recognition, image processing and remote sensing image.

About the Author—YAN LIU received his M.S. degree in computer specialty from Nanjing University of Science and Technology of China in 2004. His current interests include pattern recognition, image processing.

About the Author—PHENG-ANN HENG received his Ph.D. degree in computer science from the Indiana University of USA in 1992. He is now a professor in the Department of Computer Science and Engineering at the Chinese University of Hong Kong (CUHK). He is Director of Virtual Reality, Visualization and Imaging Research Centre at CUHK. His research interests include virtual reality applications in medicine, scientific visualization, 3D medical imaging, user interface, rendering and modeling, interactive graphics and animation.

About the Author—DE-SHEN XIA received his Ph.D. degree in Pattern Recognition and Intelligent Systems from the Rouen University of France in1987. He is now the honorary professor at ESIGELEC, France, and the professor and Ph.D. supervisor in the Department of Computer Science at NUST. He is the Director of the Laboratory of Image Processing, Analysis and Recognition at the NUST. His research interests are in the domain of image processing, remote sensing, medical image analysis and pattern recognition.