PERGAMON

Pattern Recognition 36 (2003) 1369–1381

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Feature fusion: parallel strategy vs. serial strategy

Jian Yang[a,b,*], Jing-yu Yang[a], David Zhang[b], Jian-feng Lu[a]

[a] *Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China*
[b] *Biometrics Research Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong*

Received 18 January 2002; accepted 26 August 2002

Abstract

A new strategy of parallel feature fusion is introduced in this paper. A complex vector is first used to represent the parallel combined features. Then, the traditional linear projection analysis methods, including principal component analysis, K–L expansion and linear discriminant analysis, are generalized for feature extraction in the complex feature space. Finally, the developed parallel feature fusion methods are tested on CENPARMI handwritten numeral database, NUST603 handwritten Chinese character database and ORL face image database. The experimental results indicate that the classification accuracy is increased significantly under parallel feature fusion and also demonstrate that the developed parallel fusion is more effective than the classical serial feature fusion.

Keywords: Feature fusion; Feature extraction; Complex feature space; Principal component analysis (PCA); K–L expansion; Linear discriminant analysis (LDA); Character recognition; Face recognition

1. Introduction

In recent years, data fusion has been developed rapidly and widely applied in many areas, such as object tracking and recognition [1,2], pattern analysis and classification [3,4], image processing and understanding [5,6], and so on. In this paper, we pay our main attention to the data fusion techniques used for pattern classification problems.

In practical classification applications, if the number of classes and multiple feature sets for pattern samples are given, how to achieve a desirable recognition performance based on these sets of features is a very interesting problem. Generally speaking, there exist three popular schemes. In the first one, the information derived from multiple feature sets is assimilated and integrated into a final decision directly. This technique is generally referred to as *centralized data fusion* [2] or *information fusion* [7] and widely adopted in many pattern recognition systems [7,8]. In the second, the individual decisions are first made based on different feature sets, and then they are reconciled or combined into a global decision. This technique is generally known as *distributed data fusion* or *decision fusion* [2]. In the third scheme, the given multiple feature sets are used to produce new fused feature sets, which are more helpful in the final classification [5]. The technique is usually termed *feature fusion*.

As a matter of fact, feature fusion and decision fusion are two levels of data fusion. In some cases, they are involved in the same application system [4,9]. But, in recent years, the decision level fusion, represented by multi-classifier or multi-expert combination strategies, has been of major concern [10,11]. In contrast, the feature level fusion has probably not received the amount of attention it deserves.

However, feature level fusion plays a very important role in the process of data fusion. The advantage of feature level fusion lies in two aspects: firstly, it can derive the most discriminatory information from original multiple feature sets

[*] Corresponding author. Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China. Tel.: +86-25-431-6840; fax: +86-25-4315510.

E-mail addresses: csjyang@comp.polyu.edu.hk, tuqingh@mail.njust.edu.cn (J. Yang), yangjy@mail.njust.edu.cn (J.-y. Yang), csdzhang@comp.polyu.edu.hk (D. Zhang).

0031-3203/03/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(02)00262-5

involved in fusion; secondly, it is able to eliminate the redundant information resulting from the correlation between distinct feature sets and to make the subsequent decision in real time possible. In other words, feature fusion is capable of deriving and gaining the most effective and least-dimensional feature vector sets that benefit the final decision.

In general, the existing feature fusion techniques for pattern classification can be subdivided into two basic categories. One is feature selection based, and the other is feature extraction based. In the former, all feature sets are first grouped together and then a suitable method is used for feature selection. Zhang [12] presented a fused method based on dynamic programming, Battiti [13] gave a method using supervised neural network, and, recently, Shi and Zhang [14] provided a method based on support vector machines (SVM). In the latter, the multiple feature sets are combined into one set of feature vectors that are input into a feature extractor for fusion [15].

The classical method of feature combination is to group two sets of feature vectors into one union-vector (or supervector) [15]. Recently, we proposed a novel feature combination strategy [16]. Its idea is to combine two sets of feature vectors by a complex vector rather than a real union-vector. In this paper, the feature fusion method based on the union-vector is referred to as *serial feature fusion* and that based on the complex vector is called *parallel feature fusion*.

In this paper, our focus is on feature level fusion. We continue the work in [16] and further enrich the idea of feature combination and fusion. We specify the distinction between *feature combination* and *feature fusion* and introduce the notions of *serial feature fusion* and *parallel feature fusion* as well. The idea of parallel feature fusion is outlined as follows. The two given sets of original feature vectors are first used to form a complex feature vector space. Then, the traditional linear projection methods, including principal component analysis (PCA), K–L expansion and linear discriminant analysis (LDA), are generalized for feature extraction in the complex feature space.

Besides the description of feature fusion method and the related theoretical derivation, in this paper, much attention is concentrated on the experimental comparisons. We do experiments using three databases; CENPARMI handwritten numeral database, NUST603 handwritten Chinese character database and ORL face image database, where the first two experiments involve large sample size problems and the last one involves a small sample size problem. All experimental results indicate that the classification accuracy is increased significantly and the dimension of the feature vector is reduced largely under the parallel feature fusion. What is more is that the experimental results also demonstrate that the proposed parallel fusion outperforms the classical serial feature fusion.

The remainder of this paper is organized as follows. In Section 2, some concepts and notions relevant to feature fusion and parallel feature fusion are presented. In Section 3, the generalized linear projection analysis methods, generalized PCA, generalized KLE and generalized LDA are developed. In Section 4, some feature preprocessing techniques are given. In Section 5, we analyze and reveal the symmetry property of parallel feature fusion. In Section 6, the results of the three experiments are presented and the detailed experimental comparisons are discussed. Finally, a conclusion is given in Section 7.

2. Serial and parallel feature fusion strategy

Suppose A and B are two feature spaces defined on pattern sample space Ω. For an arbitrary sample $\xi \in \Omega$, the corresponding two feature vectors are $\alpha \in A$ and $\beta \in B$. The serial combined feature of ξ is defined by $\gamma = \binom{\alpha}{\beta}$. Obviously, if feature vector α is n-dimensional and β is m-dimensional, then the serial combined feature γ is $(n+m)$-dimensional. All serial combined feature vectors of pattern samples form a $(n+m)$-dimensional serial combined feature space.

Here, we intend to combine two feature vectors by a complex vector rather than a supervector. In more detail, let α and β be two different feature vectors of the same sample ξ, the complex vector $\gamma = \alpha + \mathrm{i}\beta$ (i is an imaginary unit) is used to represent the combination of α and β and is named parallel combined feature of ξ. Note that if the dimensions of α and β are not equal, pad the lower-dimensional one with zeros until its dimension is equal to the other ones' before combination. For example, if $\alpha = (a_1, a_2, a_3)^{\mathrm{T}}$, $\beta = (b_1, b_2)^{\mathrm{T}}$, β is first turned into $(b_1, b_2, 0)^{\mathrm{T}}$ and the resulting combination form is denoted by $\gamma = (a_1 + \mathrm{i}b_1, a_2 + \mathrm{i}b_2, a_3 + \mathrm{i}0)^{\mathrm{T}}$.

Let us define the parallel combined feature space on Ω as $C = \{\alpha + \mathrm{i}\beta \mid \alpha \in A,\ \beta \in B\}$. Obviously, it is an n-dimensional complex vector space, where $n = \max\{\dim A,\ \dim B\}$. In the space, the inner product is defined by

$$(X, Y) = X^{\mathrm{H}} Y, \tag{1}$$

where $X, Y \in C$, and H is the denotation of conjugate transpose.

The complex vector space defined by the above inner product is usually called unitary space.

In unitary space, the measurement (norm) can be introduced as follows:

$$\|Z\| = \sqrt{Z^{\mathrm{H}} Z} = \sqrt{\sum_{j=1}^{n} (a_j^2 + b_j^2)}, \tag{2}$$

where $Z = (a_1 + \mathrm{i}b_1, \ldots, a_n + \mathrm{i}b_n)^{\mathrm{T}}$.

Correspondingly, the distance (called unitary distance) between the complex vectors Z_1 and Z_2 is defined by

$$\|Z_1 - Z_2\| = \sqrt{(Z_1 - Z_2)^{\mathrm{H}} (Z_1 - Z_2)}. \tag{3}$$

If samples are directly classified based on the serial combined feature $\gamma = \binom{\alpha}{\beta}$ or the parallel combined feature $\gamma = \alpha + \mathrm{i}\beta$, and Euclidean distance is adopted in the serial combined feature space while unitary distance is used in parallel combined feature space, then it is obvious that the two kinds of combined features are equivalent in essence, i.e., they will result in the same classification results. However, the combined feature vectors are always high dimensional and contain much redundant information and some conflicting information which are unfavorable for recognition. Consequently, in general, we would rather perform the classification after the process of *feature fusion* than after the process of *feature combination*.

In our opinion, feature fusion includes feature combination but more than it. That is to say, the fusion is a process of reprocessing the combined features, i.e., after dimension reduction or feature extraction, the favorable discriminatory information is remained and at the same time the unfavorable redundant or conflicting information is eliminated. According to this opinion, there are two strategies of feature fusion based on two methods of feature combination:

1. *Serial feature fusion*: The serial feature fusion is a process of feature extraction based on the serial feature combination method, and the resulting feature is called a serial fused feature.
2. *Parallel feature fusion*: The parallel feature fusion is a process of feature extraction based on the parallel feature combination method, and the resulting feature is called parallel fused feature.

As we know, the parallel combined feature vector is a complex vector. Now the problem is, how to perform the feature extraction in complex feature space? In the following section, we will discuss the linear feature extraction techniques in complex feature space.

3. Generalized linear projection analysis

3.1. Basis concepts

In the unitary space, the between-class scatter (covariance) matrix, within-class scatter matrix and total-scatter matrix are, respectively, defined as follows:

$$S_b = \sum_{i=1}^{L} P(\omega_i)(m_i - m_0)(m_i - m_0)^{\mathrm{H}}, \tag{4}$$

$$S_w = \sum_{i=1}^{L} P(\omega_i) E\{(X - m_i)(X - m_i)^{\mathrm{H}}/\omega_i\}, \tag{5}$$

$$S_t = S_b + S_w = E\{(X - m_0)(X - m_0)^{\mathrm{H}}\}, \tag{6}$$

where L denotes the number of pattern class; $P(\omega_i)$ is the prior probability of class i; $m_i = E\{X/\omega_i\}$ is the mean vector of class i; $m_0 = E\{X\} = \sum_{i=1}^{m} P(w_i)m_i$ is the mean of all training samples.

From Eqs. (4)–(6), it is obvious that S_w, S_b and S_t are all semi-positive definite Hermite matrices. What is more, S_w and S_t are all positive definite matrix when S_w is nonsingular. In this paper, S_w is assumed to be nonsingular.

Lemma 1 (Xue-ren Ding and Miao-ke Cai [17]). *Each eigenvalue of Hermite matrix is a real number.*

Since S_w, S_b and S_t are all semi-positive definite Hermite matrices, it is easy to get the following conclusion:

Corollary 1. *The eigenvalues of S_w, S_b or S_t are all nonnegative real numbers.*

3.2. Generalized PCA

Taking S_t as a generation matrix, we present the K–L transform method in complex feature space. Suppose that the orthogonal eigenvectors of S_t are $\xi_1, \dots, \xi_n$, and the associated eigenvalues are $\lambda_1, \dots, \lambda_n$, which satisfy $\lambda_1 \geqslant \cdots \geqslant \lambda_n$. Choosing the first d-maximal eigenvectors $\xi_1, \dots, \xi_d$ as projection axes, the generalized K–L transform can be defined by

$$Y = \Phi^{\mathrm{H}} X \quad \text{where } \Phi = (\xi_1, \dots, \xi_d).$$

We also call it generalized principal component analysis (GPCA).

In fact, the classical PCA is only a special case of the GPCA. That is to say, the theory of PCA developed in complex feature space is more significant, and it undoubtedly suits the case in real feature space.

3.3. Generalized K–L expansion

Except for PCA, there are three typical K–L expansion methods [18], which include: extraction of discriminatory information contained in the class means (KLE1), optimal compression of the discriminatory information contained in the class means (KLE2), and extraction of discriminatory information contained in class-centralized vectors (KLE3). In fact, these three methods can be generalized to suit feature extraction in complex space. Now, taking KLE1 as example, we discuss it in detail.

Suppose that the orthogonal eigenvectors of S_w are $\xi_1, \dots, \xi_n$, and the associated eigenvalues are $\lambda_1, \dots, \lambda_n$. As S_w is nonsingular, it is positive definite, thus $\lambda_j > 0$, $j = 1, \dots, n$. Since S_b is semi-positive, we have $\xi_j^{\mathrm{H}} S_b \xi_j \geqslant 0$. So the physical meaning of criterion $J(\xi_j) = \xi_j^{\mathrm{H}} S_b \xi_j / \lambda_j$ is the same as that defined in Euclidian space. If $J(\xi_1) \geqslant \cdots \geqslant J(\xi_d) \geqslant \cdots \geqslant J(\xi_n)$ is satisfied, then $\xi_1, \dots, \xi_d$ are selected as the projection axes. This method is named GKLE1.

3.4. Generalized LDA

In unitary space, the Fisher discriminant criterion function can be defined by

$$J_f(\varphi) = \frac{\varphi^{\mathrm{H}} S_b \varphi}{\varphi^{\mathrm{H}} S_w \varphi}, \tag{7}$$

where φ is an n-dimensional nonzero vector.

Since S_w and S_b are all semi-positive definite, for any arbitrary φ, we have $\varphi^{\mathrm{H}} S_b \varphi \geqslant 0$ and $\varphi^{\mathrm{H}} S_w \varphi \geqslant 0$. Hence, the values of $J_f(\varphi)$ are all nonnegative real number. It means that the physical meanings of the Fisher criterion defined in Unitary space is the same as that defined in Euclidian space.

If S_w is nonsingular, it is easy to prove that Fisher criterion is equivalent to the following function:

$$J(\varphi) = \frac{\varphi^{\mathrm{H}} S_b \varphi}{\varphi^{\mathrm{H}} S_t \varphi}. \tag{8}$$

For convenience, in this paper, we use the above criterion function instead of Fisher criterion defined in Eq. (7).

Recently, Jin and Yang [19–21] suggested the theory on the uncorrelated linear discriminant analysis (ULDA), and used it to solve face recognition and handwritten digit recognition problems successfully. The most outstanding advantage of ULDA is that it can eliminate the statistical correlation between the components of pattern vector. In this paper, we try to further extend Jin's theory and enable it to suit feature extraction in the combined complex feature space (unitary space). Now, we describe the uncorrelated discriminant analysis in unitary space in detail.

The uncorrelated discriminant analysis aims to find a set of vectors $\varphi_1, \ldots, \varphi_d$, which maximizes the criterion function $J(\varphi)$ under the following conjugate orthogonality constraints:

$$\varphi_j^{\mathrm{H}} S_t \varphi_i = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad i, j = 1, \ldots, d. \tag{9}$$

More formally, the uncorrelated discriminant vectors $\varphi_1, \ldots, \varphi_d$ can be chosen in this way: φ_1 is selected as a Fisher optimal projective direction. And after determining the first k discriminant vectors $\varphi_1, \ldots, \varphi_k$, the $(k+1)$th discriminant vector φ_{k+1} is the optimal solution of the following optimization problem:

$$\text{Model 1} \quad \begin{cases} \max(J(\varphi)) \\ \varphi_j^{\mathrm{H}} S_t \varphi = 0, \ j = 1, \ldots, k, \\ \varphi \in C^n \end{cases}$$

where C^n denotes the n-dimensional unitary space.

Now, we discuss how to find the optimal discriminant vectors. Since S_b, S_t are all Hermite matrices and S_t is positive definite, according to the theory in document [17], it is easy to draw the following conclusion:

Theorem 1. *Suppose that S_t is nonsingular, there exist n eigenvectors $X_1, \ldots, X_n$ corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$ of the eigenequation $S_b X = \lambda S_t X$, such that*

$$X_i^{\mathrm{H}} S_t X_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad i, j = 1, \ldots, n \tag{10}$$

and

$$X_i^{\mathrm{H}} S_b X_j = \begin{cases} \lambda_i, & i = j, \\ 0, & i \neq j, \end{cases} \quad i, j = 1, \ldots, n. \tag{11}$$

The vectors $X_1, \ldots, X_n$ satisfying the constraint (10) are called S_t-*orthogonal.*

Since S_b is semi-positive definite, from Theorem 1, it is not hard to get the following corollaries:

Corollary 2. *The generalized eigenvalues $\lambda_1, \ldots, \lambda_n$ of $S_b X = \lambda S_t X$ are all nonnegative real numbers, in which there exist q positive ones, where $q = \mathrm{rank}(S_b)$.*

Corollary 3. *The values of criterion function $J(X_j) = \lambda_j$, $j = 1, \ldots, n$.*

Corollary 4. *The eigenvectors $X_1, \ldots, X_n$ of $S_b X = \lambda S_t X$ are linearly independent, and $C^n = \mathrm{span}\{X_1, \ldots, X_n\}$.*

Without loss of generality, suppose that the eigenvalues of $S_b X = \lambda S_t X$ satisfy $\lambda_1 \geqslant \cdots \geqslant \lambda_n$.

Theorem 2. *If the first k discriminant vectors are chosen as $\varphi_1 = X_1, \ldots, \varphi_k = X_k$, then, the $(k+1)$th optimal discriminant vector φ_{k+1} can be selected as X_{k+1}.*

The proof of Theorem 2 is given in Appendix A.

Since the first discriminant vector is selected as X_1 in advance, Theorem 2 indicates that in the unitary space, the uncorrelated optimal discriminant vectors can be selected as $X_1, \ldots, X_d$, which are the S_t-orthogonal eigenvectors corresponding to the first d-maximal eigenvalues of $S_b X = \lambda S_t X$. From Corollary 1 and the physical meanings of Fisher criterion, the total number of effective discriminant vectors is at most $q = \mathrm{rank}(S_b) \leqslant L - 1$, where L is the pattern class number.

The detailed algorithm for the calculation of $X_1, \ldots, X_d$ is described as follows:

First, obtain the prewhitening transformation matrix W, such that $W^{\mathrm{H}} S_t W = I$. In fact, $W = U \Lambda^{-1/2}$, where $U = (\xi_1, \ldots, \xi_n)$, $\Lambda = \mathrm{diag}(a_1, \ldots, a_n)$, $\xi_1, \ldots, \xi_n$ are eigenvectors of S_t, and $a_1, \ldots, a_n$ are associated eigenvalues.

Next, let $\tilde{S}_b = W^{\mathrm{H}} S_b W$, and calculate its orthonormal eigenvectors $\xi_1, \ldots, \xi_n$. Suppose the associated eigenvalues satisfy $\lambda_1 \geqslant \cdots \geqslant \lambda_n$, then the optimal projection axes are $X_1 = W\xi_1, \ldots, X_d = W\xi_d$.

In unitary space, since the uncorrelated optimal discriminant vectors $X_1, \ldots, X_d$ $(d \leqslant q)$ satisfy the constraint (10) and (11), and λ_j is a positive real number, the physical meanings of the fused feature being projected onto X_j is specific, that is, the between-class scatter is λ_j, and the total scatter (i.e., the sum of the between-class scatter and the within-class scatter) is 1.

In unitary space, the uncorrelated discriminant vectors $X_1, \ldots, X_d$ form the following transformation:

$$Y = \Phi^{\mathrm{H}} X \quad \text{where } \Phi = (X_1, \ldots, X_d), \tag{12}$$

which is used for feature extraction in parallel combined feature space.

As a comparison, the method addressed in document [19,20] is only a special case of ours in this paper. That is to say, the theory of the uncorrelated discriminant analysis developed in complex vector space is more significant, and it undoubtedly suits the case in real vector space.

4. Feature preprocessing method

The difference of feature extraction or measurement might lead to the numerical unbalance between the two features α and β of the same pattern. For instance, given two feature vectors, $\alpha=(10,11,9)^{\mathrm{T}}$ and $\beta=(0.1,0.9)^{\mathrm{T}}$, corresponding to one sample, assume that they are combined as $\gamma=\alpha+\mathrm{i}\beta$ or $\gamma=\binom{\alpha}{\beta}$, which implies that the feature α plays a more important role than β in the process of fusion. Consequently, in order to counteract the numerical unbalance between features and gain satisfactory fusion performance, our suggestion is to initialize the original feature α and β, respectively, before the combination.

What is more, when the dimensions of α and β are unequal, after being initialized, we think the higher-dimensional one is still more powerful than the lower-dimensional one. The reason is that in the course of linear feature extraction after combination, the higher-dimensional one plays a more important role in the scatter matrices. So, in order to eliminate the unfavorable effect resulting from an unequal dimension, our suggestion is to adopt the weighted combination form. The serial combination is formed by $\gamma=\binom{\alpha}{\theta\beta}$ or $\gamma=\binom{\theta\alpha}{\beta}$, while the parallel combination is formed by $\gamma=\alpha+\mathrm{i}\theta\beta$ or $\theta\alpha+\mathrm{i}\beta$, where the weight θ is called a combination coefficient.

It is easy to prove that the weighted combined feature satisfies the following properties:

Property 1. If $\theta\neq 0$, the parallel combined feature $\gamma=\alpha+\mathrm{i}\theta\beta$ is equivalent to $\gamma=(1/\theta)\alpha+\mathrm{i}\beta$, and the serial combined feature $\gamma=\binom{\alpha}{\theta\beta}$ is equivalent to $\gamma=\binom{(1/\theta)\alpha}{\beta}$.

Property 2. While $\theta\to 0$, the fused feature $\gamma=\alpha+\mathrm{i}\theta\beta$ is equivalent to the single feature α; while $\theta\to\infty$ $(\theta\neq\infty)$, $\gamma=\alpha+\mathrm{i}\theta\beta$ is equivalent to the single feature β.

Two feature-preprocessing methods are introduced, respectively, as follows:

Preprocessing method I

Step 1: Let $\bar{\alpha}=\alpha/\|\alpha\|$, $\bar{\beta}=\beta/\|\beta\|$, i.e., turn α and β into unit vectors, respectively.

Step 2: Suppose the dimension of α and β is n and m, respectively; if $n=m$, let $\theta=1$; Otherwise, suppose $n>m$, let $\theta=n^2/m^2$, and the parallel combination form is $\gamma=\bar{\alpha}+\mathrm{i}\theta\bar{\beta}$ while the serial combination form is $\gamma=\binom{\bar{\alpha}}{\theta\bar{\beta}}$.

In Step 2, the evaluation of the combination coefficient $\theta=n^2/m^2$ is attributed to the following reason. When the length of two feature vectors is unequal, since the size of scatter matrix generated by feature vector α is $n\times n$, and the size of scatter matrix generated by β is $m\times m$, combination coefficient θ is considered to be the square of n/m.

Preprocessing method II

First, respectively, initialize α and β by

$$Y=\frac{X-\mu}{\sigma},$$

where X denotes the sample feature vector, μ denotes the mean vector of training samples, $\sigma=\frac{1}{n}\sum_{j=1}^{n}\sigma_j$, where n is the dimension of X and σ_j is the standard deviation of the j-th feature component of training samples.

However, if the above feature initialization method is employed, it is more difficult to evaluate the combination coefficient θ when the dimensions of α and β are unequal. Here, we give the selection scope by experience. Suppose the combination form of α and β is denoted by $\gamma=\alpha+\mathrm{i}\theta\beta$ or $\binom{\alpha}{\theta\beta}$, the dimensions of α and β are n and m, respectively, and $n>m$, let $\delta=n/m$. Then θ can be selected between δ and δ^2.

How to determine a proper combination coefficient θ for the optimal fusion performance is still a problem that deserves further investigation.

5. Symmetry property of parallel feature fusion

For parallel feature combination, two kinds of feature vectors α and β of the sample can be formed by $\alpha+\mathrm{i}\beta$ or $\beta+\mathrm{i}\alpha$. But, based on this two combination forms $\alpha+\mathrm{i}\beta$ and $\beta+\mathrm{i}\alpha$, after feature extraction via generalized linear projection analysis, there is still a question: are the final classification results identical under the same classifier? If the results are identical, we think that the parallel feature fusion satisfies a property of symmetry. That is to say, the result of parallel fusion is independent of the sequence of feature in combination. This is expected. Otherwise, if parallel feature fusion does not satisfy this property, i.e., different sequence of combined features induces different classification results which makes the problem more complicated. Fortunately, we can prove theoretically that the parallel feature fusion satisfies a desirable property of symmetry. Now, taking the GPCA (one of the feature extraction techniques mentioned above) as example, we give the proof of symmetry in parallel feature fusion.

Suppose two feature spaces defined on pattern sample space Ω are, respectively, defined by $C_1=\{\alpha+\mathrm{i}\beta\mid\alpha\in A,\beta\in B\}$ and $C_2=\{\beta+\mathrm{i}\alpha\mid\alpha\in A,\beta\in B\}$.

Lemma 2. *In unitary space, construct two matrices* $H(\alpha,\beta)=(\alpha+\mathrm{i}\beta)(\alpha+\mathrm{i}\beta)^{\mathrm{H}}$ *and* $H(\beta,\alpha)=(\beta+\mathrm{i}\alpha)(\beta+\mathrm{i}\alpha)^{\mathrm{H}}$, *then* $H(\beta,\alpha)=\overline{H(\alpha,\beta)}$, *where α and β are n-dimensional real vectors.*

Proof. Suppose $\alpha=(a_1,\dots,a_n)^{\mathrm{T}}$, $\beta=(b_1,\dots,b_n)^{\mathrm{T}}$, then

$$[H(\alpha,\beta)]_{kl}=(a_k+\mathrm{i}b_k)\overline{(a_l+\mathrm{i}b_l)}=(a_ka_l+b_kb_l)+\mathrm{i}(a_lb_k-a_kb_l),$$

$$[H(\beta,\alpha)]_{kl} = (b_k + \mathrm{i}a_k)\overline{(b_l + \mathrm{i}a_l)} = (a_k a_l + b_k b_l) - \mathrm{i}(a_l b_k - a_k b_l).$$

So $[H(\beta,\alpha)]_{kl} = \overline{[H(\alpha,\beta)]}_{kl}$.
Hence $H(\beta,\alpha) = \overline{H(\alpha,\beta)}$. □

Suppose that the within-class, between-class and total-scatter matrices in a combined feature space C_i $(i=1,2)$ are defined by S_w^i, S_b^i and S_t^i $(i = 1,2)$, respectively. By Lemma 2, it is easy to prove that they satisfy the following properties:

Property 3. $S_w^2 = \overline{S_w^1}$; $S_b^2 = \overline{S_b^1}$; $S_t^2 = \overline{S_t^1}$.

Property 4. Suppose that ξ is the eigenvector of S_t^1 (S_w^1 or S_b^1) corresponding to the eigenvalue λ, then $\bar{\xi}$ is the eigenvector of S_t^2 (S_w^2 or S_b^2) corresponding to the same eigenvalue λ.

Proof. $S_t^1\xi = \lambda\xi \Rightarrow \overline{S_t^1\xi} = \overline{\lambda\xi} \Rightarrow \overline{S_t^1}\,\bar{\xi} = \bar{\lambda}\bar{\xi}$.
By Property 3 and Corollary 1, we have $S_t^2 = \overline{S_t^1}$ and $\bar{\lambda} = \lambda$, hence $S_t^2\bar{\xi} = \lambda\bar{\xi}$. □

By Property 4, we can draw the following conclusion. If $\xi_1,\ldots,\xi_d$ are the projection axes of GPCA in combined feature space C_1, then, $\overline{\xi_1},\ldots,\overline{\xi_d}$ are the projection axes of GPCA in combined feature space C_2.

Lemma 3. *In combined feature space C_1, if the projection of vector $x + \mathrm{i}y$ onto the axis ξ* is *$p + \mathrm{i}q$, then in combined feature space C_2, the projection of sample $y + \mathrm{i}x$ onto the axis $\bar{\xi}$ is $q + \mathrm{i}p$.*

The proof of Lemma 3 is given in Appendix B.

According to Lemma 3, it is not difficult to draw the conclusion.

Property 5. In combined feature space C_1, $x + \mathrm{i}y \xrightarrow{\text{GPCA}} u + \mathrm{i}v$
In combined feature spaceC_2, $y + \mathrm{i}x \xrightarrow{\text{GPCA}} v + \mathrm{i}u$

In unitary space, by the norm defined in Eq. (2), we have $\|u + \mathrm{i}v\| = \|v + \mathrm{i}u\|$. By the unitary distance defined in Eq. (3), if $Z_1^1 = u_1 + \mathrm{i}v_1$, $Z_2^1 = u_2 + \mathrm{i}v_2$; $Z_1^2 = v_1 + \mathrm{i}u_1$, $Z_2^2 = v_2 + \mathrm{i}u_2$, then $\|Z_1^1 - Z_2^1\| = \|Z_1^2 - Z_2^2\|$. That is to say, the distance between two complex vectors depends on the values of their real part and the imaginary part but is independent of their sequence.

In summary, we draw the conclusion.

Theorem 3. *In unitary space, the parallel feature fusion based on GPCA has a property of symmetry.*

Similarly, we can prove that the parallel feature fusion based on GKLE or GLDA satisfies the property of symmetry as well. That is to say, based on the parallel feature combination form $\alpha + \mathrm{i}\beta$ or $\beta + \mathrm{i}\alpha$, after feature extraction via linear projection analysis, the classification results are identical. In other words, the fusion result is independent of the form of parallel feature combination.

6. Experiments and analysis

6.1. Experiment 1

In this experiment, our goal is to compare the performances of the parallel feature fusion and serial fusion in the case where the dimensions of two sets of feature vectors are *equal*.

The experiment is performed on NUST603 handwritten Chinese character database built in Nanjing University of Science and Technology. The database contains 19 groups of Chinese characters, which are collected from bank checks. There are 400 samples for each character (7200 in total), in which 200 are for training and the others for testing. Some original samples are shown in Fig. 1. We do the experiment based on two sets of original feature vectors: 128-dimensional cross feature [22] and 128-dimensional peripheral feature [23].

Two feature-preprocessing techniques mentioned in Section 4 are, respectively, used for preprocessing the original cross feature and peripheral feature. Since the dimensions of two features are equal, the combination coefficient $\theta = 1$. Thus, the parallel combination is formed as cross + i peripheral, and serial combination is formed as $\binom{\text{Cross}}{\text{Peripheral}}$. And, the parallel combined feature is 128-dimensional complex vector, while the serial combined feature is 256-dimensional real vector.

In this experiment, the classical PCA is used to compress the original cross feature, peripheral feature and their serial combination into k (k varies from 20 to 40)-dimensional transformed space respectively. And, the proposed GPCA is used for dimension reduction based on the parallel combined feature. Since PCA is a special case of GPCA, the above description can be uniformly described as exploiting GPCA for feature extraction in each feature space. Similarly, GKLE1 and GLDA are used for feature extraction in each feature space as well.

Finally, in each transformed space, a Bayesian classifier is employed. Here, since the prior probability of each class is equal, the Bayesian discrimination function in unitary space can be defined as

$$g_l(x) = \tfrac{1}{2}\ln|\Sigma_l| + \tfrac{1}{2}(x - \mu_l)^{\mathrm{H}}\Sigma_l^{-1}(x - \mu_l),$$

where μ_l and $\mathbf{\Sigma}_l$ denote the mean vector and covariance matrix of class l, respectively. By virtue of this function, the decision rule is: if sample x satisfies $|g_k(x)| = \min_l |g_l(x)|$, then $x \in \omega_k$. The classification results are shown in Tables 1–3. In each column of these tables, the value italicized denotes the lowest error rate based on the given feature set.

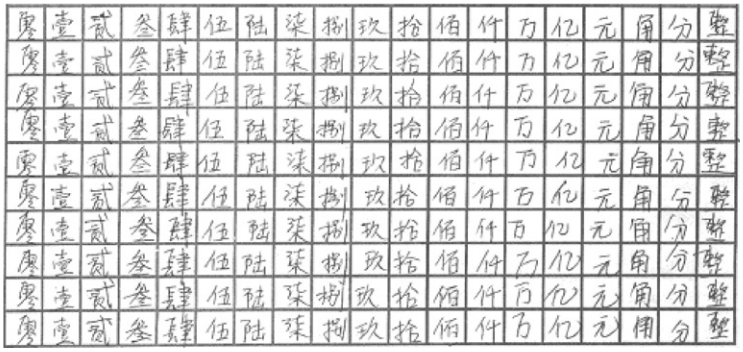
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整
零壹贰叁肆伍陆柒捌玖拾佰仟万亿元角分整

Fig. 1. Some images of Chinese characters in NUST603 handwritten database.

Table 1
Classification error rates of GPCA transformed cross feature, GPCA transformed peripheral feature, parallel fused feature and serial fused feature in Experiment 1

Number of axes	Cross feature	Peripheral feature	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
20	0.058	0.053	0.021	0.027	0.017	0.023
22	0.056	0.043	0.019	0.025	0.016	0.021
24	*0.055*	0.041	0.016	0.022	0.013	0.019
26	0.056	0.037	0.016	0.020	*0.012*	0.017
28	0.057	0.037	0.014	0.021	0.012	0.015
30	0.060	*0.034*	0.013	0.021	0.013	0.014
32	0.059	0.034	0.013	0.018	0.013	0.013
34	0.063	0.036	*0.012*	0.016	0.012	0.014
36	0.067	0.037	0.013	0.015	0.013	0.013
38	0.067	0.036	0.013	*0.013*	0.012	*0.012*
40	0.068	0.038	0.013	0.014	0.013	0.012
Average (%)	6.05	3.87	1.48	1.93	1.33	1.57

Table 2
Classification error rates of GKLE1 transformed cross feature, GKLE1 transformed peripheral feature, parallel fused feature and serial fused feature in Experiment 1

Number of axes	Cross feature	Peripheral feature	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
20	*0.058*	0.092	0.026	0.055	0.022	0.045
22	0.062	0.074	0.024	0.052	0.020	0.044
24	0.060	0.063	0.022	0.048	0.020	0.037
26	0.065	0.062	0.021	0.041	0.019	0.031
28	0.063	0.060	0.019	0.038	0.018	0.027
30	0.061	0.053	0.018	0.033	0.016	0.027
32	0.064	0.051	0.016	0.030	0.017	0.022
34	0.063	0.044	0.016	0.027	*0.015*	0.021
36	0.068	*0.039*	*0.014*	0.024	0.016	0.021
38	0.070	0.040	0.015	*0.020*	0.016	*0.018*
40	0.069	0.042	0.016	0.020	0.015	0.019
Average (%)	6.39	5.64	1.88	3.53	1.76	2.84

Table 3
Classification error rates of GLDA transformed cross feature, GLDA transformed peripheral feature, parallel fused feature and serial fused feature in Experiment 1

Number of axes	Cross feature	Peripheral feature	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
2	0.514	0.514	0.287	0.503	0.296	0.465
4	0.209	0.214	0.090	0.168	0.096	0.148
6	0.141	0.098	0.049	0.078	0.047	0.066
8	0.096	0.073	0.030	0.039	0.031	0.041
10	0.083	0.057	0.025	0.029	0.023	0.030
12	0.070	0.048	0.019	0.024	0.024	0.023
14	0.072	0.042	0.021	0.020	0.020	0.021
16	*0.069*	0.041	0.019	*0.019*	0.018	0.018
18	0.069	*0.039*	*0.018*	0.019	*0.017*	*0.017*
Average (%)	14.7	12.5	6.20	9.99	6.36	9.11

From Tables 1–3, it is evident that the classification error rate is significantly reduced after parallel or serial feature fusion. And for each feature preprocessing technique, it does not matter that the number of axes varies, the error rate of parallel feature fusion is less or no more than that of serial feature fusion. Moreover, the average classification error rate of parallel feature fusion is much less than that of serial feature fusion. This indicates that the classification results based on parallel fused feature are more robust with the dimension (the number of features or axes) varying.

Taking Table 1 as an example, the maximal recognition accuracies based on PCA transformed cross transformed feature and peripheral transformed feature are 94.5% and 96.6%, respectively, while the maximal recognition accuracy based on parallel fused feature reaches 98.8%, with an increase of 4.3% compared to cross feature and 2.2% to peripheral feature.

6.2. *Experiment 2*

In this experiment, our goal is to compare the performances of the parallel feature fusion and serial fusion in the case that the dimensions of two sets of feature vectors are *unequal*.

The experiment was done on Concordia University CENPARMI handwritten numeral database. The database contains 10 classes of digits, and each class has 600 samples. Thus, there are 6000 samples in total, in which 4000 are for training and the others for testing. Some images of original samples are shown in Fig. 2. We do the experiment based on two sets of features: 256-dimensional Gabor transformation feature [24] and 121-dimensional Legendre moment feature [25]. Two feature-preprocessing techniques mentioned in Section 4 are, respectively, used to initialize the original Gabor feature and Legendre feature. Since the dimensions of two feature vectors are unequal, it needs to evaluate the combination coefficient θ before combination. As the feature-preprocessing technique I is adopted, we evaluate $\theta = (256/121)^2 = 4.4762$, and as the feature-preprocessing technique II is adopted, we evaluate $\theta = 2.8$. Thus, the serial combination is formed as $\begin{pmatrix}\text{Gabor}\\ \theta\text{Legendre}\end{pmatrix}$ and the parallel combination is formed as Gabor$+$ iθLegendre. And, the parallel combined feature space is a 256-dimensional complex vector space, while the serial combined feature space is a 377-dimensional real vector space.

Then, GPCA, GKLE1 and GLDA are used to compress the original Gabor feature, Legendre feature and their two kinds of combined features into a low-dimensional transformed space. Finally, in each transformed space, a Bayesian classifier same as that used in Experiment 1 is employed. The classification results are shown in Tables 4–6.

From Tables 4–6, it is obvious that the classification error rate is reduced after parallel or serial feature fusion. And for each feature-preprocessing technique, it does not matter that the number of axes varies, the error rate of parallel feature fusion is less or no more than that of serial feature fusion. Moreover, the average classification error rate of parallel feature fusion is much less than that of serial feature fusion. This indicates that the classification results based on parallel fused feature are more robust with the dimension (the number of features or axes) varying.

Taking Table 4 as an example, the maximal recognition accuracies based on Gabor transformed feature and Legendre transformed feature are 90.2% and 94.4%, respectively, while, the maximal recognition accuracy based on parallel fused feature reaches 97.0%, with an increase of 6.8% compared to Gabor feature and 3.6% to Legendre feature, despite the dimension being only 42.

In addition, the results of experiment also confirm that the proposed method of evaluating combination coefficient in the feature-preprocessing step is effective.

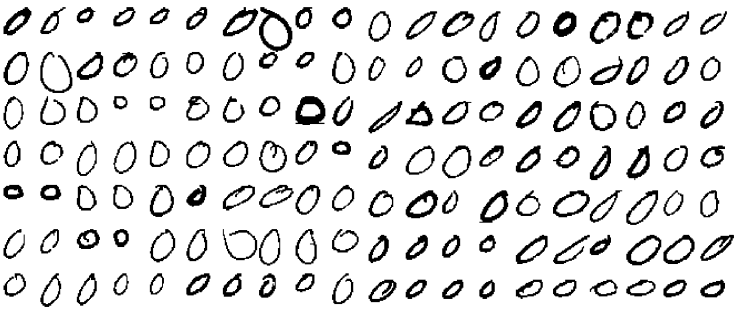

Fig. 2. Some images of digits in CENPARMI handwritten numeral database.

Table 4
Classification error rates of GPCA transformed Gabor feature, GPCA transformed Legendre feature, parallel fused feature and serial fused feature in Experiment 2

Number of axes	Gabor feature	Legendre feature	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
36	0.106	0.065	0.034	0.037	0.048	0.055
38	0.101	0.064	0.034	0.039	0.047	0.051
40	0.100	0.059	0.033	0.040	0.045	0.051
42	0.102	0.057	*0.030*	0.037	0.048	0.048
44	0.102	0.057	0.033	0.036	0.044	*0.046*
46	0.104	0.057	0.032	0.036	0.045	0.047
48	0.099	0.057	0.032	0.035	*0.042*	0.050
50	0.099	*0.056*	0.032	0.035	0.042	0.049
52	*0.098*	0.056	0.033	*0.032*	0.043	0.048
54	0.099	0.059	0.033	0.035	0.044	0.048
56	0.099	0.058	0.035	0.034	0.045	0.047
58	0.101	0.059	0.035	0.035	0.044	0.049
60	0.098	0.059	0.035	0.035	0.045	0.046
Average (%)	10.1	5.87	3.32	3.58	4.48	4.88

Table 5
Classification error rates of GKLE1 transformed Gabor feature, GKLE1 transformed Legendre feature, parallel fused feature and serial fused feature in Experiment 2

Number of axes	Gabor feature	Legendre feature	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
30	0.107	0.055	0.035	0.039	0.046	0.044
32	0.108	0.056	0.037	0.040	0.045	0.048
34	0.108	0.051	0.037	0.039	0.043	0.044
36	0.109	*0.047*	0.039	0.038	0.041	0.044
38	0.111	0.051	0.038	*0.036*	0.040	0.047
40	0.111	0.052	*0.034*	0.037	0.037	0.045
42	0.106	0.056	0.035	0.039	0.038	0.048
44	0.105	0.057	0.036	0.041	0.037	0.047
46	0.107	0.059	0.039	0.041	*0.036*	*0.043*
48	*0.100*	0.057	0.037	0.040	0.038	0.047
50	0.103	0.059	0.036	0.039	0.037	0.048
Average (%)	10.7	5.45	3.66	3.90	3.98	4.59

Table 6
Classification error rates of GLDA transformed Gabor feature, GLDA transformed Legendre feature, parallel fused feature and serial fused feature in Experiment 2

Number of axes	Cross feature	Peripheral feature	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
1	0.660	0.606	0.566	0.579	0.565	0.582
2	0.448	0.357	0.276	0.302	0.296	0.310
3	0.303	0.272	0.123	0.131	0.147	0.214
4	0.246	0.172	0.089	0.094	0.104	0.109
5	0.199	0.141	0.076	0.082	0.085	0.088
6	0.166	0.114	0.059	0.065	0.073	0.079
7	0.157	*0.097*	0.052	0.052	0.061	0.069
8	*0.153*	0.097	*0.048*	*0.051*	*0.057*	*0.061*
9	0.153	0.097	0.051	0.052	0.058	0.061
Average (%)	27.6	21.7	14.9	15.6	16.1	17.5

Based on the results of Experiments 1 and 2, we can draw the following conclusion:

(1) The classification accuracy is significantly increased after parallel feature fusion.
(2) The performance of the proposed parallel feature fusion is better or at least not worse than that of serial feature fusion.

However, the above two experiments are both performed on large sample size problems. Now, we test the effectiveness of the parallel feature fusion using a small sample size problem.

6.3. Experiment 3

The experiment is performed on the ORL (http://www.cam-orl.co.uk) database that contains a set of face images taken at the Olivetti Research Laboratory in Cambridge, UK. There are 10 different images for 40 individuals. For some people, the images were taken at different times. And the facial expression (open/closed eyes, smiling/nonsmiling) and facial details (glasses/no glasses) are variable. The images were taken against a dark homogeneous background and the people are in upright, frontal position with tolerance for some tilting and rotation of up to 20°. Moreover, there is some variation in scale of up to about 10%. All images are grayscale and normalized with a resolution of 92 × 112. Some images in ORL are shown in Fig. 3.

In this experiment, we use the first five images of each person for training and the remaining five for testing. Thus, the total amount of training samples and testing samples are both 200.

In this experiment, we intend to use two kinds of K–L transform features as original features. First, the eigenfaces method in [26] is adopted to work out the 39 orthogonal eigenvectors of the total covariance matrix Σ_t corresponding to the first 39 largest eigenvalues. Then, we use them as feature extractor to transform the $92\times112=10304$-dimensional image vector into 39-dimensional feature vector α. In a similar way, we work out the 39 orthogonal eigenvectors of the between-class covariance matrix Σ_b and use them to transform the original image vector into 39-dimensional feature vector β.

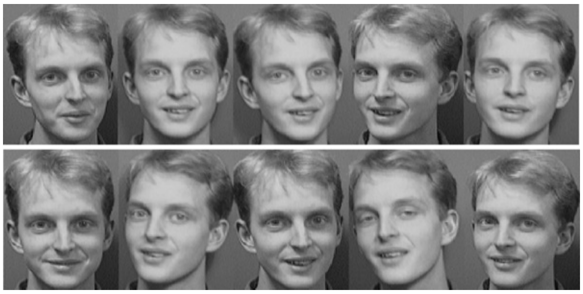
Fig. 3. Ten images of one person in ORL face database.

The two feature-preprocessing techniques mentioned in Section 4 are, respectively, used. After feature preprocessing, the parallel combination is formed as $\gamma=\alpha+i\beta$, and the serial combination is formed as $\gamma=\binom{\alpha}{\beta}$. And, the parallel combined feature is a 39-dimensional complex vector, while the serial combined feature is a 78-dimensional real vector.

The proposed GLDA is used to transform α,β and their two types of combined features into a k-dimensional (k varies from 11 to 39) transformed space, respectively. Finally, in each transformed space, a minimum Mahalanobis distance classifier is employed. Note that in unitary space, the Mahalanobis distance between the sample x and the mean vector of class l is defined by

$$g_l(x)=\sqrt{(x-\mu_l)^{\mathrm{H}}\Sigma^{-1}(x-\mu_l)},$$

where μ_l denotes the mean vector of class l and Σ denotes the total within-class covariance matrix. If sample x satisfies

Table 7
Classification errors of GLDA transformed feature α and β, parallel fused feature and serial fused feature in Experiment 3

Number of axes	α	β	Preprocessing I		Preprocessing II	
			Parallel	Serial	Parallel	Serial
11	17	31	14	36	11	23
13	15	23	14	37	11	21
15	*12*	21	15	35	*8*	20
17	13	19	14	32	8	21
19	13	20	15	31	9	21
21	12	17	14	33	9	19
23	12	17	13	32	9	20
25	12	17	13	29	9	18
27	12	18	*12*	29	10	17
29	12	17	12	29	9	17
31	12	18	12	*26*	10	*16*
33	12	17	12	27	9	17
35	12	17	12	26	9	17
37	12	*16*	12	27	9	18
39	12	16	12	27	9	17
Average	12.7	18.9	13.1	30.4	9.3	18.8

$|g_k(x)| = \min_l |g_l(x)|$, then $x \in \omega_k$. The recognition errors are shown in Table 7.

It is evident in Table 7 that the classification error of parallel fusion is much lower than that of serial fusion whether the feature-preprocessing technique I or technique II is adopted. And, based on feature-preprocessing technique II and parallel fusion, a recognition accuracy of 96% is achieved using only 15 features.

By this experiment, we can draw a conclusion that the performance of parallel feature fusion is superior to the serial feature fusion in small sample size problems. The main reason is that, in the case of small sample size, the dimension of feature vector is higher, and the accurate estimation of the scatter matrix is more difficult. Due to the dimensional increase after serial feature combination, estimation accuracy of scatter matrices is decreased. That is why the serial feature fusion strategy underperforms in this experiment. Conversely, the dimension stays invariable after parallel feature combination, so it is possible to estimate the scatter matrices more accurately in the complex feature space. This may be the key reason why parallel fusion outperforms the serial fusion in small sample size problems.

Specially, if the sum of dim α and dim β is larger than the total number of training samples, after serial feature combination $\gamma = \binom{\alpha}{\beta}$, the resulting training samples lead to a singular within-class scatter matrix. For example, if the dimensions of α and β are both 120 and the number of training samples is 200, in the serial combined feature space, the within-class scatter matrix is singular. In this case, it is very difficult to perform linear feature extraction based on LDA or KLE1. Conversely, the parallel feature fusion artfully avoids this difficulty. In fact, it is easy to prove that the training samples resulting from feature parallel combination can ensure the nonsingularity of the within-class scatter matrix as long as the training samples in original feature space ensure it. This property of parallel combination provides convenience for linear extraction.

Moreover, this experiment also indicates that the feature-preprocessing technique II is more suitable for the small sample size case.

7. Conclusion and summary

A new feature fusion strategy, *parallel feature fusion*, is introduced in this paper. A complex vector used to represent the parallel combined feature and the traditional linear projection methods, including principal component analysis, K–L expansion and linear discriminant analysis, are generalized for feature extraction in the complex feature space. In fact, the traditional projection methods are special cases of the generalized projection methods developed in this paper.

The idea and theory proposed in this paper enrich the content of feature level fusion. Thus, two styles of feature fusion techniques come into being. One is the classical serial feature fusion and the other is the presented parallel feature fusion. As a comparison, an outstanding advantage of parallel feature fusion is that the increase of dimension is avoided after feature combination. Thus, on the one hand, much computational time is saved in the process of subsequent feature extraction; on the other hand, the difficulty of within-class scatter matrix being singular is avoided in the case when the sum of dimensions of two sets of feature vectors involved in combination is larger than the total number of training samples, which is convenient for the subsequent linear extraction based on LDA or K–L expansion methods.

The experiments on CENPARMI handwritten numeral database, NUST603 handwritten Chinese character database and ORL face image database indicate that the recognition accuracy is increased significantly and the dimensionality of the feature vector is reduced largely under parallel feature fusion. What is more, the experimental results also demonstrate that the proposed parallel feature fusion is better or at least not worse than the classical serial feature fusion for large sample size problems. And, for small sample size problems, the parallel feature fusion is superior to the serial feature fusion.

In summary, this paper provides a new and effective means for feature level fusion. The developed parallel feature fusion techniques have practical significance and wide applicability. And it deserves to be emphasized that the parallel feature fusion method has more intuitive physical meaning when applied to some real-life problem. For example, in object recognition problems, if the intensity image and range image of an object are captured at the same time, and they are of same size and well matched, we can combine them together using a complex matrix, in which each element contains intensity information as well as the range information for the point. After parallel feature extraction, the resulting low-dimensional complex feature vectors containing fusion information are used for recognition. This is a problem we are setting out to research on.

Acknowledgements

We wish to thank the National Science Foundation of China, under Grant No.60072034, for supporting our research. In addition, we would like to thank the anonymous reviewers for their constructive advice.

Appendix A. Proof of Theorem 2

If the first k discriminant vectors $\varphi_1,\dots,\varphi_k$ are chosen as $X_1,\dots,X_k$, respectively, then by Corollary 4 and the S_t-orthogonal constraints in Model 1, it follows that the $(k+1)$th discriminant vectors $\varphi_{k+1}\in\operatorname{span}\{X_{k+1},\dots,X_n\}$. That is, φ_{k+1} can be denoted by $\varphi_{k+1}=c_{k+1}X_{k+1}+\cdots+c_nX_n$. Then, we have

$$J(\varphi_{k+1})=\frac{\lambda_{k+1}c_{k+1}^2+\cdots+\lambda_n c_n^2}{c_{k+1}^2+\cdots+c_n^2}\leqslant\lambda_{k+1}.$$

According to Corollary 3, $J(X_{k+1})=\lambda_{k+1}$, so φ_{k+1} can be selected as X_{k+1}. □

Appendix B. Proof of Lemma 3

$\xi=(a_1+\mathrm{i}b_1,\dots,a_n+\mathrm{i}b_n)^{\mathrm{T}}$, then

$$\begin{aligned}\xi^{\mathrm{H}}(x+\mathrm{i}y)&=(a_1-\mathrm{i}b_1,\dots,a_n-\mathrm{i}b_n)(x_1+\mathrm{i}y_1,\dots,x_n+\mathrm{i}y_n)^{\mathrm{T}}\\&=\sum_{j=1}^{n}(a_jx_j+b_jy_j)+\mathrm{i}\sum_{j=1}^{n}(a_jy_j-b_jx_j)\\&=p+\mathrm{i}q,\end{aligned}$$

$$\begin{aligned}\bar{\xi}^{\mathrm{H}}(y+\mathrm{i}x)&=(a_1+\mathrm{i}b_1,\dots,a_n+\mathrm{i}b_n)(y_1+\mathrm{i}x_1,\dots,y_n+\mathrm{i}x_n)^{\mathrm{T}}\\&=\sum_{j=1}^{n}(a_jy_j-b_jx_j)+\mathrm{i}\sum_{j=1}^{n}(a_jx_j+b_jy_j)\\&=q+\mathrm{i}p.\end{aligned}$$

So the proposition holds. □

References

[1] H.-C. Chiang, R.L. Moses, L.C. Potter, Model-based Bayesian feature matching with application to synthetic aperture radar target recognition, Pattern Recognition 34 (8) (2001) 1539–1553.

[2] T. Peli, Mon Young, R. Knox, et al., Feature level sensor fusion, Proceedings of the SPIE Sensor Fusion: Architectures, Algorithms and Applications III, Vol. 3719, pp. 332–339.

[3] N. Doi, A. Shintani, Y. Hayashi, et al., A study on month shape features suitable for HMM speech recognition using fusion of visual and auditory information, IEICE Trans. Fundam. E78-A (11) (1995) 1548–1552.

[4] A.H. Gunatilaka, B.A. Baertlein, Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 577–589.

[5] M.E. Ulug, C.L. McCullough, Feature and data level fusion of infrared and visual images, SPIE Conference on Sensor Fusion: Architectures, Algorithms and Applications III, Vol. 3719, pp. 312–318.

[6] I.S. Chang, R.-H. Park, Segmentation based on fusion of range and intensity images using robust trimmed methods, Pattern Recognition 34 (10) (2001) 1952–1962.

[7] V. Dassigi, R.C. Mann, V.A. Protopoescu, Information fusion for text classification—an experimental comparison, Pattern Recognition 34 (12) (2001) 2413–2425.

[8] H. Li, R. Deklerck, B.D. Cuyper, et al., Object recognition in brain CT-scans: knowledge-based fusion of data from multiple feature extractors, IEEE Trans. Medical Imaging 14 (2) (1995) 212–228.

[9] L.O. Jimenez, et al., Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks, IEEE Trans. Geosci. Remote Sensing 37 (3) (1999) 1360–1366.

[10] Y.S. Huang, C.Y. Suen, A method of combining multiple experts for the recognition of unconstrained handwritten numerals, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 90–94.

[11] 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.

[12] Zhang Xinhua, An information model and method of feature fusion, Int. Conf. Signal Process. 2 (1998) 1389–1392.

[13] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Network 5 (4) (1994) 537–550.

[14] Y. Shi, T. Zhang, Feature analysis: support vector machines approaches, SPIE Conference on Image Extraction, Segmentation, and Recognition, Vol. 4550, pp. 245–251.

[15] Cheng Jun Liu, Harry Wechsler, A shape- and texture-based enhanced Fisher classifier for face recognition, IEEE Trans. Image Process. 10 (4) (2001) 598–608.

[16] Jian Yang, Jing-yu Yang, Generalized K–L transform based combined feature extraction, Pattern Recognition 35 (1) (2002) 295–297.

[17] Xue-ren Ding, Miao-ke Cai, Matrix Theory in Engineering, Tianjin University Press, Tianjin, 1995, pp. 115–118.

[18] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall International, Inc., London, 1982, pp. 324–337.

[19] Zhong Jin, J.Y. Yang, et al., Face recognition based on uncorrelated discriminant transformation, Pattern Recognition 33 (7) (2001) 1405–1467.

[20] Zhong Jin, J.Y. Yang, et al., A theorem on uncorrelated optimal discriminant vectors, Pattern Recognition 33 (10) (2001) 2041–2047.

[21] Yang Jian, Yang Jing-Yu, Jin Zhong, A feature extraction approach using optimal discriminant transform and image recognition, J. Comput. Res. Dev. 38 (11) (2001) 1331–1336.

[22] Y.T. Tang, et al., Offline recognition of Chinese handwriting by multifeature and multilevel classification, IEEE Trans. Pattern Anal. Mach. Intell. 20 (5) (1998) 556–561.

[23] Y.H. Tseng, C.C. Kuo, H.J. Lee, Speeding up Chinese character recognition in an automatic document reading system, Pattern Recognition 31 (11) (1998) 1601–1612.

[24] H. Yoshihiko, et al., Recognition of handwritten numerals using Gabor features, Proceedings of the 13th ICPR, pp. 250–253.

[25] S.X. Liao, M. Pawlak, On image analysis by moments, IEEE Trans. Pattern Anal. Mach. Intell. 18 (3) (1996) 254–266.

[26] M. Turk, A. Pentland, Face recognition using eigenfaces, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.

About the Author—JIAN YANG was born in Jiangsu, China, on 3 June 1973. He received his M.S. degree in Applied Mathematics from the Changsha Railway University in 1998 and his Ph.D. degree in Pattern Recognition and Intelligent Systems from Nanjing University of Science and Technology (NUST) in 2002. At present, he is involved in a research work at the Hong Kong Polytechnic University. He is the author of more than 10 scientific papers in pattern recognition and computer vision. His current research interests include face recognition and detection, handwritten character recognition and data fusion.

About the Author—JING-YU YANG received a B.S. degree in Computer Science from NUST, Nanjing, China. From 1982 to 1984, he was a visiting scientist at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. From 1993 to 1994, he was a visiting professor at the Department of Computer Science, Missuria University. In 1998, he acted as a visiting professor at Concordia University in Canada. He is currently a professor and Chairman in the department of Computer Science at NUST. He is the author of more than 200 scientific papers in computer vision, pattern recognition, and artificial intelligence. He has won more than 20 provincial and national awards. His current research interests are in the areas of pattern recognition, robot vision, image processing, data fusion, and artificial intelligence.

About the Author—DAVID ZHANG graduated in Computer Science from Peking University and received his M.S. and Ph.D. degrees in Computer Science and Engineering from the Harbin Institute of Technology (HIT) in 1983 and 1985, respectively. He received his second Ph.D. in Electrical and Computer Engineering at the University of Waterloo, Ontario, Canada, in 1994. Currently, he is a professor at the Hong Kong Polytechnic University. He is a founder and director of both Biometrics Technology Centres supported by UGC/CRC, Hong Kong Government, and the National Nature Scientific Foundation (NSFC) of China, respectively. In addition, he is a founder and editor-in-chief of the International Journal of Image and Graphics, and an associate editor of Pattern Recognition, the International Journal of Pattern Recognition and Artificial Intelligence, and the International Journal of Robotics and Automation. So far, he has published more than 170 papers, including four books. Prof. Zhang is a senior member of the IEEE.

About the Author—JIANFENG LU received B.S., M.S. and Ph.D. degrees in Computer Science from Nanjing University of Science and Technology in 1991, 1994 and 2000, respectively. Now, he is an associate professor in NUST. His current research interests include image processing, pattern recognition and computer vision.