KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition

Jian Yang, Alejandro F. Frangi, Jing-yu Yang, David Zhang, *Senior Member*, *IEEE*, and Zhong Jin

Abstract—This paper examines the theory of kernel Fisher discriminant analysis (KFD) in a Hilbert space and develops a two-phase KFD framework, i.e., kernel principal component analysis (KPCA) plus Fisher linear discriminant analysis (LDA). This framework provides novel insights into the nature of KFD. Based on this framework, the authors propose a complete kernel Fisher discriminant analysis (CKFD) algorithm. CKFD can be used to carry out discriminant analysis in "double discriminant subspaces." The fact that, it can make full use of two kinds of discriminant information, *regular* and *irregular*, makes CKFD a more powerful discriminator. The proposed algorithm was tested and evaluated using the FERET face database and the CENPARMI handwritten numeral database. The experimental results show that CKFD outperforms other KFD algorithms.

Index Terms—Kernel-based methods, subspace methods, principal component analysis (PCA), Fisher linear discriminant analysis (LDA or FLD), feature extraction, machine learning, face recognition, handwritten digit recognition.

1 Introduction

Over the last few years, kernel-based learning machines, e.g., support vector machines (SVMs) [1], kernel principal component analysis (KPCA), and kernel Fisher discriminant analysis (KFD), have aroused considerable interest in the fields of pattern recognition and machine learning [2]. KPCA was originally developed by Schölkopf et al. in 1998 [3], while KFD was first proposed by Mika et al. in 1999 [4], [5]. Subsequent research saw the development of a series of KFD algorithms (see Baudat and Anouar [6], Roth and Steinhage [7], Mika et al. [8], [9], [10], Yang [11], Lu et al. [12], Xu et al. [13], Billings and Lee [14], Gestel et al. [15], Cawley and Talbot [16], and Lawrence and Schölkopf [17]). The KFD algorithms developed by Mika et al. are formulated for two classes, while those of Baudat and Anouar are formulated for multiple classes. Because of its ability to extract the most discriminatory nonlinear features [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], KFD has been found to be very effective in many real-world applications.

KFD, however, always encounters the *ill-posed* problem in its real-world applications [10], [18]. A number of regularization techniques that might alleviate this problem have been suggested. Mika et al. [4], [10] used the technique of making the inner product matrix nonsingular by adding a scalar matrix. Baudat and Anouar [6] employed the QR decomposition technique to avoid the singularity by removing the zero eigenvalues. Yang [11] exploited the PCA plus LDA technique adopted in Fisherface [20] to deal with the problem. Unfortunately, all of these methods discard the discriminant information contained in the null space of the within-class covariance matrix, yet this discriminant information is very effective for "small sample size" (SSS) problem [21], [22], [23], [24], [25]. Lu et al. [12] have taken this issue into account and presented kernel direct discriminant analysis (KDDA) by generalization of the direct-LDA [23].

In real-world applications, particularly in image recognition, there are a lot of SSS problems in observation space (input space). In such problems, the number of training samples is less than the dimension of feature vectors. For kernel-based methods, due to the implicit high-dimensional nonlinear mapping determined by kernel, many typical "large sample size" problems in observation space, such as handwritten digit recognition, are turned into SSS problems in *feature space*. These problems can be called generated SSS problems. Since SSS problems are common, it is necessary to develop new and more effective KFD algorithms to deal with them.

Fisher linear discriminant analysis has been well studied and widely applied to SSS problems in recent years. Many LDA algorithms have been proposed [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. The most famous method is Fisherface [19], [20], which is based on a two-phase framework: PCA plus LDA. The effectiveness of this framework in image recognition has been broadly demonstrated [19], [20], [26], [27], [28], [29]. Recently, the theoretical foundation for this framework has been laid [24], [25]. Besides, many researchers have been dedicated to searching for more effective discriminant subspaces [21], [22], [23], [24], [25]. A significant result is the finding that there exists crucial discriminative information in the null space of the

- *J. Yang, J.-y. Yang, and Z. Jin are with the Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, PR China.*
E-mail: csjyang@comp.polyu.edu.hk, yangjy@mail.njust.edu.cn.
- *A.F. Frangi is with the Computational Imaging Lab, Department of Technology, Pompeu Fabra University, Pg Circumvallacio 8, E08003 Barcelona, Spain. E-mail: alejandro.frangi@upf.edu.*
- *D. Zhang and J. Yang are with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong, PR China.*
E-mail: csdzhang@comp.polyu.edu.hk.
- *Z. Jin is with the Centre of Computer vision, University Autonoma of Barcelona, E-08193 Barcelona, Spain. E-mail: zhongjin@cvc.uab.es.*

Manuscript received 6 Aug. 2003; revised 21 Apr. 2004; accepted 19 July 2004; published online 13 Dec. 2004.
Recommended for acceptance by J. Weng.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0212-0803.

within-class scatter matrix [21], [22], [23], [24], [25]. In this paper, we call this kind of discriminative information *irregular* discriminant information, in contrast with *regular* discriminant information outside of the null space.

Kernel Fisher discriminant analysis would be likely to benefit in two ways from the state-of-the-art LDA techniques. One is the adoption of a more concise algorithm framework and the other is that it would allow the use of *irregular* discriminant information. This paper seeks to improve KFD in these ways, first of all developing a new KFD framework, KPCA plus LDA, based on a rigorous theoretical derivation in Hilbert space. Then, a complete KFD algorithm (CKFD) is proposed based on the framework. Unlike current KLD algorithms, CKFD can take advantage of two kinds discriminant information: *regular and irregular*. Finally, CKFD was used in face recognition and handwritten numeral recognition. The experimental results are encouraging.

The remainder of the paper is organized as follows: In Section 2, the idea of KPCA and KFD is given. A two-phase KFD framework, KPCA plus LDA, is developed in Section 3 and a complete KFD algorithm (CKFD) is proposed in Section 4. In Section 5, the experiments are performed on the FERET face database and CENPARMI handwritten numeral database whereby the proposed algorithm is evaluated and compared to other methods. Finally, a conclusion and discussion are offered in Section 6.

2 Outline of KPCA and KFD

For a given nonlinear mapping Φ, the *input data space* $\mathbb{R}^n$ can be mapped into the *feature space* $\mathcal{H}$:

$$\begin{aligned}\Phi : \mathbb{R}^n &\to \mathcal{H}\\ x &\mapsto \Phi(x).\end{aligned} \tag{1}$$

As a result, a pattern in the original *input space* $\mathbb{R}^n$ is mapped into a potentially much higher dimensional feature vector in the *feature space* $\mathcal{H}$. Since the *feature space* $\mathcal{H}$ is possibly infinite-dimensional and the orthogonality needs to be characterized in such a space, it is reasonable to view $\mathcal{H}$ as a Hilbert space. In this paper, $\mathcal{H}$ is always regarded as a Hilbert space.

An initial motivation of KPCA (or KFD) is to perform PCA (or LDA) in the *feature space* $\mathcal{H}$. However, it is difficult to do so directly because it is computationally very intensive to compute the dot products in a high-dimensional *feature space*. Fortunately, kernel techniques can be introduced to avoid this difficulty. The algorithm can be actually implemented in the *input space* by virtue of kernel tricks. The explicit mapping process is not required at all. Now, let us describe KPCA as follows.

Given a set of M training samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ in $\mathbb{R}^n$, the covariance operator on the *feature space* $\mathcal{H}$ can be constructed by

$$\mathbf{S}_t^\Phi = \frac{1}{M}\sum_{j=1}^{M}\left(\Phi(\mathbf{x}_j) - \mathbf{m}_0^\Phi\right)\left(\Phi(\mathbf{x}_j) - \mathbf{m}_0^\Phi\right)^{\mathrm{T}}, \tag{2}$$

where $\mathbf{m}_0^\Phi = \frac{1}{M}\sum_{j=1}^{M}\Phi(\mathbf{x}_j)$. In a finite-dimensional Hilbert space, this operator is generally called covariance matrix. The covariance operator satisfies the following properties:

Lemma 1. $\mathbf{S}_t^\Phi$ *is a*

1. *bounded operator,*
2. *compact operator,*
3. *positive operator, and*
4. *self-adjoint (symmetric) operator on Hilbert space* $\mathcal{H}$.

The proof is given in Appendix A.

Since every eigenvalue of a positive operator is nonnegative in a Hilbert space [48], from Lemma 1, it follows that all nonzero eigenvalues of $\mathbf{S}_t^\Phi$ are positive. It is these positive eigenvalues that are of interest to us. Schölkopf et al. [3] have suggested the following way to find them.

It is easy to show that every eigenvector of $\mathbf{S}_t^\Phi$, β, can be linearly expanded by

$$\beta = \sum_{i=1}^{M} a_i \Phi(\mathbf{x}_i). \tag{3}$$

To obtain the expansion coefficients, let us denote $\mathbf{Q} = [\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_M)]$ and form an $M \times M$ Gram matrix $\tilde{\mathbf{R}} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$, whose elements can be determined by virtue of kernel tricks:

$$\tilde{\mathbf{R}}_{ij} = \Phi(\mathbf{x}_i)^{\mathrm{T}}\Phi(\mathbf{x}_j) = \left(\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)\right) = \mathrm{k}(\mathbf{x}_i, \mathbf{y}_j). \tag{4}$$

Centralize $\tilde{\mathbf{R}}$ by

$$\begin{gathered}\mathbf{R} = \tilde{\mathbf{R}} - \mathbf{1}_M\tilde{\mathbf{R}} - \tilde{\mathbf{R}}\,\mathbf{1}_M + \mathbf{1}_M\tilde{\mathbf{R}}\,\mathbf{1}_M,\\ \text{where the matrix } \mathbf{1}_M = (1/M)_{M\times M}.\end{gathered} \tag{5}$$

Calculate the orthonormal eigenvectors $\gamma_1, \gamma_2, \ldots, \gamma_m$ of $\mathbf{R}$ corresponding to the m largest positive eigenvlaues, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. The orthonormal eigenvectors $\beta_1, \beta_2, \ldots, \beta_m$ of $\mathbf{S}_t^\Phi$ corresponding to the m largest positive eigenvlaues, λ_1, $\lambda_2, \ldots, \lambda_m$, then are

$$\beta_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{Q}\gamma_j, \quad j = 1, \ldots, m. \tag{6}$$

After the projection of the mapped sample $\Phi(\mathbf{x})$ onto the eigenvector system $\beta_1, \beta_2, \ldots, \beta_m$, we can obtain the KPCA-transformed feature vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}$ by

$$\mathbf{y} = \mathbf{P}^{\mathrm{T}}\Phi(\mathbf{x}), \text{ where } \mathbf{P} = (\beta_1, \beta_2, \ldots, \beta_m). \tag{7}$$

Specifically, the jth KPCA feature (component) y_j is obtained by

$$\begin{aligned}y_j &= \beta_j^{\mathrm{T}}\Phi(\mathbf{x}) = \frac{1}{\sqrt{\lambda_j}}\gamma_j^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\Phi(\mathbf{x})\\ &= \frac{1}{\sqrt{\lambda_j}}\gamma_j^{\mathrm{T}}[\mathbf{k}(\mathbf{x}_1, \mathbf{x}),\ \mathbf{k}(\mathbf{x}_2, \mathbf{x}), \ldots, \mathbf{k}(\mathbf{x}_M, \mathbf{x})], j = 1, \ldots, m.\end{aligned} \tag{8}$$

In the formulation of KFD, a similar technique is adopted again. That is, expand the Fisher discriminant vector using (3) and then formulate the problem in a space spanned by all mapped training samples. For more details, please refer to [4], [6], [11].

3 A New KFD Algorithm Framework: KPCA Plus LDA

In this section, we will build a rigorous theoretical framework for kernel Fisher discriminant analysis. This framework is important because it provides a solid theoretical foundation for our two-phased KFD algorithm that will be presented in

Section 4. That is, the presented two-phased KFD algorithm is not empirically-based but theoretically-based.

To provide more theoretical insights into KFD, we would like to examine the problems in a whole Hilbert space rather than in the space spanned by training samples. Here, an infinite-dimensional Hilbert space is preferred because any proposition that holds in an infinite-dimensional Hilbert space will hold in a finite-dimensional Hilbert space (but, the reverse might be not true). So, in this section, we will discuss the problems in an infinite-dimensional Hilbert space $\mathcal{H}$.

3.1 Fundamentals

Suppose there are c known pattern classes. The between-class scatter operator $\mathbf{S}_b^\Phi$ and the within-class scatter operator $\mathbf{S}_w^\Phi$ in the *feature space* $\mathcal{H}$ are defined below:

$$\mathbf{S}_b^\Phi = \frac{1}{M}\sum_{i=1}^{c} l_i\left(\mathbf{m}_i^\Phi - \mathbf{m}_0^\Phi\right)\left(\mathbf{m}_i^\Phi - \mathbf{m}_0^\Phi\right)^{\mathrm{T}}, \tag{9}$$

$$\mathbf{S}_w^\Phi = \frac{1}{M}\sum_{i=1}^{c}\sum_{j=1}^{l_i}\left(\Phi(\mathbf{x}_{ij}) - \mathbf{m}_i^\Phi\right)\left(\Phi(\mathbf{x}_{ij}) - \mathbf{m}_i^\Phi\right)^{\mathrm{T}}, \tag{10}$$

where $\mathbf{x}_{ij}$ denotes the $j\mathrm{th}$ training sample in class i, l_i is the number of training samples in class i, $\mathbf{m}_i^\Phi$ is the mean of the mapped training samples in class i, $\mathbf{m}_0^\Phi$ is the mean across all mapped training samples.

From the above definitions, we have $\mathbf{S}_t^\Phi = \mathbf{S}_b^\Phi + \mathbf{S}_w^\Phi$. Following along with the proof of Lemma 1, it is easy to prove that the two operators satisfy the following properties:

Lemma 2. *$\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ are both*

1. *bounded operators,*
2. *compact operators,*
3. *self-adjoint (symmetric) operators, and*
4. *positive operators on Hilbert space $\mathcal{H}$.*

Since $\mathbf{S}_b^\Phi$ is a self-adjoint (symmetric) operator in Hilbert space $\mathcal{H}$, the inner product between φ and $\mathbf{S}_b^\Phi\varphi$ satisfies $\langle\varphi, \mathbf{S}_b^\Phi\varphi\rangle = \langle\mathbf{S}_b^\Phi\varphi, \varphi\rangle$. So, we can write it as $\langle\varphi, \mathbf{S}_b^\Phi\varphi\rangle \triangleq \varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi$. Note that, if $\mathbf{S}_b^\Phi$ is not self-adjoint, this denotation is meaningless. Since $\mathbf{S}_b^\Phi$ is also a positive operator, we have $\varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi \geq 0$. Similarly, we have $\langle\varphi, \mathbf{S}_w^\Phi\varphi\rangle = \langle\mathbf{S}_w^\Phi\varphi, \varphi\rangle \triangleq \varphi^{\mathrm{T}}\mathbf{S}_w^\Phi\varphi \geq 0$. Thus, in Hilbert space $\mathcal{H}$, the Fisher criterion function can be defined by

$$J^\Phi(\varphi) = \frac{\varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi}{\varphi^{\mathrm{T}}\mathbf{S}_w^\Phi\varphi}, \quad \varphi \neq \mathbf{0}. \tag{11}$$

If the within-class scatter operator $\mathbf{S}_w^\Phi$ is invertible, $\varphi^{\mathrm{T}}\mathbf{S}_w^\Phi\varphi > 0$ always holds for every nonzero vector φ. In such a case, the Fisher criterion can be directly employed to extract a set of optimal discriminant vectors (projection axes) using the standard LDA algorithm [35]. Its physical meaning is that, after the projection of samples onto these axes, the ratio of the between-class scatter to the within-class scatter is maximized.

However, in a high-dimensional (even infinite-dimensional) *feature space* $\mathcal{H}$, it is almost impossible to make $\mathbf{S}_w^\Phi$ invertible because of the limited amount of training samples in real-world applications. That is, there always exist vectors satisfying $\varphi^{\mathrm{T}}\mathbf{S}_w^\Phi\varphi = 0$ (actually, these vectors are from the null space of $\mathbf{S}_w^\Phi$). These vectors turn out to be very effective if they satisfy $\varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi > 0$ at the same time [22], [24], [25]. This is because the positive between-class scatter makes the data become well separable when the within-class scatter is zero. In such a case, the Fisher criterion degenerates into the following between-class scatter criterion:

$$J_b^\Phi(\varphi) = \varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi, (||\varphi|| = 1). \tag{12}$$

As a special case of the Fisher criterion, the criterion given in (12) is very intuitive since it is reasonable to use the between-class scatter to measure the discriminatory ability of a projection axis when the within-class scatter is zero [22], [24].

In this paper, we will use the between-class scatter criterion defined in (12) to derive the *irregular discriminant vectors* from $\mathrm{null}(\mathbf{S}_w^\Phi)$ (i.e., the null space of $\mathbf{S}_w^\Phi$), while using the standard Fisher criterion defined in (11) to derive the *regular discriminant vectors* from the complementary set $\mathcal{H} - \mathrm{null}(\mathbf{S}_w^\Phi)$.

3.2 Strategy for Finding Fisher Optimal Discriminant Vectors in Feature Space

Now, a problem is how to find the two kinds of Fisher optimal discriminant vectors in *feature space* $\mathcal{H}$. Since $\mathcal{H}$ is very large (high or infinite-dimensional), it is computationally too intensive or even infeasible to calculate the optimal discriminant vectors directly. To avoid this difficulty, the present KFD algorithms [4], [6], [7] all formulate the problem in the space spanned by the mapped training samples. The technique is feasible when the *irregular* case is disregarded, but the problem becomes more complicated when the *irregular* discriminant information is taken into account since the *irregular* discriminant vectors exist in the null space of $\mathbf{S}_w^\Phi$. Because the null space of $\mathbf{S}_w^\Phi$ is possibly infinite-dimensional, the existing techniques for dealing with the singularity of LDA [22], [24] are inapplicable since they are all limited to a finite-dimensional space in theory.

In this section, we will examine the problem in an infinite-dimensional Hilbert space and try to find a way to solve it. Our strategy is to reduce the *feasible solution space* (search space) where two kinds of discriminant vectors might hide. It should be stressed that we would not like to lose any effective discriminant information in the process of space reduction. To this end, some theory should be developed first.

Theorem 1 (Hilbert-Schmidt Theorem [49]). *Let $\mathbf{A}$ be a compact and self-adjoint operator on Hilbert space $\mathcal{H}$. Then, its eigenvector system forms an orthonormal basis for $\mathcal{H}$.*

Since $\mathbf{S}_t^\Phi$ is compact and self-adjoint, it follows from Theorem 1 that its eigenvector system $\{\beta_i\}$ forms an orthonormal basis for $\mathcal{H}$. Suppose $\beta_1, \ldots, \beta_m$ are eigenvectors corresponding to positive eigenvalues of $\mathbf{S}_t^\Phi$, where $m = \mathrm{rank}(\mathbf{S}_t^\Phi) = \mathrm{rank}(\mathbf{R})$. Generally, $m = M - 1$, where M is the number of training samples. Let us define the subspace $\Psi_t = \mathrm{span}\{\beta_1, \beta_2, \ldots, \beta_m\}$. Suppose its orthogonal complementary space is denoted by $\Psi_t^\perp$. Actually, $\Psi_t^\perp$ is the null space of $\mathbf{S}_t^\Phi$. Since Ψ_t, due to its finite dimensionality, is a closed subspace of $\mathcal{H}$, from the *Projection theorem* [50], we have

Corollary 1. *$\mathcal{H} = \Psi_t \oplus \Psi_t^\perp$. That is, for an arbitrary vector $\varphi \in \mathcal{H}, \varphi$ can be uniquely represented in the form $\varphi = \phi + \zeta$ with $\phi \in \Psi_t$ and $\zeta \in \Psi_t^\perp$.*

Now, let us define a mapping $L : \mathcal{H} \to \Psi_t$ by

$$\varphi = \phi + \zeta \to \phi, \tag{13}$$

where ϕ is called the orthogonal projection of φ onto Ψ_t. It is easy to verify that L is a linear operator from $\mathcal{H}$ onto its subspace Ψ_t.

Theorem 2. *Under the mapping $L : \mathcal{H} \rightarrow \Psi_t$ determined by $\varphi = \phi + \zeta \rightarrow \phi$, the Fisher criterion satisfies the following properties:*

$$J_b^\Phi(\varphi) = J_b^\Phi(\phi) \text{ and } J^\Phi(\varphi) = J^\Phi(\phi). \tag{14}$$

The proof is given in Appendix B.

According to Theorem 2, we can conclude that both kinds of discriminant vectors can be derived from Ψ_t without any loss of effective discriminatory information with respect to the Fisher criterion. Since the new search space Ψ_t is finite-dimensional and much smaller (less dimensional) than $\mathrm{null}(\mathbf{S}_w^\Phi)$ and $\mathcal{H} - \mathrm{null}(\mathbf{S}_w^\Phi)$, it is feasible to derive discriminant vectors from it.

3.3 Idea of Calculating Fisher Optimal Discriminant Vectors

In this section, we will offer our idea of calculating Fisher optimal discriminant vectors in the reduced search space Ψ_t. Since the dimension of Ψ_t is m, according to functional analysis theory [47], Ψ_t is isomorphic to m-dimensional Euclidean space IR^m The corresponding *isomorphic mapping* is

$$\varphi = \mathbf{P}\eta, \text{ where } \mathbf{P} = (\beta_1, \beta_2, \ldots, \beta_m), \eta \in \mathrm{IR}^m, \tag{15}$$

which is a one-to-one mapping from IR^m onto Ψ_t.

Under the isomorphic mapping $\varphi = \mathbf{P}\eta$, the criterion functions $J^\Phi(\varphi)$ and $J_b^\Phi(\varphi)$ in feature space are, respectively, converted into

$$J^\Phi(\varphi) = \frac{\eta^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_b^\Phi\mathbf{P})\eta}{\eta^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_w^\Phi\mathbf{P})\eta} \text{ and } J_b^\Phi(\varphi) = \eta^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_b^\Phi\mathbf{P})\eta. \tag{16}$$

Now, based on (16), let us define two functions:

$$J(\eta) = \frac{\eta^{\mathrm{T}}\mathbf{S}_b\eta}{\eta^{\mathrm{T}}\mathbf{S}_w\eta}, (\eta \neq \mathbf{0}) \text{ and } J_b(\eta) = \eta^{\mathrm{T}}\mathbf{S}_b\eta, \ (||\eta|| = 1), \tag{17}$$

where $\mathbf{S}_b = \mathbf{P}^{\mathrm{T}}\mathbf{S}_b^\Phi\mathbf{P}$ and $\mathbf{S}_w = \mathbf{P}^{\mathrm{T}}\mathbf{S}_w^\Phi\mathbf{P}$.

It is easy to show that $\mathbf{S}_b$ and $\mathbf{S}_w$ are both $m \times m$ semipositive definite matrices. This means that $J(\eta)$ is a generalized Rayleigh quotient [34] and $J_b(\eta)$ is a Rayleigh quotient in the isomorphic space IR^m. Note that $J_b(\eta)$ is viewed as a Rayleigh quotient because the formula $\eta^{\mathrm{T}}\mathbf{S}_b\,\eta\,(||\eta|| = 1)$ is equivalent to $\frac{\eta^{\mathrm{T}}\mathbf{S}_b\eta}{\eta^{\mathrm{T}}\eta}$ [34].

Under the isomorphic mapping mentioned above, the stationary points (optimal solutions) of the Fisher criterion have the following intuitive property:

Theorem 3. *Let $\varphi = \mathbf{P}\,\eta$ be an isomorphic mapping from IR^m onto Ψ_t. Then, $\varphi * = P\eta *$ is the stationary point of $J^\Phi(\varphi)$ $(J_b^\Phi(\varphi))$ if and only if $\eta *$ is the stationary point of $J(\eta)$ $(J_b(\eta))$.*

From Theorem 3, it is easy to draw the following conclusion:

Corollary 2. *If $\eta_1, \ldots, \eta_d$ is a set of stationary points of the function $J(\eta)(J_b(\eta))$, then, $\varphi_1 = \mathbf{P}\eta_1, \ldots, \varphi_d = \mathbf{P}\eta_d$ is a set of* regular *(*irregular*) optimal discriminant vectors with respect to the Fisher criterion $J^\Phi(\varphi)$ $(J_b^\Phi(\varphi))$.*

Now, the problem of calculating the optimal discriminant vectors in subspace Ψ_t is transformed into the extremum problem of the (generalized) Rayleigh quotient in the isomorphic space IR^m.

3.4 A Concise KFD Framework: KPCA Plus LDA

The obtained optimal discriminant vectors are used for feature extraction in *feature space*. Given a sample $\mathbf{x}$ and its mapped image $\Phi(\mathbf{x})$, we can obtain the discriminant feature vector $\mathbf{z}$ by the following transformation:

$$\mathbf{z} = \mathbf{W}^{\mathrm{T}}\Phi(\mathbf{x}), \tag{18}$$

where

$$\begin{aligned}\mathbf{W}^{\mathrm{T}} &= (\varphi_1, \varphi_2, \ldots, \varphi_d)^{\mathrm{T}} = (\mathbf{P}\eta_1, \mathbf{P}\eta_2, \ldots, \mathbf{P}\eta_d)^{\mathrm{T}} \\ &= (\eta_1, \eta_2, \ldots, \eta_d)^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}.\end{aligned}$$

The transformation in (18) can be decomposed into two transformations:

$$\mathbf{y} = \mathbf{P}^{\mathrm{T}}\Phi(\mathbf{x}), \text{ where } \mathbf{P} = (\beta_1, \beta_2, \ldots, \beta_m), \tag{19}$$

and

$$\mathbf{z} = \mathbf{G}^{\mathrm{T}}\mathbf{y}, \text{ where } \mathbf{G} = (\eta_1, \eta_2, \ldots, \eta_d). \tag{20}$$

Since $\beta_1, \beta_2, \ldots, \beta_m$ are eigenvectors of $\mathbf{S}_t^\Phi$ corresponding to positive eigenvalues, the transformation in (19) is exactly KPCA; see (7) and (8). This transformation transforms the input space IR^n into space IR^m.

Now, let us view the issues in the KPCA-transformed space IR^m. Looking back at (17) and considering the two matrices $\mathbf{S}_b$ and $\mathbf{S}_w$, it is easy to show that they are between-class and within-class scatter matrices in IR^m. In fact, we can construct them directly by

$$\mathbf{S}_b = \frac{1}{M}\sum_{i=1}^{c} l_i(\mathbf{m}_i - \mathbf{m}_0)(\mathbf{m}_i - \mathbf{m}_0)^{\mathrm{T}}, \tag{21}$$

$$\mathbf{S}_w = \frac{1}{M}\sum_{i=1}^{c}\sum_{j=1}^{l_i}(\mathbf{y}_{ij} - \mathbf{m}_i)\,(\mathbf{y}_{ij} - \mathbf{m}_i)^{\mathrm{T}}, \tag{22}$$

where $\mathbf{y}_{ij}$ denotes the jth training sample in class i, l_i is the number of training samples in class i, $\mathbf{m}_i$ is the mean of the training samples in class i, $\mathbf{m}_0$ the mean across all training samples.

Since $\mathbf{S}_b$ and $\mathbf{S}_w$ are between-class and within-class scatter matrices in IR^m, the functions $J(\eta)$ and $J_b(\eta)$ can be viewed as Fisher criterions and, their stationary points $\eta_1, \ldots, \eta_d$ are the associated Fisher optimal discriminant vectors. Correspondingly, the transformation in (20) is the Fisher linear discriminant transformation (LDA) in the KPCA-transformed space IR^m.

Up to now, the essence of KFD has been revealed. That is, KPCA is first used to reduce (or increase) the dimension of the *input space* to m, where m is the rank of $\mathbf{S}_t^\Phi$ (i.e., the rank of the centralized Gram matrix $\mathbf{R}$). Next, LDA is used for further feature extraction in the KPCA-transformed space IR^m.

In summary, a new KFD framework, i.e., KPCA plus LDA, is developed in this section. This framework offers us a new insight into the nature of kernel Fisher discriminant analysis.

4 Complete KFD Algorithm

In this section, we will develop a complete KFD algorithm based on the two-phase KFD framework. Two kinds of discriminant information, *regular* and *irregular*, will be derived and fused for classification tasks.

4.1 Extraction of Two Kinds of Discriminant Features

Our task is to explore how to perform LDA in the KPCA-transformed space $\mathbb{R}^m$. After all, the standard LDA algorithm [35] remains inapplicable since the within-class scatter matrix $\mathbf{S}_w$ is still singular in $\mathbb{R}^m$. We would rather take advantage of this *singularity* to extract more discriminant information than avoid it by means of the previous regularization techniques [4], [6], [11]. Our strategy is to split the space $\mathbb{R}^m$ into two subspaces: the null space and the range space of $\mathbf{S}_w$. We then use the Fisher criterion to derive the *regular* discriminant vectors from the range space and use the between-class scatter criterion to derive the *irregular* discriminant vectors from the null space.

Suppose $\alpha_1, \ldots, \alpha_m$ are the orthonormal eigenvectors of $\mathbf{S}_w$ and assume that the first q ones corresponde to nonzero eigenvalues, where $q = \text{rank}\ (\mathbf{S}_w)$. Let us define a subspace $\Theta_w = \text{span}\{\alpha_{q+1}, \ldots, \alpha_m\}$. Its orthogonal complementary space is $\Theta_w^\perp = \text{span}\{\alpha_1, \ldots, \alpha_q\}$.

Actually, Θ_w is the null space and $\Theta_w^\perp$ is the range space of $\mathbf{S}_w$ and $\mathbb{R}^m = \Theta_w \oplus \Theta_w^\perp$. The dimension of the subspace $\Theta_w^\perp$ is p. Generally, $q = M - c = m - c + 1$. The dimension of the subspace Θ_w is $p = m - q$. Generally, $p = c - 1$.

Lemma 3. *For every nonzero vector* $\eta \in \Theta_w$*, the inequality* $\eta^{\mathrm{T}}\mathbf{S}_b\eta > 0$ *always holds.*

The proof is given in Appendix C.

Lemma 3 tells us there indeed exists *irregular* discriminant information in the null space of $\mathbf{S}_w, \Theta_w$, since the within-class scatter is zero while the between-class scatter is always positive. Thus, the optimal *irregular* discriminant vectors must be derived from this space. On the other hand, since every nonzero vector $\eta \in \Theta_w^\perp$ satisfies $\eta^{\mathrm{T}}\mathbf{S}_w\eta > 0$, it is feasible to derive the optimal regular discriminant vectors from $\Theta_w^\perp$ using the standard Fisher criterion.

The idea of isomorphic mapping discussed in Section 3.3 can still be used for calculations of the optimal *regular* and *irregular* discriminant vectors.

Let us first consider the calculation of the optimal *regular* discriminant vectors in $\Theta_w^\perp$. Since the dimension of $\Theta_w^\perp$ is q, $\Theta_w^\perp$ is isomorphic to Euclidean space $\mathbb{R}^q$ and the corresponding isomorphic mapping is

$$\eta = \mathbf{P}_1\xi, \text{ where } \mathbf{P}_1 = (\alpha_1, \ldots, \alpha_q). \tag{23}$$

Under this mapping, the Fisher criterion $J(\eta)$ in (17) is converted into

$$\tilde{J}(\xi) = \frac{\xi^{\mathrm{T}}\tilde{\mathbf{S}}_b\xi}{\xi^{\mathrm{T}}\tilde{\mathbf{S}}_w\xi}, \quad (\xi \neq \mathbf{0}), \tag{24}$$

where $\tilde{\mathbf{S}}_b = \mathbf{P}_1^{\mathrm{T}}\mathbf{S}_b\mathbf{P}_1$ and $\tilde{\mathbf{S}}_w = \mathbf{P}_1^{\mathrm{T}}\mathbf{S}_w\mathbf{P}_1$. It is easy to verify that $\tilde{\mathbf{S}}_b$ is semipositive definite and $\tilde{\mathbf{S}}_w$ is positive definite (must be invertible) in $\mathbb{R}^q$. Thus, $\tilde{J}(\xi)$ is a standard generalized Rayleigh quotient. Its stationary points $\mathbf{u}_1, \ldots, \mathbf{u}_d$ $(d \leq c - 1)$ are actually the generalized eigenvectors of the generalized eigenequation $\tilde{\mathbf{S}}_b\xi = \lambda\tilde{\mathbf{S}}_w\xi$ corresponding to the d largest positive eigenvalues [34]. It is easy to calculate them using the standard LDA algorithm [33], [35]. After working out $\mathbf{u}_1, \ldots, \mathbf{u}_d$, we can obtain $\tilde{\eta}_j = \mathbf{P}_1\mathbf{u}_j$ $(j = 1, \ldots, d)$ using (23). From the property of isomorphic mapping, we know $\tilde{\eta}_1, \ldots, \tilde{\eta}_d$ are the optimal *regular* discriminant vectors with respect to $J(\eta)$.

In a similar way, we can calculate the optimal *irregular* discriminant vectors within Θ_w. Θ_w is isomorphic to Euclidean space $\mathbb{R}^p$ and the corresponding isomorphic mapping is

$$\eta = \mathbf{P}_2\xi, \text{ where } \mathbf{P}_2 = (\alpha_{q+1}, \ldots, \alpha_m). \tag{25}$$

Under this mapping, the criterion $J_b(\eta)$ in (17) is converted into

$$\hat{J}_b(\xi) = \xi^{\mathrm{T}}\hat{\mathbf{S}}_b\xi, \quad (||\xi|| = 1), \tag{26}$$

where $\hat{\mathbf{S}}_b = \mathbf{P}_2^{\mathrm{T}}\mathbf{S}_b\mathbf{P}_2$. It is easy to verify that $\hat{\mathbf{S}}_b$ is positive definite in $\mathbb{R}^p$. The stationary points $\mathbf{v}_1, \ldots, \mathbf{v}_d$ $(d \leq c - 1)$ of $\hat{J}_b(\xi)$ are actually the orthonormal eigenvectors of $\hat{\mathbf{S}}_b$ corresponding to d largest eigenvalues. After working out $\mathbf{v}_1, \ldots, \mathbf{v}_d$, we can obtain $\hat{\eta}_j = \mathbf{P}_2\mathbf{v}_j$ $(j = 1, \ldots, d)$ using (25). From the property of isomorphic mapping, we know $\hat{\eta}_1, \ldots, \hat{\eta}_d$ are the optimal *irregular* discriminant vectors with respect to $J_b(\eta)$.

Based on the derived optimal discriminant vectors, the linear discriminant transformation in (20) can be performed in $\mathbb{R}^m$. Specifically, after the projection of the sample $\mathbf{y}$ onto the *regular* discriminant vectors $\tilde{\eta}_1, \ldots, \tilde{\eta}_d$, we can obtain the *regular* discriminant feature vector:

$$\mathbf{z}^1 = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)^{\mathrm{T}}\mathbf{y} = \mathbf{U}^{\mathrm{T}}\mathbf{P}_1^{\mathrm{T}}\mathbf{y}, \tag{27}$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$, $\mathbf{P}_1 = (\alpha_1, \ldots, \alpha_q)$.

After the projection of the sample $\mathbf{y}$ onto the *irregular* discriminant vectors $\hat{\eta}_1, \ldots, \hat{\eta}_d$, we can obtain the *irregular* discriminant feature vector:

$$\mathbf{z}^2 = (\hat{\eta}_1, \ldots, \hat{\eta}_d)^{\mathrm{T}}\mathbf{y} = \mathbf{V}^{\mathrm{T}}\mathbf{P}_2^{\mathrm{T}}\mathbf{y}, \tag{28}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$, $\mathbf{P}_2 = (\alpha_{q+1}, \ldots, \alpha_m)$.

4.2 Fusion of Two Kinds of Discriminant Features for Classification

Since, for any given sample, we can obtain two d-dimensional discriminant feature vectors, it is possible to fuse them in the decision level. Here, we suggest a simple fusion strategy based on a summed normalized-distance.

Suppose the distance between two samples $\mathbf{z}_i$ and $\mathbf{z}_j$ is given by

$$g(\mathbf{z}_i, \mathbf{z}_j) = ||\mathbf{z}_i - \mathbf{z}_j||, \tag{29}$$

where $||\cdot||$ is the notation of norm. The norm determines what measure is used. For example, the Euclidean norm $||\cdot||_2$ defines the usual Euclidean distance. For simplicity, the Euclidean measure is adopted in this paper.

Let us denote a pattern $\mathbf{z} = [\mathbf{z}^1, \mathbf{z}^2]$, where $\mathbf{z}^1, \mathbf{z}^2$ are *regular* and *irregular* discriminant feature vectors of the same pattern. The summed normalized-distance between sample $\mathbf{z}$ and the training sample $\mathbf{z}_i = [\mathbf{z}_i^1, \mathbf{z}_i^2]$ $(i = 1, \ldots, M)$ is defined by

$$\bar{g}(\mathbf{z},\ \mathbf{z}_i) = \theta\frac{||\mathbf{z}^1 - \mathbf{z}_i^1||}{\sum\limits_{j=1}^{M}||\mathbf{z}^1 - \mathbf{z}_j^1||} + \frac{||\mathbf{z}^2 - \mathbf{z}_i^2||}{\sum\limits_{j=1}^{M}||\mathbf{z}^2 - \mathbf{z}_j^2||}, \tag{30}$$

where θ is the fusion coefficient. This coefficient determines the weight of *regular* discriminant information in the decision level.

When a nearest neighbor classifier is used, if a sample $\mathbf{z}$ satisfies $\bar{g}(\mathbf{z},\ \mathbf{z}_j) = \min_i \bar{g}(\mathbf{z},\ \mathbf{z}_i)$ and $\mathbf{z}_j$ belongs to class k, then $\mathbf{z}$ belongs to class k. When a minimum distance classifier is used, the mean vector $\mu_i = [\mu_i^1, \mu_i^2]$ of class i is viewed as a prototype of samples in such a class. If a sample $\mathbf{z}$ satisfies $\bar{g}(\mathbf{z},\ \mu_k) = \min_i \bar{g}(\mathbf{z},\ \mu_i)$, then $\mathbf{z}$ belongs to class k.

4.3 Complete KFD Algorithm

In summary of the discussion so far, the complete KFD algorithm is given below:

CKFD Algorithm

Step 1. Use KPCA to transform the input space IR^n into an m-dimensional space IR^m, where $m = \mathrm{rank}(\mathbf{R}), \mathbf{R}$ is the centralized Gram matrix. Pattern $\mathbf{x}$ in IR^n is transformed to be KPCA-based feature vector $\mathbf{y}$ in IR^m.

Step 2. In IR^m, construct the between-class and within-class scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$. Calculate $\mathbf{S}_w$'s orthonormal eigenvectors, $\alpha_1, \ldots, \alpha_m$, assuming the first $q\,(q = \mathrm{rank}(\mathbf{S}_w))$ ones are corresponding to positive eigenvalues.

Step 3. Extract the *regular* discriminant features: Let $\mathbf{P}_1 = (\alpha_1, \ldots, \alpha_q)$. Define $\tilde{\mathbf{S}}_b = \mathbf{P}_1^{\mathrm{T}} \mathbf{S}_b \mathbf{P}_1$ and $\tilde{\mathbf{S}}_w = \mathbf{P}_1^{\mathrm{T}} \mathbf{S}_w \mathbf{P}_1$ and calculate the generalized eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_d\ (d \leq c-1)$ of $\tilde{\mathbf{S}}_b \xi = \lambda \tilde{\mathbf{S}}_w \xi$ corresponding to the d largest positive eigenvalues using the algorithm in [33], [35]. Let $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$. The *regular* discriminant feature vector is $\mathbf{z}^1 = \mathbf{U}^{\mathrm{T}} \mathbf{P}_1^{\mathrm{T}} \mathbf{y}$.

Step 4. Extract the *irregular* discriminant features: Let $\mathbf{P}_2 = (\alpha_{q+1}, \ldots, \alpha_m)$. Define $\hat{\mathbf{S}}_b = \mathbf{P}_2^{\mathrm{T}} \mathbf{S}_b \mathbf{P}_2$ and calculate $\hat{\mathbf{S}}_b$'s orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d\ (d \leq c-1)$ corresponding to the d largest eigenvalues. Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$. The *irregular* discriminant feature vector is $\mathbf{z}^2 = \mathbf{V}^{\mathrm{T}} \mathbf{P}_2^{\mathrm{T}} \mathbf{y}$.

Step 5. Fuse the *regular* and *irregular* discriminant features using summed normalized-distance for classification.

Concerning the implementation of the CKFD algorithm, a remark should be made. For numerical robustness, in Step 2 of the CKFD algorithm, q could be selected as a number that is properly less than the real rank of $\mathbf{S}_w$ in practical applications. In this paper, we choose q as the number of eigenvalues that are less than $\frac{\lambda_{\max}}{2{,}000}$, where $\lambda_{\max}$ is the maximal eigenvalue of $\mathbf{S}_w$.

4.4 Relationship to Other KFD (or LDA) Algorithms

In this section, we will review some other KFD (LDA) methods and explicitly distinguish them from the proposed CKFD. Let us begin with the linear discriminant analysis methods. Liu et al. [21] first claimed that there exist two kinds of discriminant information for LDA in small sample size cases, irregular discriminant information (within the null space of *within-class* scatter matrix) and regular discriminant information (beyond the null space). Chen et al. [22] emphasized the irregular information and proposed a more effective way to extract it, but overlooked the regular information. Yu and Yang [23] took two kinds of discriminatory information into account and suggested extracting them within the range space of the *between-class* scatter matrix. Since the dimension of the range space is up to $c-1$, Yu et al.'s algorithm (DLDA) is computationally more efficient for SSS problems in that the computational complexity is reduced to be $\mathcal{O}(c^3)$.

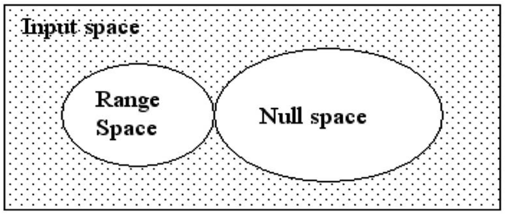


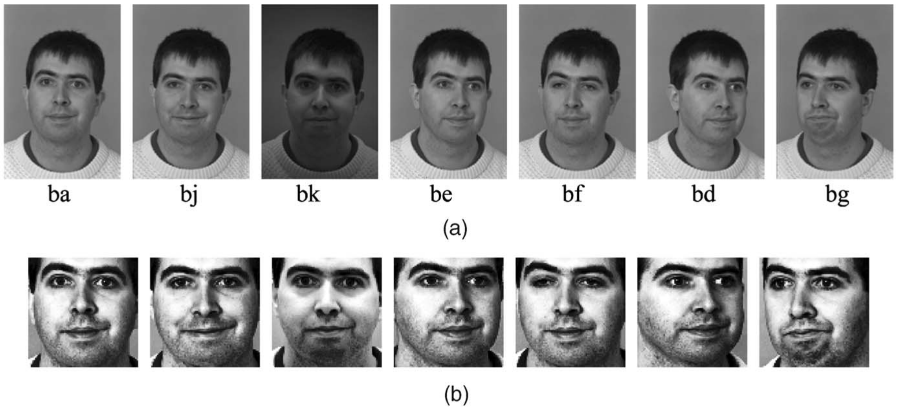
Fig. 1. Illustration the subspaces of DLDA.

DLDA, however, is suboptimal in theory. Although there is no discriminatory information within the null space of the *between-class* scatter matrix, no theory (like Theorem 2) can guarantee that all discriminatory information must exist in the range space because there is a large space beyond the null and the range space which may contain crucial discriminant information; see the shadow area in Fig. 1. For two-class problems (such as gender recognition), the weakness of DLDA becomes more noticeable. The range space is only one-dimensional and spanned by the difference of the two class mean vectors. This subspace is too small to contain enough discriminant information. Actually, in such a case, the resulting discriminant vector of DLDA is the difference vector itself, which is not optimal with respect to the Fisher criterion, let alone the ability to extract two kinds of discriminant information.

Lu et al. [12] generalized DLDA using the idea of kernels and presented kernel direct discriminant analysis (KDDA). KDDA was demonstrated effective for face recognition, but, as a nonlinear version of DLDA, KDDA unavoidably suffers the weakness of DLDA. On the other hand, unlike DLDA, which can significantly reduce computational complexity of LDA (as discussed above), KDDA has the same computational complexity, i.e., $\mathcal{O}(M^3)$, as other KFD algorithms [4], [5], [6], [7], [8], [9], [10], [11] because KDDA still needs to calculate the eigenvectors of an $M \times M$ Gram matrix.

Like Liu et al.'s [21] method, our previous LDA algorithm [24] can obtain more than $c-1$ features, that is, all $c-1$ irregular discriminant features plus some regular ones. This algorithm turned out to be more effective than Chen and Yu's methods, which can extract at most $c-1$ features. In addition, our LDA algorithm [24] is more powerful and simpler than Liu et al.'s [21] method [52]. The algorithm in the literature [32] can be viewed as a nonlinear generalization of that in [24]. However, the derivation of the algorithm is based on an assumption that the feature space is assumed to be a finite dimensional space. This assumption is no problem for polynomial kernels, but is unsuitable for other kernels which determine mappings that might lead to an infinite-dimensional feature space.

Compared to our previous idea [32] and Lu et al.'s KDDA, CKFD has two prominent advantages. One is in the theory and other is in the algorithm itself. The theoretical derivation of the algorithm does not need any assumption. The developed theory in Hilbert space lays a solid foundation for the algorithm. The derived discriminant information is guaranteed not only optimal but also complete (lossless) with respect to the Fisher criterion. The completeness of discriminant information enables CKFD to be used to perform discriminant analysis in "double discriminant subspaces." In each subspace, the number of discriminant features can be up to $c-1$. This means $2(c-1)$ features can be obtained in total. This is different from the KFD (or LDA) algorithms discussed above and beyond [4], [5], [6], [7], [8], [9], [10], [11],



(a)

(b)

Fig. 2. Images of one person in the FERET database. (a) Original images. (b) Cropped images (after histogram equalization) corresponding to images in (a).

[12], [13], [14], [15], [16], [17], [22], [23], which can yield only one discriminant subspace containing at most $c-1$ discriminant features. What is more, CKFD provides a new mechanism for decision fusion. This mechanism makes it possible to take advantage of the two kinds of discriminant information and to determine their contribution to decision by modifying the fusion coefficient.

CKFD has a computational complexity of $\mathcal{O}(M^3)$ (M is the number of training samples), which is the same as the existing KFD algorithms [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. The reason for this is that the KPCA phase of CKFD is actually carried out in the space spanned by M training samples, so its computational complexity still depends on the operations of solving $M \times M$ sized eigenvalue problems [3], [10]. Despite this, compared to other KFD algorithms, CKFD indeed requires additional computation mainly owing to its space decomposition process performed in the KPCA-transformed space. In such a space, all eigenvectors of $\mathbf{S}_w$ should be calculated.

5 Experiments

In this section, three experiments are designed to evaluate the performance of the proposed algorithm. The first experiment is on face recognition and the second one is on handwritten digit recognition. Face recognition is typically a small sample size problem, while handwritten digit classification is a "large sample size" problem in observation space. We will demonstrate that the proposed CKFD algorithm is applicable to both of these kinds of problems. In the third experiment, CKFD is applied to two-class problems, i.e., the classification of digit-pairs. We will show that CKFD is capable of exhibiting data in two-dimensional space for two-class problems.

5.1 Experiment on Face Recognition Using the FERET Database

The FERET face image database is a result of the FERET program, which was sponsored by the US Department of Defense through the DARPA Program [36], [37]. It has become a standard database for testing and evaluating state-of-the-art face recognition algorithms.

The proposed algorithm was tested on a subset of the FERET database. This subset includes 1,400 images of 200 individuals (each individual has seven images). It is composed of the images whose names are marked with two-character strings: "ba," "bj," "bk," "be," "bf," "bd," and "bg" [51]. This subset involves variations in facial expression, illumination, and pose. In our experiment, the facial portion of each original image was automatically cropped based on the location of eyes and the cropped image was resized to 80×80 pixels and preprocessed by histogram equalization. Some example images of one person are shown in Fig. 2.

Three images of each subject are randomly chosen for training, while the remaining four images are used for testing. Thus, the training sample set size is 600 and the testing sample set size is 800. In this way, we run the system 20 times and obtain 20 different training and testing sample sets. The first 10 are used for model selection and the others for performance evaluation.

Two popular kernels are involved in our tests. One is the polynomial kernel $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+1)^r$ and the other is the Gaussian RBF kernel $k(\mathbf{x},\mathbf{y}) = \exp(-||\mathbf{x}-\mathbf{y}||^2/\delta)$. Three methods, namely, Kernel Eigenface [11], Kernel Fisherface [11], and the proposed CKFD algorithm, are tested and compared. In order to gain more insights into our algorithm, two additional versions, 1) CKFD: *regular*, where only the *regular* discriminant features are used, and 2) CKFD: *irregular*, where only the *irregular* discriminant features are used, are also evaluated. Two simple classifiers, a minimum distance classifier (MD) and a nearest neighbor classifier (NN), are employed in the experiments.

In the phase of model selection, our goal is to determine proper kernel parameters (i.e., the order r of the polynomial kernel and the width δ of the Gaussian RBF kernel), the dimension of the projection subspace for each method, and the fusion coefficient θ for CKFD. Since it is very difficult to determine these parameters at the same time, a stepwise selection strategy is more feasible and thus is adopted here [12]. Specifically, we fix the dimension and the fusion coefficient (only for CKFD) in advance and try to find the optimal kernel parameter for a given kernel function. Then, based on the chosen kernel parameters, the selection of the subspace sizes is performed. Finally, the fusion coefficient of CKFD is determined with respect to other chosen parameters.

To determine proper parameters for kernels, we use the global-to-local search strategy [2]. After globally searching over a wide range of the parameter space, we find a candidate

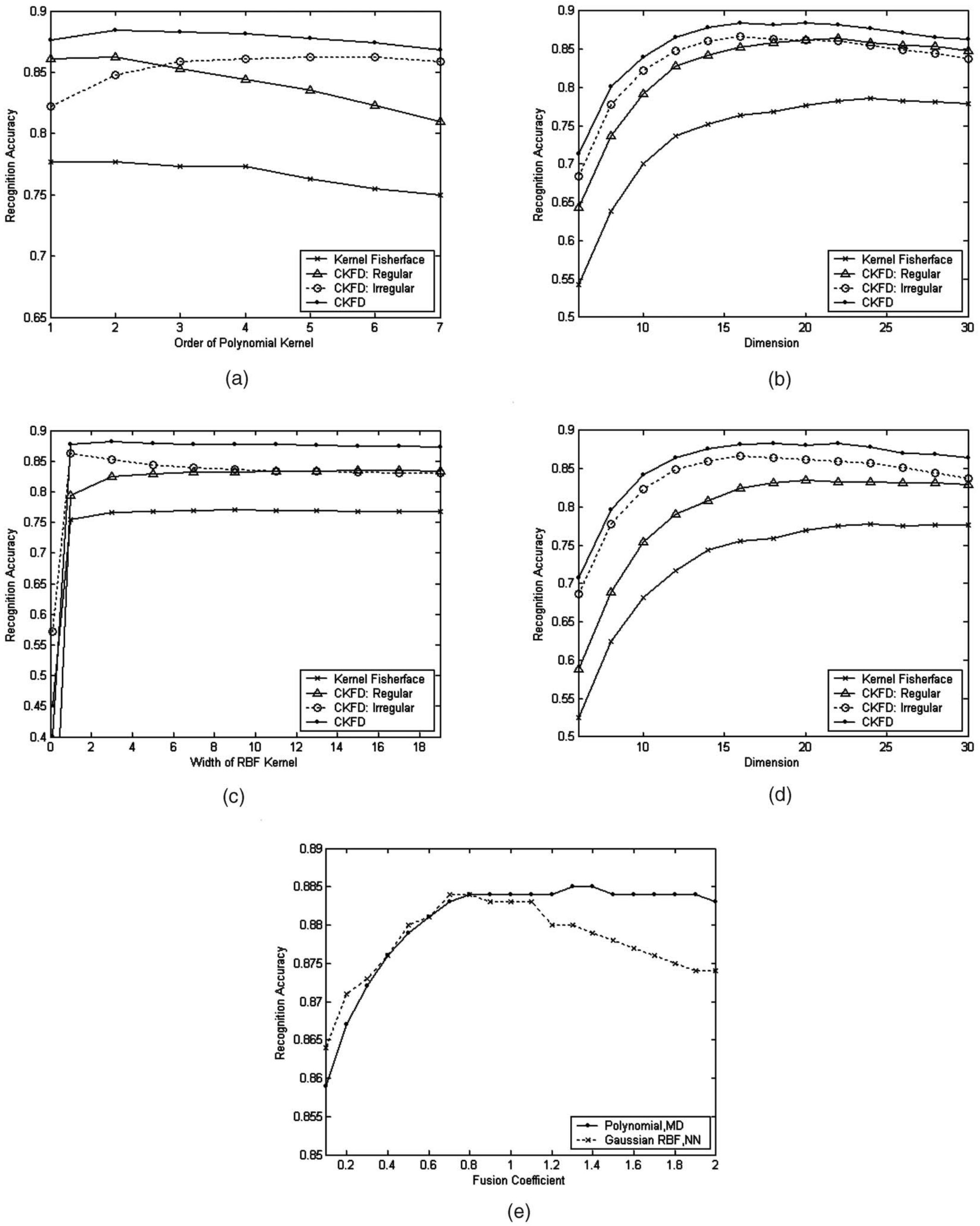


Fig. 3. Illustration of the recognition rates over the variation of kernel parameters, subspace dimensions and fusion coefficients in the model selection stage. (a) Recognition rates versus the order of the polynomial kernel, using minimum distance classifiers. (b) Recognition rates versus the subspace dimension, using the polynomial kernel and minimum distance classifiers. (c) Recognition rates versus the width of the Gaussian kernel, using nearest neighbor classifiers, (d) Recognition rates versus the subspace dimension, using the Gaussian kernel and nearest neighbor classifiers. (e) Recognition rates of CKFD versus the fusion coefficients.

interval where the optimal parameters might exist. Here, for the polynomial kernel, the candidate order interval is from 1 to 7 and, for the Gaussian RBF kernel, the candidate width interval is from 0.1 to 20. Then, we try to find the optimal kernel parameters within these intervals. Figs. 3a and 3c show the recognition accuracy versus the variation of kernel parameters corresponding to four methods with a fixed dimension of 20 and $\theta = 1$ for CKFD. From these figures, we can determine the proper kernel parameters. For example, the order of polynomial kernel should be two for CKFD with respect to a minimum distance classifier and the width of Gaussian kernel should be three for CKFD with respect to a nearest neighbor classifier.

By kernel parameter selection, we find that the nonlinear kernels are really helpful for improving the performance of CKFD. The results of CKFD with the linear kernel (i.e., the first order polynomial kernel), second order polynomial kernel, and Gaussian RBF kernel ($\delta = 3$) are listed in Table 1. From this table, it can be seen that CKFD with nonlinear kernels achieves better results under two different classifiers.

TABLE 1
Performance of CKFD (%) Using Different Kernels in the Model Selection Process

Classifier	Linear kernel	Polynomial kernel (r =2)	Gaussian kernel (δ =3)
MD	87.6	88.4	88.1
NN	87.1	88.1	88.5

TABLE 2
Optimal Parameters Corresponding to Each Method with Respect to Two Different Kernels and Classifiers

Method		Kernel Eigenface	Kernel Fisherface	CKFD: Irregular	CKFD: Regular	CKFD
Polynomial kernel	MD	[1, 190]	[2, 24]	[5, 16]	[2, 22]	[2, 20, 1.4]
	NN	[1, 180]	[2, 30]	[5, 16]	[1, 24]	[2, 18, 0.9]
Gaussian kernel	MD	[15, 190]	[19, 28]	[1, 16]	[7, 20]	[3, 20, 1]
	NN	[19, 180]	[9, 24]	[1, 16]	[15, 20]	[3, 18, 0.8]

(Note that the parameter set is arranged as [Degree (or Width), Subspace Dimension, Fusion Coefficient].)

TABLE 3
The Average Recognition Rates (%) of Kernel Eigenface, Kernel Fisherface, CKFD: *Regular*, CKFD: *Irregular*, and CKFD across 10 Tests and Their Standard Deviations (std)

Method		Kernel Eigenface	Kernel Fisherface	CKFD: Irregular	CKFD: Regular	CKFD
Polynomial kernel	MD	29.51 ± 2.47	78.11 ± 1.65	86.18 ± 2.03	85.95 ± 1.94	88.08 ± 1.73
	NN	25.38 ± 1.05	77.61 ± 1.68	86.18 ± 2.03	82.92 ± 1.86	88.26 ± 1.43
Gaussian kernel	MD	29.50 ± 2.45	77.93 ± 1.39	86.27 ± 1.91	86.33 ± 1.84	88.06 ± 1.59
	NN	25.33 ± 1.05	77.35 ± 1.21	86.27 ± 1.91	83.23 ± 1.46	88.38 ± 1.57

Moreover, Figs. 3a and 3c also show that 1) two kinds of discriminant features, *regular* and *irregular*, are both effective for discrimination. But, the variation trends of their performance versus the kernel parameters are different. When the polynomial kernel is used, the *regular* discriminant information degrades with the increase of orders (when the order is over 2), whereas the *irregular* discriminant information improves until the order is more than six. When the Gaussian kernel is used, the *regular* discriminant information enhances with the increase of widths, while the *irregular* discriminant information degrades after the width is more than 1. 2) After the fusion of two kinds of discriminant information, the performance is improved irrespective of the variation in the kernel parameters. This indicates that the *regular* and *irregular discriminant* features are complimentary for achieving a better result. 3) CKFD (even "CKFD: *regular*" or "CKFD: *irregular*") consistently outperforms Kernel Fisherface no matter what kernel is used.

After determining the kernel parameters, we set out to select the dimension of discriminant subspace. Let us depict the performance of each method over the variation of dimensions and show them in Figs. 3b and 3d. From these figures, we can choose the optimal subspace dimension for each method with respect to different kernels and classifiers. Besides, we find that CKFD irregular features seem more effective than regular ones when the subspace size is small (less than 16). Anyway, they both contribute to better results by fusion and are more powerful than Kernel Fisherface, no matter how many features are used. In order to fuse two kinds of discriminant information more effectively, in particular when they have significant different performance, we need to choose the fusion coefficients. The variation of performance of CKFD versus fusion coefficients with respect to two kernels and two classifiers is shown in Fig.3e. Obviously, the optimal fusion coefficient θ should be 1.4 for the polynomial kernel with a minimum distance classifier and 0.8 for the Gaussian kernel with a nearest neighbor classifier.

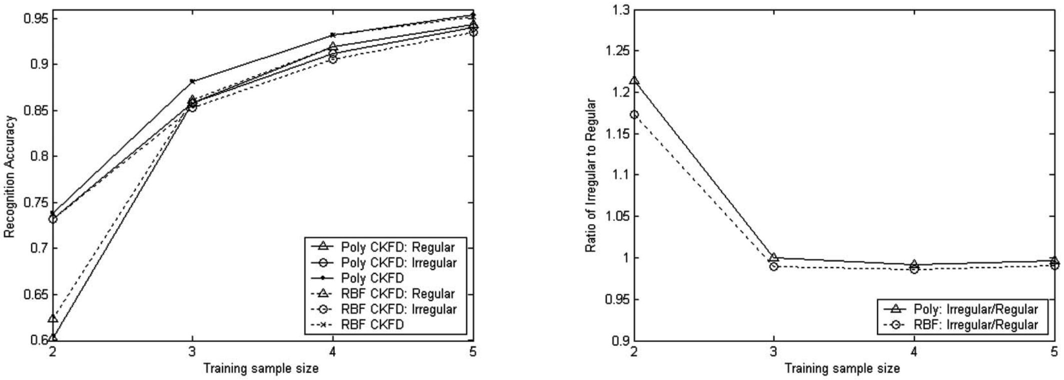
After model selection, we determine all parameters for each method with respect to different kernels and classifiers and list them in Table 2. With these parameters, all methods are reevaluated using an other 10 sets of training and testing samples. The average recognition rate and standard deviation (std) across 10 tests are shown in Table 3. From Table 3, we can see that the *irregular* discriminant features stand comparison with the regular ones with respect to their discriminatory power. Both kinds of discriminant features contribute to a better classification performance by virtue of fusion. All three CKFD versions outperform Kernel Eigenface and Kernel Fisherface.

Is CKFD statistically significantly better than other methods in terms of its recognition rate? To answer this question, let us evaluate the experimental results in Table 3 using McNemar's [39], [40], [41] significance test. McNemar's test is essentially a null hypothesis statistical test based on a Bernoulli model. If the resulting $p\text{-value}$ is below the desired significance level (for example, 0.02), the null hypothesis is rejected and the performance difference between two algorithms is considered to be statistically significant. By this test, we find that CKFD statistically significantly outperforms Kernel Eigenface and Kernel Fisherface at a significance level of $p = 1.036 \times 10^{-7}$.

In order to evaluate the computational efficiency of algorithms, we also give the average total CPU time of each

TABLE 4
The Average Total CPU Time (s) for Training and Testing Corresponding to Each Method under a Minimum Distance Classifier (CPU: Pentium 2.4HHz, RAM: 1Gb)

Method	Kernel Eigenface	Kernel Fisherface	CKFD: Irregular	CKFD: Regular	CKFD
Polynomial	76.959	79.878	89.392	91.544	93.834
Gaussian	116.047	118.988	128.328	131.113	133.028



(a) (b)

Fig. 4. (a) The performance of CKFD (regular, irregular, and fusion) with the variation of training sample sizes. (b) The ratio of the performance of "CKFD: irregular" to "CKFD: regular" with the variation of training sample sizes.

TABLE 5
The Optimal Parameters for Each Method

Method	GDA	CKFD: Irregular	CKFD: Regular	CKFD
Polynomial	[2, 9]	[2, 9]	[2, 9]	[2, 9, 0.9]
Gaussian	[60, 9]	[50, 9]	[90, 9]	[80, 9, 0.3]

(Note that the parameter set is arranged as [Degree (or Width), Subspace Dimension, Fusion Coefficient].)

method involved. Table 4 shows that CKFD (regular, irregular and fusion) algorithms are slightly slower than Kernel Fisherface and Kernel Eigenface.

By far, we can conclude that both regular and irregular discriminant features are effective for classification. They might, however, perform differently with the variation of kernel parameters, dimensions, and classifiers. If we fix these factors and allow the training sample size to vary, how about their performance? To address this issue, k $(k = 2, 3, 4, 5)$ images of each subject are randomly chosen for training and the others for testing. For each k, 10 training sample sets and the corresponding testing sample sets are generated. Then, we perform experiments using these sample sets. Here, a third-order polynomial kernel and Gaussian kernel with width $\delta = 3$ are adopted. The subspace dimension is fixed to be 20 and a minimum distance classifier is employed. The performance of CKFD (regular, irregular, and fusion) is depicted and shown in Fig. 4a. The ratio of the performance of "CKFD: irregular" to "CKFD: regular" is shown in Fig. 4b.

Fig. 4a indicates that the regular and irregular discriminant information both increases with the increase of training sample sizes. The fusion strategy remains effective irrespective of the variation in training sample sizes. Fig. 4b indicates that the ratio of the performance of "CKFD: irregular" to "CKFD: regular" is large when the training sample size is small (equal to 2 here). That is, the smaller the training sample size is, the more powerful the irregular discriminant features are. When the training sample size becomes larger (more than 2), the ratio curve levels off.

5.2 Experiment on Handwritten Digit Classification Using CENPARMI Database

In this experiment, we use the Concordia University CENPARMI handwritten numeral database [42], [44]. This database contains 6,000 samples of 10 numeral classes (each class has 600 samples). Here, our experiment is performed based on 256-dimensional Gabor transformation features [43], [44], which turned out to be effective for handwritten digit classification.

In our experiments, 100 samples are randomly chosen from each class for training, while the remaining 500 samples are used for testing. Thus, the training sample set size is 1,000 and the testing sample set size is 5,000. We run the system 10 times and obtain 10 different training and testing sample sets for performance evaluation. Here, the polynomial kernel and Gaussian RBF kernel are both involved. The standard LDA [35], GDA [6], and three versions of CKFD (regular, irregular, and fusion) are tested and evaluated. A minimum distance classifier is employed for computational efficiency.

The model selection process is performed using the same method described in Section 5.1. The optimal parameters corresponding to each method are obtained and listed in Table 5. Based on these parameters, GDA and three

TABLE 6
The Average Recognition Rates (%) of Each Method across 10 Tests on the CENPARMI Database and the Standard Deviations (std)

Method	GDA	CKFD: Irregular	CKFD: Regular	CKFD	LDA
Polynomial	81.87 ± 2.95	84.48 ± 0.84	83.80 ± 0.62	86.96 ± 0.50	76.34 ± 0.77
Gaussian	87.64 ± 2.35	88.59 ± 0.26	84.27 ± 0.58	88.79 ± 0.31	

TABLE 7
The Average Total CPU Time (s) for Training and Testing of Each Method

Method	GDA	CKFD: Irregular	CKFD: Regular	CKFD	LDA
Polynomial	177.536	244.583	245.101	247.953	14.508
Gaussian	220.216	294.294	298.438	305.538	

TABLE 8
Comparison of LDA, KFD [4], KPCA, and CKFD on Some Digit-Pairs Drawn from the CENPARMI Database

Digit-pairs	LDA	KFD [4]	KPCA	CKFD
{1, 7}	92.3	93.8	57.7	97.9
{1, 9}	93.5	95.3	55.4	98.2
{2, 3}	73.1	87.0	51.1	92.5
{3, 5}	79.6	92.3	75.2	93.8
{4, 9}	89.8	95.3	73.1	96.8
{0, 6}	83.8	91.5	67.3	92.1

versions of CKFD are tested. The average recognition rates across 10 tests and the standard deviations (std) are listed in Table 6. The average total CPU time consumed by each method is shown in Table 7.

From Table 6, it can be seen that 1) both kinds of discriminant features (regular and irregular) are very effective for classification. 2) After their fusion, the discriminatory power is dramatically enhanced with the polynomial kernel and slightly enhanced with the Gaussian RBF kernel. 3) CKFD performs better than GDA and LDA and the performance difference between CKFD and GDA is statistically significant when polynomial kernel is used. Besides, the standard deviation of CKFD is much smaller than that of GDA. These conclusions are consistent with those in face recognition experiments on the whole. Table 7 indicates that CKFD versions (regular, irregular, and fusion) consume more CPU time than GDA for training and testing. This is because CKFD needs additional computation for space decomposition in Step 2. In addition, all kernel-based methods are much more time-consuming than linear method LDA.

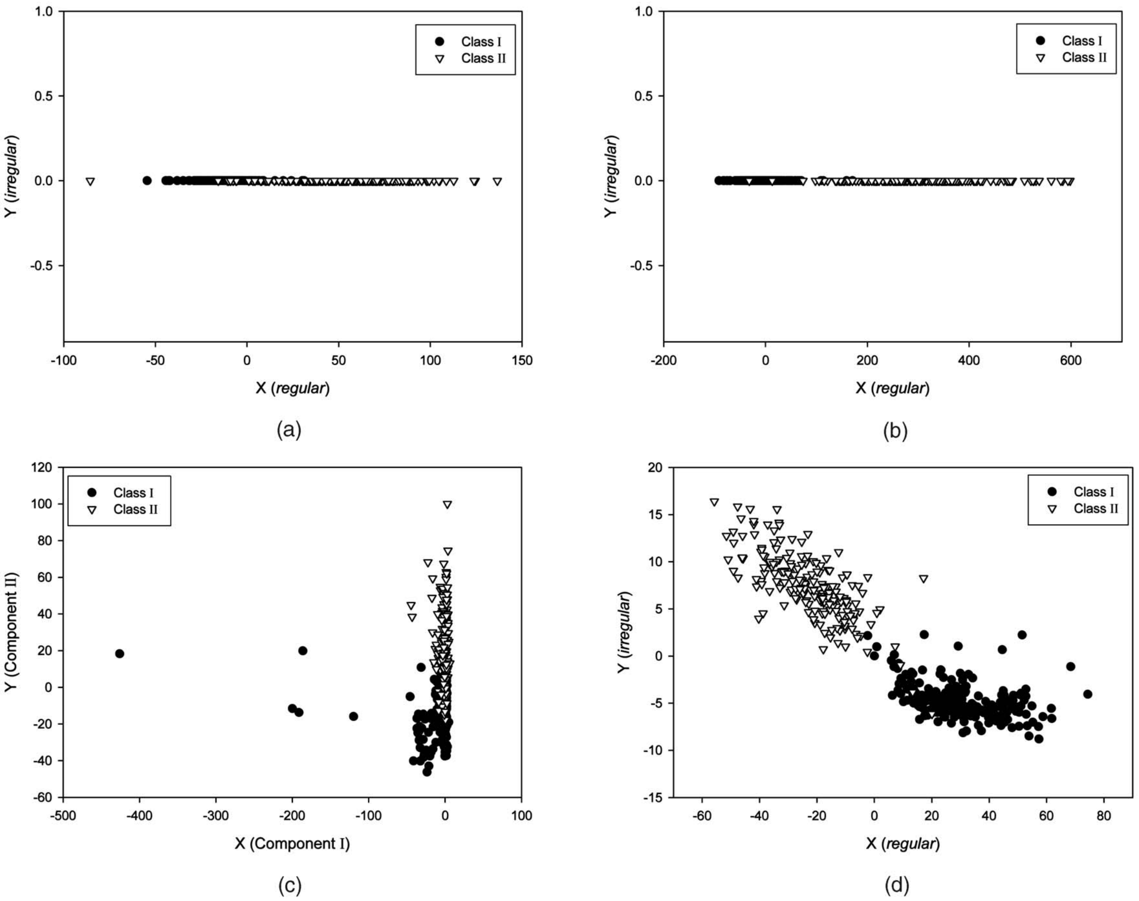
5.3 Experiments on Two-Class Problems: An Example of CKFD-Based Two-Dimensional Exhibition

In this section, we will do some tests on two-class problems. For convenience, the CENPARMI handwritten numeral database is employed again. We will not use a whole database, but draw the samples of some digit-pairs for experiments this time. For instance, all samples of "1" and "7" are taken out to form a digit-pair subset. The algorithms are tested on this subset and used to classify the two-class patterns: "1" and "7".

Some easily confused digit-pairs are first chosen, as listed in the first column of Table 8. Then, LDA, KFD [4], KPCA [3], and CKFD are tested on these digit-pairs subsets. For each subset, the first 100 samples per class are used for training and the remaining 500 samples are for testing. Thus, the total number of training samples is 200 while the total number of testing samples is 1,000. For KFD [4], to overcome the singularity, the inner product matrix $\mathbf{N}$ (induced by the within-class scatter matrix) is added by a multiple of the identity matrix, i.e., $\mathbf{N}_\mu = \mathbf{N} + \mu\,\mathbf{I}$. Here, the parameter μ is chosen as $\mu = 10^{-3}$. Since there are only 200 training samples, the within-class scatter matrix of LDA is also singular. We adopt the same technique as that in KFD to regularize it.

Note that, for two-class cases, LDA and KFD can get only one discriminant axis, whereas CKFD can obtain two discriminant axes (one is regular and the other is irregular). After the projection of samples onto the discriminant axes, the resulting LDA and KFD features are one-dimensional while CKFD features are two-dimensional. Although KPCA can extract multidimensional features, for convenience of comparison in the following data visualization example, only two principal components are used in our tests. The classification results corresponding to each method with a second order polynomial kernel and a minimum distance classifier are listed in Table 8. Table 8 shows CKFD outperforms other methods.

Now, taking digit-pair $\{1, 7\}$ as an example, we plot the scatter of 400 testing samples (200 ones per class) as they are



(a)

(b)

(c)

(d)

Fig. 5. The scatter plots of two-class samples as they are projected onto LDA, KFD [4], KPCA (two principal components are used), and CKFD spaces. (a) The scatter plot of LDA, where 29 samples are misclassified. (b) The scatter plot of KFD, where 20 samples are misclassified. (c) The scatter plot of KPCA, where 174 samples are misclassified. (d) The scatter plot of CKFD, where six samples are misclassified.

projected onto the discriminant space (or principal component space), as shown in Fig. 5. For LDA and KFD, since there is only one discriminant axis, the projected samples scatter on a line. Differently, CKFD enables the data to scatter on a plane that is spanned by the *regular* and *irregular* discriminant axes. It can obviously be seen from Fig. 5 that the data are more separable in the two-dimensional CKFD space than in the one-dimensional KFD or LDA space. Actually, this separability can be simply measured by the classification errors. There are only six errors of CKFD while there are 20 errors of KFD and 29 errors of LDA. This indicates that the dimensional increase of discriminant space is really helpful for discrimination. However, KPCA does not perform well, although it can also exhibit the data in two-dimensional mode. Fig. 5c shows the two-class patterns are badly overlapped in the KPCA-based space. Therefore, we can conclude that the KPCA principal component features are not very discriminatory for classification tasks.

6 Conclusion, Discussion, and Future Work

A new KFD framework—KPCA plus LDA—is developed in this paper. Under this framework, a two-phase KFD algorithm is presented. Actually, based on the developed KFD framework, a series of existing KFD algorithms can be reformulated in alternative ways. In other words, it is easy to give the equivalent versions of the previous KFD algorithms. Taking kernel Fisherface as an example, we can first use KPCA to reduce the dimension to l (note that here only l components are used; l is subject to $c \leq l \leq M - c$, where M is the number of training samples and c is the number of classes) and then perform standard LDA in the KPCA-transformed space. Similarly, we can construct alternative versions for others. These versions make it easier to understand and implement kernel Fisher discriminant, particularly for the new investigator or programmer.

A complete KFD algorithm (CKFD) is proposed to implement the KPCA plus LDA strategy. This algorithm allows us to perform discriminant analysis in "double discriminant subspaces": regular and irregular. The previous KFD algorithms always emphasize the former and neglect the latter. In fact, the irregular discriminant subspace contains important discriminative information which is as powerful as the regular discriminant subspace. This has been demonstrated by our experiments. It should be emphasized that, for kernel-based discriminant analysis, the two kinds of discriminant information (particularly the irregular one) are widely existent, not limited to small sample size problems like face recognition. Our experiment on handwritten digit recognition shows that CKFD is suitable for "large sample size" problems (in observation space) as well. The underlying reason is that the implicit nonlinear mapping determined by "kernel" always turns large sample size problems in observation space into small sample size ones in *feature space*. More interestingly, the two discriminant subspaces of CKFD turn out to be mutually complementary for discrimination despite the fact that each of them can work well independently. The fusion of two kinds of discriminant information can achieve better results.

Especially for small sample size problems, CKFD is exactly in tune with the existing two-phase LDA algorithms that are based on *PCA plus LDA* framework. Actually, if a linear kernel, i.e., $k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$, is adopted instead of nonlinear kernels, CKFD would degenerate to be a *PCA plus LDA algorithm* like that in [24]. Therefore, the existing two-phase LDA (*PCA plus LDA*) algorithms can be viewed as a special case of CKFD.

Finally, we have to point out that the computational efficiency of CKFD is a problem deserving further investigation. Actually, all kernel-based methods, including KPCA [3], GDA [6], and KFD [4], encounter the same problem. This is because all kernel-based discriminant methods have to solve an $M \times M$ sized eigenproblem (or generalized eigenproblem). When the sample size M is fairly large, it becomes very computationally intensive [10]. Several ways suggested by Mika et al. [10] and Burges and Schölkopf [45] can be used to deal with this problem, but the optimal implementation scheme (e.g., developing a more efficient numerical algorithm for large scale eigenproblems) is still open.

Appendix A
The Proof of Lemma 1

Proof. For simplicity, let us denote $\mathbf{T} = M\mathbf{S}_t^\Phi$ and $g_j = \Phi(\mathbf{x}_j) - \mathbf{m}_0^\Phi$. Then, $\mathbf{T} = \sum_{j=1}^M g_j g_j^{\mathrm{T}}$.

1. For every $f \in \mathcal{H}$, we have $\mathbf{T}f = \sum_{j=1}^M \langle g_j, f \rangle g_j$. Since

$$||\mathbf{T}f|| \leq \sum_{j=1}^M |\langle g_j, f \rangle| \; ||g_j|| \leq \; ||f|| \; \sum_{j=1}^M ||g_j||^2,$$

$\mathbf{T}$ is bounded and $||\mathbf{T}|| \leq \; \sum_{j=1}^M ||g_j||^2$.

2. Let us consider the range of the operator $\mathbf{T}$: $\mathcal{R}(\mathbf{T}) = \{\mathbf{T}f, f \in \mathcal{H}\}$. Since

$$\mathbf{T}f = \sum_{j=1}^M \langle g_j, f \rangle g_j, \mathcal{R}(\mathbf{T}) = \mathcal{L}(g_1, \ldots, g_M),$$

which is the generated space by $g_1, \ldots, g_M$. So, dim $\mathcal{R}(\mathbf{T}) \leq M < \infty$, which implies that $\mathbf{T}$ is a compact operator [47].

3. For every $f \in \mathcal{H}$, we have

$$\langle \mathbf{T}f, f \rangle = \sum_{j=1}^M \langle g_j, f \rangle \langle g_j, f \rangle = \sum_{j=1}^M \langle g_j, f \rangle^2 \geq 0.$$

Thus, $\mathbf{T}$ is a positive operator on Hilbert space $\mathcal{H}$.

4. Since $\mathbf{T}$ is a positive operator, it must be self-adjoint (symmetric) because its adjoint $\mathbf{T}^* = \mathbf{T}$ (see [48]).

Since $\mathbf{S}_t^\Phi$ has the same properties as $\mathbf{T}$, Lemma 1 is proven. □

Appendix B
The Proof of Theorem 2

In order to verify Theorem 2, let us introduce two lemmas first.

Lemma B1. *$\varphi^{\mathrm{T}}\mathbf{S}_t^\Phi\varphi = 0$ if and only if $\varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi = 0$ and $\varphi^{\mathrm{T}}\mathbf{S}_w^\Phi\varphi = 0$.*

Proof. Since $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ are both positive and $\mathbf{S}_t^\Phi = \mathbf{S}_b^\Phi + \mathbf{S}_w^\Phi$, it is easy to verify this. □

Lemma B2 [47]. *Suppose that $\mathbf{A}$ is a positive operator. Then, $x^{\mathrm{T}}\mathbf{A}x = 0$ if and only if $\mathbf{A}x = \mathbf{0}$.*

Proof. If $\mathbf{A}x = \mathbf{0}$, it is obvious that $x^{\mathrm{T}}\mathbf{A}x = 0$. So, we only need to prove that $x^{\mathrm{T}}\mathbf{A}x = 0 \Rightarrow \mathbf{A}x = \mathbf{0}$. Since $\mathbf{A}$ is a positive operator, it must have a positive square root T [47], such that $\mathbf{A} = \mathbf{T}^2$. Thus, $\langle \mathbf{T}x, \mathbf{T}x \rangle = \langle \mathbf{A}x, x \rangle = x^{\mathrm{T}}\mathbf{A}x = 0$. So, $\mathbf{T}x = \mathbf{0}$, from which it follows that $\mathbf{A}x = \mathbf{T}(\mathbf{T}x) = \mathbf{0}$. □

The Proof of Theorem 2. Since $\Psi_t^\perp$ is the null space of $\mathbf{S}_t^\Phi$, for every $\zeta \in \Psi_t^\perp$, we have $\zeta^{\mathrm{T}}\mathbf{S}_t^\Phi\zeta = 0$.

From Lemma B1, it follows that $\zeta^{\mathrm{T}}\mathbf{S}_b^\Phi\zeta = 0$. Since $\mathbf{S}_b^\Phi$ is a positive operator according to Lemma 2, we have $\mathbf{S}_b^\Phi\zeta = 0$ by Lemma B2. Hence,

$$\varphi^{\mathrm{T}}\mathbf{S}_b^\Phi\varphi = \phi^{\mathrm{T}}\mathbf{S}_b^\Phi\phi + 2\phi^{\mathrm{T}}\mathbf{S}_b^\Phi\zeta + \zeta^{\mathrm{T}}\mathbf{S}_b^\Phi\zeta = \phi^{\mathrm{T}}\mathbf{S}_b^\Phi\phi.$$

Similarly, we have

$$\varphi^{\mathrm{T}}\mathbf{S}_w^\Phi\varphi = \phi^{\mathrm{T}}\mathbf{S}_w^\Phi\phi + 2\phi^{\mathrm{T}}\mathbf{S}_w^\Phi\zeta + \zeta^{\mathrm{T}}\mathbf{S}_w^\Phi\zeta = \phi^{\mathrm{T}}\mathbf{S}_w^\Phi\phi.$$

So, $J_b^\Phi(\varphi) = J_b^\Phi(\phi)$ and $J^\Phi(\varphi) = J^\Phi(\phi)$. □

Appendix C
The Proof of Lemma 3

Proof. Since $\mathbf{S}_t^\Phi$ is a compact and positive operator from Lemma 1, the total scatter matrix $\mathbf{S}_t$ in the KPCA-transformation space $\mathbb{R}^m$ can be represented by $\mathbf{S}_t = \mathbf{P}^{\mathrm{T}}\mathbf{S}_t^\Phi\mathbf{P} = diag(\lambda_1, \lambda_2, \ldots, \lambda_m)$, where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the positive eigenvalues of $\mathbf{S}_t^\Phi$.

So, $\mathbf{S}_t$ is a positive definite matrix in $\mathbb{R}^m$. This means that, for every nonzero vector $\eta \in \mathbb{R}^m$, $\eta^{\mathrm{T}}\mathbf{S}_t\eta > 0$ always holds.

Obviously, for every nonzero vector $\eta \in \Theta_w$, $\eta^{\mathrm{T}}\mathbf{S}_w\eta = 0$ always holds.

Since $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$, for every nonzero vector $\eta \in \Theta_w$, we have $\eta^{\mathrm{T}}\mathbf{S}_b\eta = \eta^{\mathrm{T}}\mathbf{S}_t\eta - \eta^{\mathrm{T}}\mathbf{S}_w\eta > 0$. □

Acknowledgments

This research was supported by the UGC/CRC fund from the HKSAR Government and the central fund from the Hong Kong Polytechnic University, the National Science Foundation of China under Grant No. 60332010, 60472060, 60473039, Grant TIC2002-04495-C02, from the Spanish Ministry of Science and Technology (MCyT), and Grant IM3-G03/185 from the Spanish ISCIII. Dr. Jin and Dr. Frangi are also supported by RyC Research Fellowships from MCyT. The authors would also like to thank all of the anonymous reviewers and the editor for their constructive advice.

REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer, 1995.

[2] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Trans. Neural Networks,* vol. 12, no. 2, pp. 181-201, 2001.

[3] B. Schölkopf, A. Smola, and K.R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. 10, no. 5, pp. 1299-1319, 1998.

[4] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher Discriminant Analysis with Kernels," *Proc. IEEE Int'l Workshop Neural Networks for Signal Processing IX,* pp. 41-48, Aug. 1999.

[5] S. Mika, G. Rätsch, B. Schölkopf, A. Smola, J. Weston, and K.-R. Müller, "Invariant Feature Extraction and Classification in Kernel Spaces," *Advances in Neural Information Processing Systems 12,* Cambridge, Mass.: MIT Press, 1999.

[6] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," *Neural Computation,* vol. 12, no. 10, pp. 2385-2404, 2000.

[7] V. Roth and V. Steinhage, "Nonlinear Discriminant Analysis Using Kernel Functions," *Advances in Neural Information Processing Systems,* S.A. Solla, T.K. Leen, and K.-R. Mueller, eds., vol. 12, pp. 568-574, MIT Press, 2000.

[8] S. Mika, G. Ratsch, and K.-R. Müller, "A Mathematical Programming Approach to the Kernel Fisher Algorithm," *Advances in Neural Information Processing Systems 13,* T.K. Leen, T.G. Dietterich, and V. Tresp, eds., pp. 591-597, MIT Press, 2001.

[9] S. Mika, A.J. Smola, and B. Schölkopf, "An Improved Training Algorithm for Kernel Fisher Discriminants," *Proc. Eighth Int'l Workshop Artificial Intelligence and Statistics,* T. Jaakkola and T. Richardson, eds., pp. 98-104, 2001.

[10] S. Mika, G. Rätsch, J Weston, B. Schölkopf, A. Smola, and K.-R. Müller, "Constructing Descriptive and Discriminative Nonlinear Features: Rayleigh Coefficients in Kernel Feature Spaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 5, pp. 623-628, May 2003.

[11] M.H. Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods," *Proc. Fifth IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 215-220, May 2002.

[12] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms," *IEEE Trans. Neural Networks,* vol. 14, no. 1, pp. 117-126, 2003.

[13] J. Xu, X. Zhang, and Y. Li, "Kernel MSE Algorithm: A Unified Framework for KFD, LS-SVM, and KRR," *Proc. Int'l Joint Conf. Neural Networks,* pp. 1486-1491, July 2001.

[14] S.A. Billings and K.L Lee, "Nonlinear Fisher Discriminant Analysis Using a Minimum Squared Error Cost Function and the Orthogonal Least Squares Algorithm," *Neural Networks,* vol. 15, no. 2, pp. 263-270, 2002.

[15] T.V. Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vanderwalle, "Bayesian Framework for Least Squares Support Vector Machine Classifiers, Gaussian Processes and Kernel Fisher Discriminant Analysis," *Neural Computation,* vol. 15, no. 5, pp. 1115-1148, May 2002.

[16] G.C. Cawley and N.L.C. Talbot, "Efficient Leave-One-Out Cross-Validation of Kernel Fisher Discriminant Classifiers," *Pattern Recognition,* vol. 36, no. 11, pp. 2585-2592, 2003.

[17] N.D. Lawrence and B. Schölkopf, "Estimating a Kernel Fisher Discriminant in the Presence of Label Noise," *Proc. 18th Int'l Conf. Machine Learning,* pp. 306-313, 2001.

[18] A.N. Tikhonov and V.Y. Arsenin, *Solution of Ill-Posed Problems.* New York: Wiley, 1997.

[19] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, Aug. 1996.

[20] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriengman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[21] K. Liu, Y.-Q. Cheng, J.-Y. Yang, and X. Liu, "An Efficient Algorithm for Foley-Sammon Optimal Set of Discriminant Vectors by Algebraic Method," *Int'l J. Pattern Recognition and Artificial Intelligence,* vol. 6, no. 5, pp. 817-829, 1992.

[22] L.F. Chen, H.Y.M. Liao, J.C. Lin, M.D. Kao, and G.J. Yu, "A New LDA-Based Face Recognition System which Can Solve the Small Sample Size Problem," *Pattern Recognition,* vol. 33, no. 10, pp. 1713-1726, 2000.

[23] H. Yu and J. Yang, "A Direct LDA Algorithm for High-Dimensional Data—With Application to Face Recognition," *Pattern Recognition,* vol. 34, no. 10, pp. 2067-2070, 2001.

[24] J. Yang and J.Y. Yang, "Why Can LDA Be Performed in PCA Transformed Space?" *Pattern Recognition,* vol. 36, no. 2, pp. 563-566, 2003.

[25] J. Yang and J.Y. Yang, "Optimal FLD Algorithm for Facial Feature Extraction," *Proc. SPIE Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision,* pp. 438-444, Oct. 2001.

[26] C.J. Liu and H. Wechsler, "A Shape- and Texture-Based Enhanced Fisher Classifier for Face Recognition," *IEEE Trans. Image Processing,* vol. 10, no. 4, pp. 598-608, 2001.

[27] C.J. Liu and H. Wechsler, "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases," *IEEE Trans. Image Processing,* vol. 9, no. 1, pp. 132-137, 2000.

[28] W. Zhao, R. Chellappa, and J. Phillips, "Subspace Linear Discriminant Analysis for Face Recognition," Technical Report CS-TR4009, Univ. of Maryland, 1999.

[29] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng, "Discriminant Analysis of Principal Components for Face Recognition," *Face Recognition: From Theory to Applications,* H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie and T. S. Huang, eds., pp. 73-85, Springer-Verlag, 1998.

[30] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[31] M. Kirby and L. Sirovich, "Application of the KL Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp. 103-108, Jan. 1990.

[32] J. Yang, A.F. Frangi, and J.-Y. Yang, "A New Kernel Fisher Discriminant Algorithm with Application to Face Recognition," *Neurocomputing,* vol. 56, pp. 415-421, 2004.

[33] G.H. Golub and C.F. Van Loan, *Matrix Computations,* third ed. Baltimore and London: The Johns Hopkins Univ. Press, 1996.

[34] P. Lancaster and M. Tismenetsky, *The Theory of Matrices,* second ed. Orlando, Fla.: Academic Press, 1985.

[35] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Boston: Academic Press, 1990.

[36] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

[37] P.J. Phillips, "The Facial Recognition Technology (FERET) Database," http://www.itl.nist.gov/iad/humanid/feret/feret_master.html, 2004.

[38] J. Yang, D. Zhang, and J.-y. Yang, "A Generalized K-L Expansion Method which Can Deal with Small Sample Size and High-Dimensional Problems," *Pattern Analysis and Application,* vol. 6, no. 1, pp. 47-54, 2003.

[39] W. Yambor, B. Draper, and R. Beveridge, "Analyzing PCA-Based Face Recognition Algorithms: Eigenvector Selection and Distance Measures," *Empirical Evaluation Methods in Computer Vision,* H. Christensen and J. Phillips, eds., Singapore: World Scientific Press, 2002.

[40] J. Devore and R. Peck, *Statistics: The Exploration and Analysis of Data,* third ed. Brooks Cole, 1997.

[41] B.A. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge, "Recognizing Faces with PCA and ICA," *Computer Vision and Image Understanding,* vol. 91, no. 1-2, pp. 115-137, 2003.

[42] Z. Lou, K. Liu, J.Y. Yang, and C.Y. Suen, "Rejection Criteria and Pairwise Discrimination of Handwritten Numerals Based on Structural Features," *Pattern Analysis and Applications,* vol. 2, no. 3, pp. 228-238, 1992.

[43] Y. Hamamoto, S. Uchimura, M. Watanabe, T. Yasuda, and S. Tomita, "Recognition of Handwritten Numerals Using Gabor Features," *Proc. 13th Int'l Conf. Pattern Recognition,* pp. 250-253, Aug. 1996.

[44] J. Yang, J.-y. Yang, D. Zhang, and J.F. Lu, "Feature Fusion: Parallel Strategy vs. Serial Strategy," *Pattern Recognition,* vol. 36, no. 6, pp. 1369-1381, 2003.

[45] C.J.C. Burges and B. Schölkopf, "Improving the Accuracy and Speed of Support Vector Learning Machines," *Advances in Neural Information Processing Systems 9,* M. Mozer, M. Jordan, and T. Petsche, eds., pp. 375-381, Cambridge, Mass.: MIT Press, 1997.

[46] B. Schölkopf and A. Smola, *Learning with Kernels.* Cambridge, Mass.: MIT Press, 2002.

[47] E. Kreyszig, *Introductory Functional Analysis with Applications.* John Wiley & Sons, 1978.

[48] W. Rudin, *Functional Analysis.* McGraw-Hill, 1973.

[49] V. Hutson and J.S. Pym, *Applications of Functional Analysis and Operator Theory.* London: Academic Press, 1980.

[50] J. Weidmann, *Linear Operators in Hilbert Spaces.* New York: Springer-Verlag, 1980.

[51] J. Yang, J.-y. Yang, and A.F. Frangi, "Combined Fisherfaces Framework," *Image and Vision Computing,* vol. 21, no. 12, pp. 1037-1044, 2003.

[52] J. Yang, J.-y. Yang, and H. Ye, "Theory of Fisher Linear Discriminant Analysis and Its Application," *Acta Automatica Sinica,* vol. 29, no. 4, pp. 481-494, 2003 (in Chinese).

Jian Yang received the BS degree in mathematics from the Xuzhou Normal University in 1995. He received the MS degree in applied mathematics from the Changsha Railway University in 1998 and the PhD degree from the Nanjing University of Science and Technology (NUST) Department of Computer Science, on the subject of pattern recognition and intelligence systems in 2002. From March to September 2002, he worked as a research assistant in the Department of Computing, Hong Kong Polytechnic University. From January to December 2003, he was a postdoctoral researcher at the University of Zaragoza and affiliated with the Division of Bioengineering of the Aragon Institute of Engineering Research (I3A). In the same year, he was awarded the RyC program Research Fellowship, sponsored by the Spanish Ministry of Science and Technology. Now, he is a research associate at the Hong Kong Polytechnic University with the biometrics center and a postdoctoral research fellow at NUST. He is the author of more than 30 scientific papers in pattern recognition and computer vision. His current research interests include pattern recognition, computer vision, and machine learning.

Alejandro F. Frangi received the MSc degree in telecommunication engineering from the Universitat Politecnica de Catalunya, Barcelona, in 1996, where he subsequently did research on electrical impedance tomography for image reconstruction and noise characterization. In 2001, he received the PhD degree from the Image Sciences Institute of the University Medical Center Utrecht on model-based cardiovascular image analysis. From 2001 to 2004, he was an assistant professor and Ramón y Cajal Research Fellow at the University of Zaragoza, Spain, where he cofounded and codirected the Computer Vision Group of the Aragon Institute of Engineering Research (I3A). He is now a Ramon y Cajal Research Fellow at the Pompeu Fabra University in Barcelona, Spain, where he directs the Computational Imaging Lab in the Department of Technology. His main research interests are in computer vision and medical image analysis, with particular emphasis on model and registration-based techniques. He is an associate editor of the *IEEE Transactions on Medical Imaging* and has served twice as guest editor for special issues of the same journal. He is also member of the Biosecure European Excellence Network on Biometrics for Secure Authentication (www.biosecure.info).

Jing-yu Yang received the BS degree in computer science from Nanjing University of Science and Technology (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 in the Department of Computer Science, Missuria University. And, 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 300 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.

David Zhang graduated in computer science from Peking University in 1974 and received the MSc and PhD degrees in computer science and engineering from the Harbin Institute of Technology (HIT) in 1983 and 1985, respectively. He received a second PhD degree in electrical and computer engineering from the University of Waterloo, Ontario, Canada, in 1994. After that, he was an associate professor at the City University of Hong Kong and a professor at the Hong Kong Polytechnic University. Currently, he is a founder and director of the Biometrics Technology Centre supported by the UGC of the Government of the Hong Kong SAR. He is the founder and editor-in-chief of the *International Journal of Image and Graphics* and an associate editor for some international journals such as the *IEEE Transactions on Systems, Man, and Cybernetics*, *Pattern Recognition*, and *International Journal of Pattern Recognition and Artificial Intelligence*. His research interests include automated biometrics-based identification, neural systems and applications, and image processing and pattern recognition. So far, he has published more than 180 papers as well as 10 books, and won numerous prizes. He is a senior member of the IEEE and the IEEE Computer Society.

Zhong Jin received the BS degree in mathematics, the MS degree in applied mathematics, and the PhD degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 1982, 1984, and 1999, respectively. He is a professor in the Department of Computer Science, NUST. He visited the Department of Computer Science and Engineering, The Chinese University of Hong Kong from January 2000 to June 2000 and from November 2000 to August 2001 and visited the Laboratoire HEUDIASYC, UMR CNRS 6599, Universite de Technologie de Compiegne, France, from October 2001 to July 2002. Dr. Jin is now visiting the Centre de Visio per Computador, Universitat Autonoma de Barcelona, Spain, as the Ramon y Cajal program Research Fellow. His current interests are in the areas of pattern recognition, computer vision, face recognition, facial expression analysis, and content-based image retrieval.