



Kernel quadratic discriminant analysis for small sample size problem

Jie Wang^{a,*}, K.N. Plataniotis^a, Juwei Lu^b, A.N. Venetsanopoulos^c

^a*Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Canada M5A 3G4*

^b*Vidient Systems, Inc., 4000 Burton Dr., Santa Clara, CA 95054, USA*

^c*Ryerson University, 360 Victoria St, Toronto, Canada M5B 2K3*

Received 13 January 2007; received in revised form 31 July 2007; accepted 23 October 2007

Abstract

It is generally believed that quadratic discriminant analysis (QDA) can better fit the data in practical pattern recognition applications compared to linear discriminant analysis (LDA) method. This is due to the fact that QDA relaxes the assumption made by LDA-based methods that the covariance matrix for each class is identical. However, it still assumes that the class conditional distribution is Gaussian which is usually not the case in many real-world applications. In this paper, a novel kernel-based QDA method is proposed to further relax the Gaussian assumption by using the kernel machine technique. The proposed method solves the complex pattern recognition problem by combining the QDA solution and the kernel machine technique, and at the same time, tackles the so-called small sample size problem through a regularized estimation of the covariance matrix. Extensive experimental results indicate that the proposed method is a more sophisticated solution outperforming many traditional kernel-based learning algorithms.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Linear discriminant analysis; Quadratic discriminant analysis; Small sample size; Kernel machine technique

1. Introduction

Linear discriminant analysis (LDA) is one of the most effective feature extraction methods in statistical pattern recognition learning [1–5]. LDA-based methods extract the discriminant features by maximizing the so-called Fisher's criterion which is defined as the ratio of between- and within-class scatter matrices. It has been shown that LDA can be viewed as a special case of the optimal Bayesian classifier when the class conditional distribution of the data are Gaussian with identical covariance structure [6,7]. In such cases, the resulting class boundaries are restricted to linear hyper-planes. In many practical pattern recognition problems, however, the distribution of the data is far more complicated than Gaussian, usually multi-modal and non-convex. In such cases, the performance of LDA-based techniques deteriorates dramatically [6]. To this end, nonlinear techniques appear to be more effective to handle complex distributed data samples.

Quadratic discriminant analysis (QDA) is one of the most commonly used nonlinear techniques for pattern classification.

In the QDA framework, the class conditional distribution is assumed to be Gaussian, however, with an allowance for different covariance matrices. In such cases, a more complex quadratic boundary can be formed. It is therefore reasonable to believe that QDA better fits the real data structure. However, due to the fact that more free parameters are to be estimated (C covariance matrices, where C denotes the number of classes) compared to those in an LDA-based solution (1 covariance matrix), QDA is more susceptible to the so-called small sample size (SSS) problem such that the number of training samples is smaller or comparable to the dimensionality of the sample space [8–10,2]. Under the SSS scenario, a reliable estimation of the system parameters, such as the within- and between-class matrices in LDA-based methods, becomes a difficult task. The SSS problem is a realistic problem existing in many real-world applications such as face recognition. Under the SSS scenario, a direct application of QDA is difficult or even mathematically intractable as the estimates of the covariance matrices for each class become either ill- or poorly posed [6]. In order to apply QDA to the pattern recognition problem under the SSS scenario, regularized direct QDA (denoted as RD-QDA) [6] has been proposed to tackle the SSS problem in two

* Corresponding author. Tel.: +1 416 893 0323; fax: +1 416 978 4425.
E-mail address: wjiewjie@gmail.com (J. Wang).

steps: dimensionality reduction and regularized covariance matrix estimation. It has been shown that the discriminatory information resides in the intersection of the null space of the within-class scatter (denoted as \mathcal{A}) and the complement space of the null space of the between-class scatter (denoted as \mathcal{B}'), i.e., $\mathcal{A} \cap \mathcal{B}'$ [3–5]. At the same time, no significant information, in terms of maximization of Fisher's criterion, will be lost if the null space of the between-class scatter is discarded. The low-dimensional between-class subspace \mathcal{B}' is therefore first extracted. The RD-QDA method is thus a QDA solution performed in \mathcal{B}' , where a regularized covariance matrix is defined for each class according to the suggestions in Ref. [11]. By adjusting the regularization parameters in the estimation of the covariance matrix, a set of traditional or novel linear discriminant analysis methods could be obtained.

Although RD-QDA relaxes the identical covariance assumption, it still assumes, however, that each class is subject to a Gaussian distribution, which may not be the case in reality. To further relax the Gaussian assumption, in this paper, a kernel-based RD-QDA solution, denoted as KRQDA, is proposed by using the kernel machine technique [12–16]. Kernel machine technique is considered the most important tool in the design of nonlinear statistical learning techniques. The premise behind the kernel machine technique is to find a nonlinear map from the original sample space to a higher dimensional kernel feature space by using a nonlinear function. In the kernel feature space, pattern distribution is expected to be simplified so that better classification performance can be achieved by applying traditional linear or quadratic methodologies [15,14]. The nonlinear mapping could be performed implicitly by replacing dot products of the feature representations in the kernel space with a kernel function defined in the original sample space [15,14]. Implementing the RD-QDA method in the kernel feature space, KRQDA combines the strengths of the quadratic learning technique and the kernel machine technique. The proposed method provides a sophisticated solution to the complex pattern recognition problem by producing quadratic boundaries in the kernel space and at the same time tackles the SSS problem by incorporating regularization settings through RD-QDA. It is not difficult to see that RD-QDA is a special case of the KRQDA method when a linear kernel is considered. To this end, KRQDA can be viewed as a more general learning algorithm from which a set of linear or nonlinear discriminant analysis methods could be obtained by adjusting the kernel function as well as the regularization parameters. Extensive experiments have been performed and the results indicate that the proposed method outperforms many traditional kernel-based algorithms and the RD-QDA solution.

The rest of this paper is organized as follows. Section 2 briefly reviews the RD-QDA method. Following that, the detailed derivation of RD-QDA in the kernel space is described in Section 3. In Section 4, extensive experiments are conducted on the FERET face database and other pattern classification data sets to demonstrate the effectiveness of the proposed method followed by a conclusion drawn in Section 5.

2. Review of regularized direct quadratic discriminant analysis

In Ref. [6], a regularized direct quadratic discriminant analysis (RD-QDA) has been proposed to apply QDA solution under the SSS scenario. The RD-QDA method includes two main steps: dimensionality reduction and the application of the regularized QDA in the reduced feature subspace.

Let $\mathcal{Z} = \{\mathcal{Z}_i\}_{i=1}^C$ denote the training set, containing C classes with each class $\mathcal{Z}_i = \{\mathbf{z}_{ij}\}_{j=1}^{C_i}$, consisting of C_i samples $\mathbf{z}_{ij} \in \mathbb{R}^J$, where \mathbb{R}^J denotes the J -dimensional real space. Therefore, there are $N = \sum_{i=1}^C C_i$ examples available in the training set in total.

Let \mathbf{S}_b and \mathbf{S}_w be the between- and within-class scatter matrices, the low-dimensional complement space of the null space of \mathbf{S}_b , denoted as \mathcal{B}' , is first extracted. Let $\mathbf{V}_b = [\mathbf{v}_{b1}, \dots, \mathbf{v}_{bM}]$ be M eigenvectors of \mathbf{S}_b corresponding to M non-zero eigenvalues $\Lambda = [\lambda_{b1}, \dots, \lambda_{bM}]$, where $M = \min(C - 1, J)$. The \mathbf{S}_b subspace \mathcal{B}' is thus spanned by \mathbf{V}_b , which is further scaled by $\mathbf{U} = \mathbf{V}_b \Lambda_b^{-1/2}$ so that $\mathbf{U}^T \mathbf{S}_b \mathbf{U} = \mathcal{I}$, where $\Lambda_b = \text{diag}(\Lambda)$, $\text{diag}(\cdot)$ denotes the diagonalization operator and \mathcal{I} is the $(M \times M)$ identity matrix. In order to perform QDA in \mathcal{B}' , all training samples \mathbf{z}_{ij} are first projected into the subspace spanned by \mathbf{U} , obtaining the corresponding feature representations expressed as $\mathbf{y}_{ij} = \mathbf{U}^T \mathbf{z}_{ij}$, where \mathbf{y}_{ij} is the feature representation of \mathbf{z}_{ij} in the subspace \mathcal{B}' . The regularized sample covariance matrix of i th class, denoted by $\hat{\Sigma}_i(\alpha, \gamma)$, is thus defined as follows [11]:

$$\hat{\Sigma}_i(\alpha, \gamma) = (1 - \gamma) \hat{\Sigma}_i(\alpha) + \frac{\gamma}{M} \text{tr}[\hat{\Sigma}_i(\alpha)] \mathbf{I},$$

$$\hat{\Sigma}_i(\alpha) = \frac{1}{C_i(\alpha)} [(1 - \alpha) \mathbf{S}_i + \alpha \mathbf{S}], \quad (1)$$

where (α, γ) are two regularization parameters and M is the dimensionality of \mathcal{B}' . $C_i(\alpha) = (1 - \alpha)C_i + \alpha N$ and \mathbf{S}_i is the covariance matrix of i th class estimated in \mathcal{B}' , i.e. $\mathbf{S}_i = \sum_{j=1}^{C_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T$, $\mathbf{y}_{ij} = \mathbf{U}^T \mathbf{z}_{ij}$, $\bar{\mathbf{y}}_i = (1/C_i) \sum_{j=1}^{C_i} \mathbf{y}_{ij}$ and $\mathbf{S} = \sum_{i=1}^C \mathbf{S}_i$.

In the classification session, the input test \mathbf{p} is projected into the subspace \mathcal{B}' , obtaining the corresponding representation $\mathbf{q} = \mathbf{U}^T \mathbf{p}$. The class label of \mathbf{p} is then determined by comparing the Mahalanobis distance between the input \mathbf{q} and each class center $\bar{\mathbf{y}}_i$, $i = 1, \dots, C$, i.e., $ID(\mathbf{p}) = \arg \min_i d_i(\mathbf{q})$, where

$$d_i(\mathbf{q}) = (\mathbf{q} - \bar{\mathbf{y}}_i)^T \hat{\Sigma}_i^{-1}(\alpha, \gamma) (\mathbf{q} - \bar{\mathbf{y}}_i) + \ln |\hat{\Sigma}_i(\alpha, \gamma)| - 2 \ln \pi_i, \quad (2)$$

where $\pi_i = C_i/N$ is the prior probability of class i .

It was shown in Ref. [6] that a series of traditional linear discriminant analysis methods could be obtained by adjusting the parameters of α and γ . These include the direct nearest center method (D-NC) ($\alpha = 1$ and $\gamma = 1$); the direct weighted nearest center method (D-WNC) ($\alpha = 0$ and $\gamma = 1$); the direct quadratic discriminant analysis (D-QDA) ($\alpha = 0$ and $\gamma = 0$) and two direct linear discriminant analysis solutions by Yu and Yang (YD-LDA) [3] ($\alpha = 1$ and $\gamma = 0$) and by Lu et al. [5] (LD-LDA) ($\alpha = 1$ and $\gamma = (\text{tr}[\mathbf{S}/N] + M)/M$), respectively.

3. Kernel regularized quadratic discriminant analysis

RD-QDA relaxes the assumption made by LDA-based solutions that the covariance matrix of each class is identical. It still assumes, however, that each class is subject to a Gaussian distribution, which is obviously not the case in many real-world pattern classification applications. Motivated by the kernel machine technique, a kernel based RD-QDA solution, denoted as KRQDA, is proposed in this paper. Implementing the RD-QDA method in the kernel feature space, KRQDA further relaxes the Gaussian assumption by introducing the kernel machine technique and at the same time tackles the SSS problem through the regularization settings in RD-QDA.

3.1. Dimensionality reduction in the kernel space

Let $\Phi = [\phi(\mathbf{z}_{11}), \dots, \phi(\mathbf{z}_{CC})]$ be the corresponding feature representations of the training samples in the kernel space \mathbb{F}^F . Let \mathbf{K} be the $N \times N$ Gram matrix (kernel matrix), i.e., $\mathbf{K} = (K_{lh})_{l=1, \dots, C}^{h=1, \dots, C}$. K_{lh} is a $C_l \times C_h$ sub-matrix of \mathbf{K} composed of the samples from classes \mathcal{Z}_l and \mathcal{Z}_h , i.e., $K_{lh} = (k_{ij})_{i=1, \dots, C_l}^{j=1, \dots, C_h}$, where $k_{ij} = k(\mathbf{z}_{li}, \mathbf{z}_{hj})$ and $k(\cdot)$ denotes the kernel function defined in \mathbb{R}^J . Let $\tilde{\mathbf{S}}_b$ be the between-class scatter in \mathbb{F}^F , defined as

$$\tilde{\mathbf{S}}_b = \frac{1}{N} \sum_{i=1}^C C_i (\bar{\phi}_i - \bar{\phi})(\bar{\phi}_i - \bar{\phi})^T, \quad (3)$$

where $\bar{\phi}_i = (1/C_i) \sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij})$ is the mean of \mathcal{Z}_i in \mathbb{F}^F and $\bar{\phi} = (1/N) \sum_{i=1}^C \sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij})$ is the mean of training samples in \mathbb{F}^F .

Following the RD-QDA framework, the complement space of the null space of $\tilde{\mathbf{S}}_b$, denoted as $\tilde{\mathcal{B}}'$, is first extracted. $\tilde{\mathcal{B}}'$ is spanned by the first M ($M \leq C - 1$) eigenvectors of $\tilde{\mathbf{S}}_b$, i.e., $\tilde{\mathbf{V}}_b = [\tilde{\mathbf{v}}_{b1}, \dots, \tilde{\mathbf{v}}_{bM}]$, corresponding to the M largest eigenvalues. $\tilde{\mathbf{V}}_b$ can be obtained by solving the eigenvalue problem of $\tilde{\mathbf{S}}_b$, which can be rewritten as follows:

$$\begin{aligned} \tilde{\mathbf{S}}_b &= \sum_{i=1}^C \left(\sqrt{\frac{C_i}{N}} (\bar{\phi}_i - \bar{\phi}) \right) \left(\sqrt{\frac{C_i}{N}} (\bar{\phi}_i - \bar{\phi}) \right)^T \\ &= \sum_{i=1}^C \hat{\phi}_i \hat{\phi}_i^T = \Phi_b \Phi_b^T, \end{aligned} \quad (4)$$

where $\hat{\phi}_i = \sqrt{C_i/N} (\bar{\phi}_i - \bar{\phi})$ and $\Phi_b = [\hat{\phi}_1, \dots, \hat{\phi}_C]$. It can be observed that $\tilde{\mathbf{S}}_b$ is a matrix of size $F \times F$, where F denotes the dimensionality of the kernel space. Due to the high dimensionality of \mathbb{F}^F , a direct computation of the eigenvectors of $\tilde{\mathbf{S}}_b$ is impossible. Fortunately, the problem can be solved by an algebraic transform from the eigenvectors of a $C \times C$ matrix $\Phi_b^T \Phi_b$ [17]. Let $\tilde{\lambda}_{bi}$ and $\tilde{\mathbf{e}}_{bi}$ be i th eigenvalue and eigenvector of $\Phi_b^T \Phi_b$ sorted in a descending eigenvalue order, it can be seen that $(\Phi_b \Phi_b^T)(\Phi_b \tilde{\mathbf{e}}_{bi}) = \tilde{\lambda}_{bi} (\Phi_b \tilde{\mathbf{e}}_{bi})$. Therefore, it can be deduced that $(\Phi_b \tilde{\mathbf{e}}_{bi})$ is i th eigenvector of $\tilde{\mathbf{S}}_b = \Phi_b \Phi_b^T$.

As shown in Ref. [17], $\Phi_b^T \Phi_b$ can be expressed as

$$\begin{aligned} \Phi_b^T \Phi_b &= \frac{1}{N} \mathbf{B} \cdot \left(\mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{A}_{NC} - \frac{1}{N} (\mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{1}_{NC}) \right. \\ &\quad \left. - \frac{1}{N} (\mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{A}_{NC}) + \frac{1}{N^2} (\mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{1}_{NC}) \right) \cdot \mathbf{B}, \end{aligned} \quad (5)$$

where $\mathbf{B} = \text{diag}[\sqrt{C_1}, \dots, \sqrt{C_C}]$, $\mathbf{1}_{NC}$ is an $N \times C$ matrix with all elements equal to 1, $\mathbf{A}_{NC} = \text{diag}[\mathbf{a}_{C_1}, \dots, \mathbf{a}_{C_C}]$ is an $N \times C$ block diagonal matrix and \mathbf{a}_{C_i} is a $C_i \times 1$ vector with all elements equal to $1/C_i$. Let $\tilde{\mathbf{E}}_{bM} = [\tilde{\mathbf{e}}_{b1}, \dots, \tilde{\mathbf{e}}_{bM}]$ consist of M significant eigenvectors of $\Phi_b^T \Phi_b$ corresponding to the M largest eigenvalues $\tilde{\lambda}_{b1} > \dots > \tilde{\lambda}_{bM}$ and $\tilde{\mathbf{V}}_b = \Phi_b \tilde{\mathbf{E}}_{bM}$, it is not difficult to derive that $\tilde{\mathbf{V}}_b^T \tilde{\mathbf{S}}_b \tilde{\mathbf{V}}_b = \tilde{\Lambda}_b$, where $\tilde{\Lambda}_b = \text{diag}[\tilde{\lambda}_{b,1}^2, \dots, \tilde{\lambda}_{b,M}^2]$. Thus, the transformation matrix $\tilde{\mathbf{U}}$ such that $\tilde{\mathbf{U}}^T \tilde{\mathbf{S}}_b \tilde{\mathbf{U}} = \mathbf{I}$ can be computed as follows:

$$\tilde{\mathbf{U}} = \tilde{\mathbf{V}}_b \tilde{\Lambda}_b^{-1/2}, \quad \tilde{\mathbf{V}}_b = \Phi_b \tilde{\mathbf{E}}_{bM}. \quad (6)$$

After obtaining the low-dimensional feature subspace $\tilde{\mathcal{B}}'$ spanned by $\tilde{\mathbf{U}}$, all training samples are projected to $\tilde{\mathcal{B}}'$ to get the corresponding feature representations $\tilde{\mathbf{y}}_{ij}$, as follows:

$$\tilde{\mathbf{y}}_{ij} = \tilde{\mathbf{U}}^T \phi(\mathbf{z}_{ij}) = \tilde{\Lambda}_b^{-1/2} \tilde{\mathbf{E}}_{bM}^T \Phi_b^T \phi(\mathbf{z}_{ij}), \quad (7)$$

where $\Phi_b^T \phi(\mathbf{z}_{ij})$ can be expressed as [17]

$$\Phi_b^T \phi(\mathbf{z}_{ij}) = \frac{1}{\sqrt{N}} \mathbf{B} \left(\mathbf{A}_{NC} \cdot v(\phi(\mathbf{z}_{ij})) - \frac{1}{N} \mathbf{1}_{NC}^T \cdot v(\phi(\mathbf{z}_{ij})) \right), \quad (8)$$

where $v(\phi(\mathbf{z}_{ij})) = [\phi(\mathbf{z}_{11})\phi(\mathbf{z}_{ij}), \phi(\mathbf{z}_{12})\phi(\mathbf{z}_{ij}), \dots, \phi(\mathbf{z}_{CC})\phi(\mathbf{z}_{ij})]^T$ can be computed implicitly through the kernel function defined in \mathbb{R}^J , i.e., $\phi(\mathbf{z}_{mn})\phi(\mathbf{z}_{ij}) = k(\mathbf{z}_{mn}, \mathbf{z}_{ij})$.

3.2. Perform QDA in $\tilde{\mathcal{B}}'$

In order to perform QDA solution in $\tilde{\mathcal{B}}'$, the regularized sample covariance matrix of i th class, denoted as $\tilde{\Sigma}_i(\alpha, \gamma)$, is defined as follows [11]:

$$\tilde{\Sigma}_i(\alpha, \gamma) = (1 - \gamma) \tilde{\Sigma}_i(\alpha) + \frac{\gamma}{M} \text{tr}[\tilde{\Sigma}_i(\alpha)] \mathbf{I},$$

$$\tilde{\Sigma}_i(\alpha) = \frac{1}{C_i(\alpha)} [(1 - \alpha) \tilde{\mathbf{S}}_i + \alpha \tilde{\mathbf{S}}],$$

$$C_i(\alpha) = (1 - \alpha) C_i + \alpha N,$$

$$\tilde{\mathbf{S}}_i = \sum_{j=1}^{C_i} (\tilde{\mathbf{y}}_{ij} - \bar{\tilde{\mathbf{y}}}_i)(\tilde{\mathbf{y}}_{ij} - \bar{\tilde{\mathbf{y}}}_i)^T,$$

$$\tilde{\mathbf{S}} = \sum_{i=1}^C \tilde{\mathbf{S}}_i. \quad (9)$$

$\bar{\tilde{\mathbf{y}}}_i = (1/C_i) \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij}$ and (α, γ) is a pair of regularization parameters.

Table 1
Kernel-based algorithms derived from KRQDA

Algorithms	KNC	KWNC	KQDA	KDDA ¹	KDDA ²
α	1	0	0	1	1
γ	1	1	0	0	$\frac{M}{\text{tr}\left(\frac{\tilde{\mathbf{S}}_i}{N}\right) + M}$
$\tilde{\Sigma}_i$	$\frac{1}{M}\text{tr}[\tilde{\mathbf{S}}_i/N]\mathcal{I}$	$\frac{1}{M}\text{tr}[\tilde{\mathbf{S}}_i/N]\mathcal{I}$	$\tilde{\mathbf{S}}_i/C_i$	$\tilde{\mathbf{S}}_i/N$	$\left(\frac{\text{tr}[\tilde{\mathbf{S}}_i/N]}{\text{tr}[\tilde{\mathbf{S}}_i/N] + M}\right)(\tilde{\mathbf{S}}_i/N + \mathcal{I})$

The key component in the calculation of $\tilde{\Sigma}_i(\alpha, \gamma)$ is to derive the covariance matrix of i th class, i.e., $\tilde{\mathbf{S}}_i$, which could be expressed as follows:

$$\begin{aligned} \tilde{\mathbf{S}}_i &= \sum_{j=1}^{C_i} (\tilde{\mathbf{y}}_{ij} - \tilde{\mathbf{y}}_i)(\tilde{\mathbf{y}}_{ij} - \tilde{\mathbf{y}}_i)^T \\ &= \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij}\tilde{\mathbf{y}}_{ij}^T - \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_{ij}^T - \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij}\tilde{\mathbf{y}}_i^T + \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T \\ &= \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij}\tilde{\mathbf{y}}_{ij}^T - C_i\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T - C_i\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T + C_i\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T \\ &= \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij}\tilde{\mathbf{y}}_{ij}^T - C_i\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T \\ &= \mathbf{J}_1 - C_i \times \mathbf{J}_2, \end{aligned} \tag{10}$$

where $\mathbf{J}_1 = \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij}\tilde{\mathbf{y}}_{ij}^T$ and $\mathbf{J}_2 = \tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^T$. The detailed derivation of \mathbf{J}_1 and \mathbf{J}_2 can be found in Appendices A and B.

In the classification session, the input test image \mathbf{p} is firstly projected to the lower dimensional between class subspace i.e., $\tilde{\mathbf{q}} = \tilde{\mathbf{U}}^T \phi(\mathbf{p})$. The identity of the test image is then determined by comparing the Mahalanobis distance between the feature representation of the test image $\tilde{\mathbf{q}}$ and each class center $\tilde{\mathbf{y}}_i$, i.e.,

$$\begin{aligned} ID(\mathbf{p}) &= \arg \min_i d_i(\tilde{\mathbf{q}}), \\ d_i(\tilde{\mathbf{q}}) &= (\tilde{\mathbf{q}} - \tilde{\mathbf{y}}_i)^T \tilde{\Sigma}_i^{-1}(\alpha, \gamma) (\tilde{\mathbf{q}} - \tilde{\mathbf{y}}_i) + \ln|\tilde{\Sigma}_i(\alpha, \gamma)| - 2\ln\pi_i, \end{aligned} \tag{11}$$

where $\pi_i = C_i/N$.

It was shown in Ref. [6] that the two parameters α ($0 \leq \alpha \leq 1$) and γ ($0 \leq \gamma \leq 1$) regularize the estimation of $\tilde{\mathbf{S}}_i$ to $\tilde{\mathbf{S}}$ (by α) and to a multiple of the identity matrix (by γ). Therefore, in a way similar to RD-QDA, a set of traditional or novel kernel-based discriminant learning methods can be derived by adjusting the parameters of α and γ which are summarized in Table 1. When ($\alpha = 0, \gamma = 0$), no regularization is applied. The proposed KRQDA is equivalent to a standard quadratic discriminant analysis solution in the kernel feature space. Thus, a novel kernel-based QDA method is derived, denoted as KQDA. While ($\alpha = 1, \gamma = 1$), a strong regularization is applied so that the covariance matrix $\tilde{\Sigma}_i$ becomes a multiple of the identity matrix, i.e., $\tilde{\Sigma}_i = \text{tr}(\tilde{\mathbf{S}}_i/N)\mathbf{I}/M$. Thus the proposed method is

reduced to a nearest center solution performed in the kernel space, denoted as kernel nearest center classifier KNC. When ($\alpha = 0, \gamma = 1$), $\tilde{\Sigma}_i = \text{tr}(\tilde{\mathbf{S}}_i/N)\mathbf{I}/M$. In such a case, the proposed method becomes a weighted nearest center classifier, denoted as KWNC, with the weight for each class defined as $\text{tr}(\tilde{\mathbf{S}}_i/N)/M$. If ($\alpha = 1, \gamma = 0$), it is not difficult to see that the proposed method is equivalent to the well-known kernel direct discriminant analysis [17] with the traditional Fisher's criterion ($\tilde{\mathbf{A}} = \arg \max_{\tilde{\mathbf{A}}} |\tilde{\mathbf{A}}^T \tilde{\mathbf{S}}_b \tilde{\mathbf{A}}| / |\tilde{\mathbf{A}}^T \tilde{\mathbf{S}}_w \tilde{\mathbf{A}}|$), denoted as KDDA¹. If a modified Fisher's criterion is considered as proposed in Ref. [17], i.e., ($\tilde{\mathbf{A}} = \arg \max_{\tilde{\mathbf{A}}} |\tilde{\mathbf{A}}^T \tilde{\mathbf{S}}_b \tilde{\mathbf{A}}| / (|\tilde{\mathbf{A}}^T \tilde{\mathbf{S}}_b \tilde{\mathbf{A}}| + |\tilde{\mathbf{A}}^T \tilde{\mathbf{S}}_w \tilde{\mathbf{A}}|)$), it can be derived, that the corresponding KDDA solution, denoted as KDDA², is also a special case of KRQDA, when ($\alpha = 1, \gamma = (\text{tr}(\tilde{\mathbf{S}}_i/N) + M)/M$).

In addition to manipulating the regularization parameters (α, γ), another flexibility of the proposed method lies in the choosing of kernel functions $k(\cdot)$. The variation in the kernel functions results in different kernel spaces. It can be observed that when a linear kernel is used, i.e., $k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \mathbf{z}_j$, the proposed KRQDA is reduced to RD-QDA. Therefore, KRQDA is a more general solution and RD-QDA is a special case corresponding to the linear kernel functions.

4. Experiment

In order to demonstrate the effectiveness of the proposed method, a set of experiments are conducted to examine the performance of the proposed KRQDA method with respect to the regularization parameters (α, γ) as well as the kernel functions.

4.1. Experiment I—performance as functions of regularization parameters

In order to examine how the regularization parameters (α, γ) affect the performance of KRQDA under different SSS scenarios, a subset of the well-known FERET face database is used for experimentation [18,19]. The data set used in this experiment contains 960 images. It is composed of 91 subjects (i.e., 91 classes) with each subject having at least eight images. According to the FERET protocol guidelines, for each face image, the following preprocessing operations are performed: (1) images are rotated and scaled so that the centers of the eyes are placed on specific pixels and the image size is normalized to 150×130 ; (2) a standard mask is applied to remove

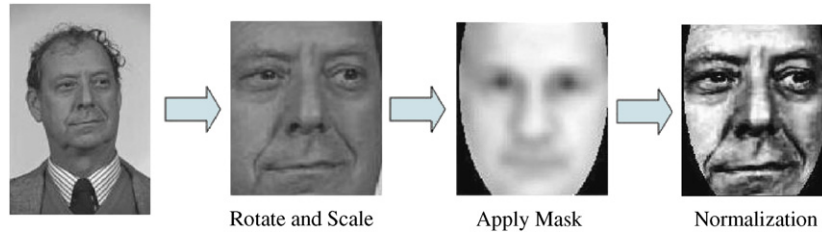
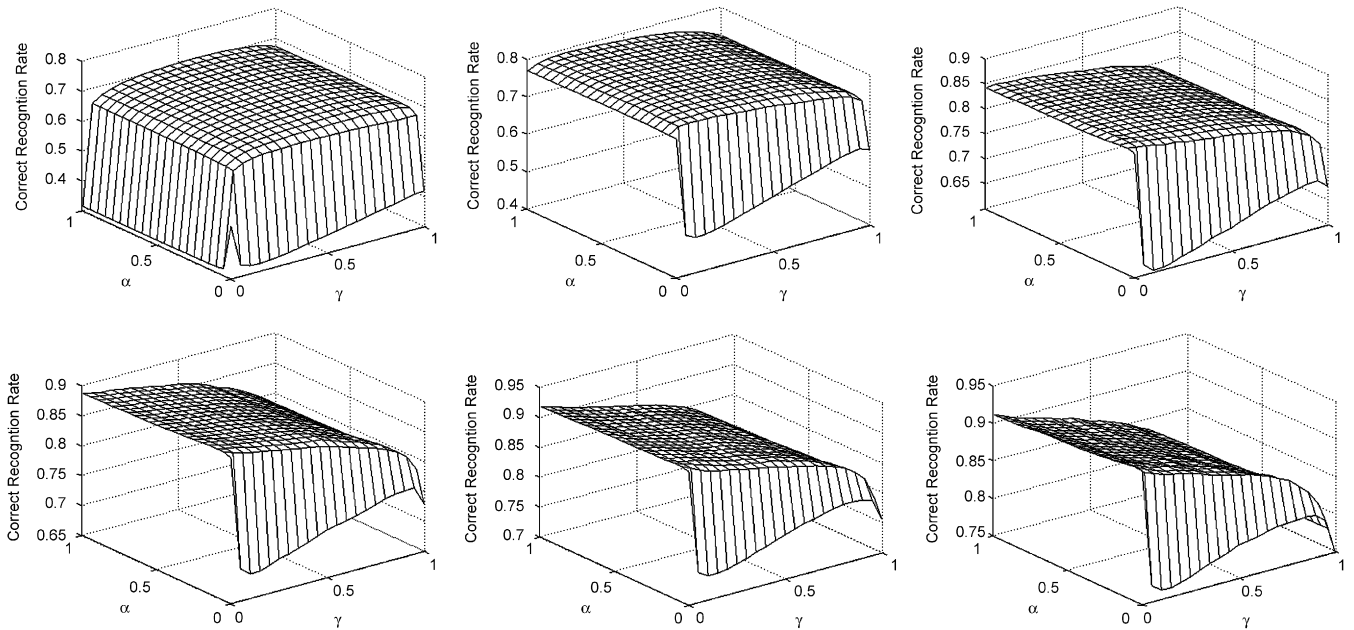


Fig. 1. Preprocessing procedure.

Fig. 2. *CRR* of *KRQDA* as functions of (α, γ) ; top: $L = 2, 3, 4$; bottom $L = 5, 6, 7$.

non-face portions; (3) histogram equalization is performed and image intensity values are normalized to zero mean and unit standard deviation; (4) each image is finally represented, after the application of the mask, as a vector of dimensionality 17 154. Fig. 1 depicts the preprocessing procedure.

For each subject, L images are extracted to form the training set while the remaining images are treated as tests. In order to examine the effectiveness of the proposed method in handling the SSS problem, a set of L values are tested with $L = \{2, 3, 4, 5, 6, 7\}$. For each L value, the partitioning of training and test sets is repeated 10 times and the reported results are the average of 10 repetitions.

The testing grid of (α, γ) values for the proposed *KRQDA* solution is defined as $\alpha = [10^{-4} : 0.0495 : 1]$ and $\gamma = [10^{-4} : 0.0495 : 1]$, where $[10^{-4} : 0.0495 : 1]$ denotes a vector consisting 21 elements from 10^{-4} to 1 with a step of 0.0495. Therefore, 21×21 (α, γ) combinations are examined. A Gaussian kernel is used as the kernel function for all kernel-based algorithms. The kernel parameter σ^2 is set to 10^8 based on a “trial-and-error” method.

The recognition performance is measured by the correct recognition rate (*CRR*). Fig. 2 depicts the *CRRs* with respect to different regularization parameters (α, γ) . Also, a quantitative

comparison of various kernel-based algorithms is performed. Table 2 summarizes the *CRRs* obtained by various kernel-based algorithms. The *KRQDA* performance reported here represents best found *CRRs* across all (α, γ) combinations. The optimal (α^*, γ^*) values are also listed. In addition to *KRQDA* and its derived algorithms (listed in Table 1), kernel principal component analysis (*KPCA*) [20] and generalized discriminant analysis (*GDA*) [21] are also implemented for comparison purposes. For *KPCA* and *GDA*, it is well known that the corresponding performance is a function of the number of extracted features. Therefore, the performance reported here is the best found *CRR* corresponding the optimal feature number (N_{opt}) obtained by exhaustively searching all possible numbers. It can be observed from Fig. 2, when $L = 2$, the peak value of the *CRR* appears in the central area of the (α, γ) grid. As the value of L increases, the *CRR* peak moves toward the $(0, 0)$ corner. A similar observation can be obtained from Table 2. In a severe SSS scenario, i.e., when $L = 2$, the best *CRR* of *KRQDA* is obtained when α and γ have large values ($\lambda = 0.64$ and $\gamma = 0.54$). Correspondingly, solutions like *KNC* ($\alpha = 1, \gamma = 1$) and *KDDA*² ($\alpha = 1, \gamma = M/(\text{tr}(\tilde{\mathbf{S}}_i/N) + M)$) which include a strong regularization process outperform less regularized solutions such as *KWNC* ($\alpha = 0, \gamma = 1$), *KDDA*¹ ($\alpha = 1, \gamma = 0$) and *KQDA*

Table 2
CRRs (%) of various kernel based algorithms on FERET set

	$L = 2$	$L = 3$	$L = 4$	$L = 5$	$L = 6$	$L = 7$
KPCA	59.10	65.59	67.13	68.18	71.06	71.61
N_{opt}	159.3	225.2	261.1	247.9	335	355.7
GDA	58.52	68.95	77.47	86.75	89.71	89.32
N_{opt}	89.7	90	89.9	89.9	89.7	84.5
KNC	67.42	72.55	75.30	77.74	78.24	77.68
KWNC	41.95	60.26	67.80	72.97	75.70	74.89
KQDA	47.52	77.37	84.77	89.37	92.42	92.79
$KDDA^1$	24.31	77.05	84.36	88.61	91.74	91.18
$KDDA^2$	69.92	77.02	81.73	84.99	87.17	86.69
KRQDA	70.85	79.61	85.25	89.72	92.58	92.97
(α^*, γ^*)	(0.64, 0.54)	(0.45, 0.25)	(0.05, 10^{-4})	(0.10, 10^{-4})	(0.10, 10^{-4})	(0.10, 10^{-4})

($\alpha = 0, \gamma = 0$). If sufficient samples are available, e.g., $L = 7$, KQDA gives better performance. This indicates when the training sample size (L) is small, the SSS problem is a more critical concern. In such cases, a regularization step that replaces individual covariance matrices $\tilde{\Sigma}_i$, with a common within-class scatter matrix \tilde{S} estimated from all training samples is an appropriate choice. Although it is a biased solution, it significantly reduces the variance in the estimation of $\tilde{\Sigma}_i$ resulting in an improved performance. However, when sufficient samples are available for each class, methods like KQDA demonstrate their effectiveness in tackling the complex distribution problem by producing quadratic boundaries in the kernel feature space. Therefore, in such cases, unbiased solutions associated with small (α, γ) values are appropriate choices. Compared to other kernel-based algorithms, KRQDA gives the best performance for all cases due to the fact that a flexible regularization scheme has been integrated.

4.2. Experiment II—performance as functions of kernels

In addition to regularization parameters, another flexibility in the proposed KRQDA method is the manipulation of the kernels. In this experiment, the influence of the selected kernel functions will be examined. In addition to the FERET face database, two additional data sets are used for this experiment including the vowel data set and the vehicle data set collected from UCI benchmark repository [22]. The vowel data set consists of 11 classes with each class containing 90 samples. Each sample is represented by a 10-dimensional vector. The vehicle data set consists of four classes with each class containing over 200 samples. Each sample is represented by an 18-dimensional vector.

For face data set, $L = \{2, 4, 6\}$ images per subject are extracted as training samples while the rest are treated as tests. Again, the configuration of the training and test sets is repeated 10 times and the reported results are the average over 10 trials. For the vowel and vehicle data sets, a 10-fold testing is performed. In other words, for each class, the samples are separated into 10 non-overlapping subsets, nine of which are used as training samples while the remaining single subset is treated as the test set. The experiment is repeated 10 times

Table 3
Best found CRR (%) of KRQDA with Gaussian kernel using various σ^2 values

$\sigma^2 =$	10^4	10^5	10^6	10^7	10^8
$L = 2$	65.35	71.00	71.09	70.89	70.85
$L = 4$	77.33	84.41	85.13	85.23	85.25
$L = 6$	84.08	91.59	92.58	92.56	92.58
$\sigma^2 =$	10	50	100	500	1000
Vowel	89.49	90.91	90.81	90.00	89.39
Vehicle	25.34	43.38	56.76	55.81	55.07

corresponding to 10 different test sets and the results reported here are the average over 10 trials.

The testing grid of (α, γ) values for vowel and vehicle data set is set to $\alpha = [0 : 0.0495 : 1]$ and $\gamma = [0 : 0.0495 : 1]$ so that the extreme cases $\alpha = 0$ or $\gamma = 0$ will be tested. However, for face data set, the corresponding test grid remains the same as that used in the first experiment, i.e., $\alpha = [10^{-4} : 0.0495 : 1]$ and $\gamma = [10^{-4} : 0.0495 : 1]$. This is due to the fact that for face data set, ($\alpha = 0, \gamma = 0$) will trigger computation difficulty. If ($\alpha = 0, \gamma = 0$), according to Eq. (9), it can be derived that $\tilde{\Sigma}_i(\alpha, \gamma) = \tilde{S}_i / C_i$, which is singular. This is due to the fact that the number of training samples for each subject is smaller than the dimensionality of the reduced between-class subspace, i.e., $C - 1 = 90$, where C is the number of subjects. Then in the recognition step, the computation of Mahalanobis distance d_i (Eq. (11)) becomes inapplicable.

Table 3 depicts the best found CRRs of KRQDA with respect to different σ^2 values. It can be seen that different kernel σ^2 values results in different recognition performance. For the vowel data set, performance difference due to the different σ^2 values is small. However, for other two data sets, such difference is large. It can be observed that the gap between the best and worst CRRs are over 5% and 30% for the FERET and vehicle data sets, respectively. This indicates that choosing an appropriate kernel parameter is paramount in ensuring good performance in kernel-based algorithms. Table 4 lists the best found CRRs of KRQDA with respect to different kernels, including the Gaussian kernel, polynomial kernel and linear kernel. For the Gaussian kernel, σ^2 is set to 10^8 , 50 and 100 for the face, vowels and vehicle sets. While for the polynomial

Table 4
Best found CRR (%) of KRQDA with various kernel functions

	FERET			Vowel	Vehicle
	$L = 2$	$L = 4$	$L = 6$		
Gaussian (α^* , γ^*)	70.85 (0.64, 0.54)	85.25 (0.05, 10^{-4})	92.58 (0.10, 10^{-4})	90.91 (0, 0)	56.76 (0.6435, 0.9900)
Polynomial (α^* , γ^*)	68.84 (0.40, 0.64)	84.87 (0.10, 10^{-4})	92.61 (0.10, 10^{-4})	87.78 (0, 0)	51.86 (0.4950, 0)
Linear (α^* , γ^*)	70.86 (0.74, 0.64)	85.22 (0.05, 10^{-4})	92.63 (0.05, 10^{-4})	88.19 (0, 0)	52.66 (0.1980, 0)

kernel, defined as $k(\mathbf{z}_i, \mathbf{z}_j) = (a(\mathbf{z}_i \cdot \mathbf{z}_j) + b)^d$, the parameters are set to $\{a = 10^{-9}, b = 1, d = 3\}$ for the FERET face data set as suggested in Ref. [17], $\{a = 1, b = 1, d = 3\}$ for the vowel data set and $\{a = 1, b = 1, d = 0.01\}$ for the vehicle data set. The kernel parameters are selected empirically through a “trial-and-error” procedure. The linear kernel is defined as $k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \mathbf{z}_j$. It is not difficult to understand, that if a linear kernel is considered, the proposed KRQDA is reduced to RD-QDA.

Compared to the RD-QDA solution, no significant improvement can be observed by the proposed KRQDA for the FERET face data set. The performance of KRQDA with either the Gaussian kernel or the polynomial kernel is very close to that of RD-QDA. This can be explained as follows. The superiority of the proposed KRQDA method to the traditional RD-QDA method lies in its capability of relaxing the Gaussian assumption by using kernel machine technique. However, as a more complicated algorithm, it is more easily to overfit in particular for face recognition application where only limited number of training samples are available. In addition, as any kernel-based algorithm, its performance depends on the selected kernel function and its associated parameters. However, how to systematically choose an appropriate kernel function is still a challenging problem far from well solved. The kernel parameters used in this experiment is selected empirically through a “trial-and-error” procedure. The lack of a good kernel selection algorithm also prevents the KRQDA method to demonstrate its potential power. Therefore, its superiority has been discounted. For the vowel and vehicle data sets, where the SSS problem is not that critical with more training samples available, better performance can be observed by using the KRQDA method with Gaussian kernel. In order to examine how KRQDA improves the recognition performance, we define $IMP = CRR_{Gaussian} - CRR_{Linear} / 1 - CRR_{Linear}$ (%) to measure to comparative improvement of the KRQDA method with Gaussian kernel with respect to that with linear kernel, i.e., RD-QDA. It can be observed that $IMP_{Vowel} = 23\%$ and $IMP_{Vehicle} = 9\%$. This indicates that the KRQDA method provides better performance than the traditional RD-QDA method, in particular on the vowel data set.

In comparing with computational cost, KRQDA is more time consuming than RD-QDA as it is a more complicated algorithm. We compare both the training and testing time consumed by the KRQDA method and the RD-QDA method on

the FERET face data set, when $L = 2$. The simulation is implemented on a PC with a 3.0 GHz Intel Pentium 4 processor and 2.0 GB RAM. All programs are written in Matlab v7.0 and executed in MS Windows XP. It takes 15.23 s to train a KRQDA algorithm and 3.79 s to train a RD-QDA algorithm. In terms of testing time, KRQDA needs 0.046 s and RD-QDA needs 0.031 s. It can be observed that KRQDA takes much more time than RD-QDA in the training session. The extra time is mostly spent on the calculation of the kernel matrix. Since the training is an off-line operation, a higher computational cost is durable. While looking at the testing time, however, KRQDA is comparable to RD-QDA. This indicates that KRQDA is a realistic solution which can be applied in real-world applications.

5. Conclusion

In this paper, a novel kernel regularized quadratic learning algorithm, KRQDA, has been developed. The proposed method combines the strengths of the kernel machine technique and the regularized quadratic learning technique to tackle the complex pattern classification problem under the SSS scenario. KRQDA is a comprehensive and general pattern recognition solution from which a set of traditional or novel, linear or nonlinear learning algorithms can be derived by adjusting the regularization parameters as well as the kernel functions. Experimentation on different data sets indicates that the proposed method outperforms many traditional discriminant learning algorithms.

As many other kernel-based algorithms, different kernel functions as well as different kernel parameters significantly affect the performance of the KRQDA method. However, how to choose an appropriate kernel function still remains a difficult problem in the kernel machine research. In addition, systematically choosing appropriate regularization parameters α and γ is another challenging task due to the lack of priori knowledge regarding the underlying distribution of the patterns. These problems will be left for future research.

Acknowledgments

This work is partially supported by a grant provided by the Bell University Laboratory at the University of Toronto.

Portions of the research in this paper use the FERET database of facial images collected under the FERET program. The authors would like to thank the FERET Technical Agent, the U.S. National Institute of Standards and Technology (NIST) for providing the FERET database.

Appendix A. Derivation of \mathbf{J}_1

$$\begin{aligned} \mathbf{J}_1 &= \sum_{j=1}^{C_i} \tilde{\mathbf{y}}_{ij} \tilde{\mathbf{y}}_{ij}^T = \sum_{j=1}^{C_i} \tilde{\mathbf{U}}^T \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \tilde{\mathbf{U}} \\ &= (\tilde{\mathbf{E}}_{bm} \tilde{\Lambda}_b^{-1/2})^T \Phi_b^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \\ &\quad \times \Phi_b (\tilde{\mathbf{E}}_{bm} \tilde{\Lambda}_b^{-1/2}), \end{aligned} \quad (12)$$

where $\tilde{\mathbf{E}}_{bm}$ and $\tilde{\Lambda}_b$ can be obtained from Eq. (5).

Express $\Phi_b^T (\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T) \Phi_b$ as follows:

$$\begin{aligned} &\Phi_b^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \Phi_b \\ &= \left(\hat{\Phi}_m^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \hat{\Phi}_n \right)_{n=1, \dots, C}^{m=1, \dots, C}, \end{aligned} \quad (13)$$

where

$$\begin{aligned} &\hat{\Phi}_m^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \hat{\Phi}_n \\ &= \frac{\sqrt{C_m C_n}}{N} \left[\bar{\Phi}_m^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\Phi}_n \right. \\ &\quad - \bar{\Phi}_m^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\Phi} \\ &\quad - \bar{\Phi}^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\Phi}_n \\ &\quad \left. + \bar{\Phi}^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\Phi} \right] \\ &= \frac{\sqrt{C_m C_n}}{N} (I_1 - I_2 - I_3 + I_4), \end{aligned} \quad (14)$$

where I_1, I_2, I_3, I_4 can be expanded as follows:

$$\begin{aligned} I_1 &= \bar{\Phi}_m^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\Phi}_n \\ &= \frac{1}{C_m} \sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \frac{1}{C_n} \sum_{y=1}^{C_n} \phi(\mathbf{z}_{ny}) \\ &= \frac{1}{C_m C_n} \sum_{j=1}^{C_i} \left[\sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \phi(\mathbf{z}_{ij}) \right] \left[\sum_{y=1}^{C_n} \phi(\mathbf{z}_{ij})^T \phi(\mathbf{z}_{ny}) \right] \\ &= \frac{1}{C_m C_n} \sum_{j=1}^{C_i} \left[\sum_{x=1}^{C_m} (k_{xj})_{mi} \sum_{y=1}^{C_n} (k_{jy})_{in} \right] \\ &= \sum_{j=1}^{C_i} (\mathbf{Q}_{N1}^m)^T \mathbf{K} \mathbf{Q}_{N1}^{i,j} (\mathbf{Q}_{N1}^{i,j})^T \mathbf{K} \mathbf{Q}_{N1}^n \\ &= (\mathbf{Q}_{N1}^m)^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{Q}_{N1}^n. \end{aligned}$$

Therefore,

$$(I_1)_{n=1, \dots, C}^{m=1, \dots, C} = \mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC},$$

$$\begin{aligned} I_2 &= \bar{\Phi}_m^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\Phi} \\ &= \frac{1}{C_m} \sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \frac{1}{N} \sum_{n=1}^C \sum_{y=1}^{C_n} \phi(\mathbf{z}_{ny}) \\ &= \frac{1}{C_m N} \sum_{j=1}^{C_i} \left\{ \left[\sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \phi(\mathbf{z}_{ij}) \right] \right. \\ &\quad \left. \times \left[\sum_{n=1}^C \sum_{y=1}^{C_n} \phi(\mathbf{z}_{ij})^T \phi(\mathbf{z}_{ny}) \right] \right\} \\ &= \frac{1}{C_m N} \sum_{j=1}^{C_i} \left[\sum_{x=1}^{C_m} (k_{xj})_{mi} \sum_{n=1}^C \sum_{y=1}^{C_n} (k_{jy})_{in} \right] \\ &= \frac{1}{N} \sum_{j=1}^{C_i} (\mathbf{Q}_{N1})^T \mathbf{K} \mathbf{Q}_{N1}^{i,j} (\mathbf{Q}_{N1}^{i,j})^T \mathbf{K} \mathbf{1}_{N1} \\ &= \frac{1}{N} (\mathbf{Q}_{N1}^m)^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{N1} \\ &\Rightarrow (I_2)_{n=1, \dots, C}^{m=1, \dots, C} = \frac{1}{N} \mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC}, \\ (I_3)_{n=1, \dots, C}^{m=1, \dots, C} &= (I_2)_{n=1, \dots, C}^{m=1, \dots, C} = \frac{1}{N} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC}, \end{aligned}$$

$$\begin{aligned}
 I_4 &= \bar{\phi}^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \bar{\phi} \\
 &= \frac{1}{N} \sum_{m=1}^C \sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \frac{1}{N} \sum_{n=1}^C \sum_{y=1}^{C_n} \phi(\mathbf{z}_{ny}) \\
 &= \frac{1}{N^2} \sum_{j=1}^{C_i} \left\{ \left[\sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \phi(\mathbf{z}_{ij}) \right] \right. \\
 &\quad \left. \times \left[\sum_{n=1}^C \sum_{y=1}^{C_n} \phi(\mathbf{z}_{ij})^T \phi(\mathbf{z}_{ny}) \right] \right\} \\
 &= \frac{1}{N^2} \sum_{j=1}^{C_i} \left[\sum_{x=1}^{C_m} (k_{xj})_{mi} \sum_{n=1}^C \sum_{y=1}^{C_n} (k_{jy})_{in} \right] \\
 &= \frac{1}{N^2} \sum_{j=1}^{C_i} (\mathbf{1}_{N1})^T \mathbf{K} \mathbf{Q}_{N1}^{i,j} (\mathbf{Q}_{N1}^{i,j})^T \mathbf{K} \mathbf{1}_{N1} \\
 &= \frac{1}{N^2} (\mathbf{1}_{N1})^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{N1}
 \end{aligned}$$

$$\Rightarrow (I_4)_{n=1, \dots, C}^{m=1, \dots, C} = \frac{1}{N^2} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC}.$$

Define $\mathbf{Q}_{N1} = [\mathbf{q}_{C1}, \dots, \mathbf{q}_{CC}]$ is a $N \times 1$ vector and \mathbf{q}_{C_i} is a $C_i \times 1$ vector with all elements equal to 0. Thus, \mathbf{Q}_{N1}^m is equivalent to \mathbf{Q}_{N1} except that the elements in \mathbf{q}_{C_m} equal to $1/C_m$ while $\mathbf{Q}_{N1}^{i,j}$ is equivalent to \mathbf{Q}_{N1} except that the j th element in \mathbf{q}_{C_i} equals to 1. $\mathbf{D}_{NN}^i = \text{diag}[\mathbf{0}_{C_1}, \dots, \mathbf{0}_{C_{i-1}}, \mathbf{1}_{C_i}, \mathbf{0}_{C_{i+1}}, \dots, \mathbf{0}_{C_C}]$ is a $N \times N$ diagonal matrix, $\mathbf{1}_{C_i}$ is a $C_i \times 1$ vector with all elements equal to 1 and $\mathbf{0}_{C_k}$ is a $C_k \times 1$ vector with all elements equal to 0.

Therefore

$$\begin{aligned}
 &\Phi_b^T \left(\sum_{j=1}^{C_i} \phi(\mathbf{z}_{ij}) \phi(\mathbf{z}_{ij})^T \right) \Phi_b \\
 &= \frac{1}{N} \mathbf{B} \cdot (\mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC} \\
 &\quad - \frac{1}{N} \mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC} \\
 &\quad - \frac{1}{N} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC} \\
 &\quad + \frac{1}{N^2} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{D}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC}) \cdot \mathbf{B}^T. \tag{15}
 \end{aligned}$$

Appendix B. Derivation of J_2

$$\begin{aligned}
 J_2 &= \bar{\mathbf{y}}_i \bar{\mathbf{y}}_i^T = \tilde{\mathbf{U}}^T \bar{\phi}_i \bar{\phi}_i^T \tilde{\mathbf{U}} \\
 &= (\tilde{\mathbf{E}}_{bm} \mathbf{A}_b^{-1/2})^T \Phi_b^T \bar{\phi}_i \bar{\phi}_i^T \Phi_b (\tilde{\mathbf{E}}_{bm} \mathbf{A}_b^{-1/2}), \tag{16}
 \end{aligned}$$

where $\Phi_b^T \bar{\phi}_i \bar{\phi}_i^T \Phi_b$ can be expressed as follows:

$$\Phi_b^T \bar{\phi}_i \bar{\phi}_i^T \Phi_b = (\hat{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \hat{\phi}_n)_{n=1, \dots, C}^{m=1, \dots, C}, \tag{17}$$

where

$$\begin{aligned}
 \hat{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \hat{\phi}_n &= \frac{\sqrt{C_m C_n}}{N} (\bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}_n - \bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi} \\
 &\quad - \bar{\phi}^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}_n + \bar{\phi}^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}). \tag{18}
 \end{aligned}$$

Express each item in Eq. (18) as follows:

$$\begin{aligned}
 &\bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}_n \\
 &= \left(\frac{1}{C_m} \sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \right) \left(\frac{1}{C_i} \sum_{y=1}^{C_i} \phi(\mathbf{z}_{iy}) \right) \\
 &\quad \times \left(\frac{1}{C_i} \sum_{s=1}^{C_i} \phi(\mathbf{z}_{is})^T \right) \left(\frac{1}{C_n} \sum_{t=1}^{C_n} \phi(\mathbf{z}_{tn}) \right) \\
 &= \frac{1}{C_m C_i} \sum_{x=1}^{C_m} \sum_{y=1}^{C_i} (k_{xy})_{mi} \cdot \frac{1}{C_i C_n} \sum_{s=1}^{C_i} \sum_{t=1}^{C_n} (k_{st})_{in} \\
 &= (\mathbf{Q}_{N1}^m)^T \cdot \mathbf{K} \cdot (\mathbf{Q}_{N1}^i) \cdot (\mathbf{Q}_{N1}^i)^T \cdot \mathbf{K} \cdot (\mathbf{Q}_{N1}^n) \\
 &\Rightarrow (\bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}_n)_{n=1, \dots, C}^{m=1, \dots, C} = \mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC},
 \end{aligned}$$

$$\begin{aligned}
 &\bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi} \\
 &= \left(\frac{1}{C_m} \sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \right) \left(\frac{1}{C_i} \sum_{y=1}^{C_i} \phi(\mathbf{z}_{iy}) \right) \\
 &\quad \times \left(\frac{1}{C_i} \sum_{s=1}^{C_i} \phi(\mathbf{z}_{is})^T \right) \left(\frac{1}{N} \sum_{n=1}^C \sum_{t=1}^{C_n} \phi(\mathbf{z}_{tn}) \right) \\
 &= \frac{1}{C_m C_i} \sum_{x=1}^{C_m} \sum_{y=1}^{C_i} (k_{xy})_{mi} \cdot \frac{1}{C_i N} \sum_{s=1}^{C_i} \sum_{n=1}^C \sum_{t=1}^{C_n} (k_{st})_{in} \\
 &= (\mathbf{Q}_{N1}^m)^T \cdot \mathbf{K} \cdot (\mathbf{Q}_{N1}^i) \cdot (\mathbf{Q}_{N1}^i)^T \cdot \mathbf{K} \cdot \left(\frac{1}{N} \mathbf{1}_{N1} \right) \\
 &\Rightarrow (\bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi})_{n=1, \dots, C}^{m=1, \dots, C} = \frac{1}{N} \mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC},
 \end{aligned}$$

$$(\bar{\phi} \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}_n)_{n=1, \dots, C}^{m=1, \dots, C} = ((\bar{\phi}_m^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi})^T)_{n=1, \dots, C}^{m=1, \dots, C}$$

$$= \frac{1}{N} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC},$$

$$\begin{aligned}
& \bar{\phi}^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi} \\
&= \left(\frac{1}{N} \sum_{m=1}^C \sum_{x=1}^{C_m} \phi(\mathbf{z}_{mx})^T \right) \left(\frac{1}{C_i} \sum_{y=1}^{C_i} \phi(\mathbf{z}_{iy}) \right) \\
&\quad \times \left(\frac{1}{C_i} \sum_{s=1}^{C_i} \phi(\mathbf{z}_{is})^T \right) \left(\frac{1}{N} \sum_{n=1}^C \sum_{t=1}^{C_n} \phi(\mathbf{z}_{tn}) \right) \\
&= \frac{1}{NC_i} \sum_{m=1}^C \sum_{x=1}^{C_m} \sum_{y=1}^{C_i} (k_{xy})_{mi} \cdot \frac{1}{C_i N} \sum_{s=1}^{C_i} \sum_{n=1}^C \sum_{t=1}^{C_n} (k_{st})_{in} \\
&= \left(\frac{1}{N} \mathbf{1}_{N1} \right)^T \cdot \mathbf{K} \cdot (\mathbf{Q}_{N1}^i) \cdot (\mathbf{Q}_{N1}^i)^T \cdot \mathbf{K} \cdot \left(\frac{1}{N} \mathbf{1}_{N1} \right) \\
\Rightarrow (\bar{\phi}^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi})_{n=1, \dots, C}^{m=1, \dots, C} &= \frac{1}{N^2} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC},
\end{aligned}$$

where $\mathbf{W}_{NN}^i = \text{diag}[\mathbf{0}_{C_1}, \dots, \mathbf{0}_{C_{i-1}}, \mathbf{1}_{C_i}, \mathbf{0}_{C_{i+1}}, \dots, \mathbf{0}_{C_C}]$ is a $N \times N$ block diagonal matrix and $\mathbf{0}_{C_k}$ is a $C_k \times C_k$ matrix with all elements equal to 0 and $\mathbf{0}_{C_i}$ is a $C_i \times C_i$ matrix with all elements equal to $1/C_i^2$.

Therefore,

$$\begin{aligned}
\bar{\phi}_b^T \bar{\phi}_i \bar{\phi}_i^T \bar{\phi}_b &= \frac{1}{N} \mathbf{B} \cdot \left(\mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC} \right. \\
&\quad - \frac{1}{N} \mathbf{A}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC} \\
&\quad - \frac{1}{N} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{A}_{NC} \\
&\quad \left. + \frac{1}{N^2} \mathbf{1}_{NC}^T \cdot \mathbf{K} \cdot \mathbf{W}_{NN}^i \cdot \mathbf{K} \cdot \mathbf{1}_{NC} \right) \cdot \mathbf{B}. \quad (19)
\end{aligned}$$

References

- [1] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach. Intell.* 19 (7) (1997) 711–720.
- [2] L.-F. Chen, H.-Y.M. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, *Pattern Recognition* 33 (2000) 1713–1726.

- [3] H. Yu, J. Yang, A direct lda algorithm for high-dimensional data with application to face recognition, *Pattern Recognition* 34 (2001) 2067–2070.
- [4] J. Lu, K. Plataniotis, A. Venetsanopoulos, Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, *Pattern Recognition Lett.* 26 (2) (2005) 181–191.
- [5] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using LDA based algorithms, *IEEE Trans. Neural Networks* 14 (1) (2003) 195–200.
- [6] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Regularized discriminant analysis for the small sample size problem in face recognition, *Pattern Recognition Lett.* 24 (16) (2003) 3079–3087.
- [7] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, second ed., Wiley, New York, NY, 2000.
- [8] L. Kanal, B. Chandrasekaran, On dimensionality and sample size in statistical pattern classification, *Pattern Recognition* 3 (1971) 238–255.
- [9] P. Wald, R. Kronmal, Discriminant functions when covariance are unequal and sample sizes are moderate, *Biometrics* 33 (1977) 479–484.
- [10] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, *IEEE Trans. Pattern Anal. Mach. Intell.* 13 (3) (1991) 252–264.
- [11] J.H. Friedman, Regularized discriminant analysis, *J. Am. Statist. Assoc.* 84 (1989) 165–175.
- [12] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- [13] B. Scholkopf, C. Burges, A.J. Smola, *Advances in Kernel Methods—Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
- [14] A. Ruiz, P.L. de Teruel, Nonlinear kernel-based statistical pattern analysis, *IEEE Trans. Neural Networks* 12 (1) (2001) 16–32.
- [15] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, *IEEE Trans. Neural Networks* 12 (2) (2001) 181–201.
- [16] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.
- [17] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans. Neural Networks* 14 (1) (2003) 117–126.
- [18] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, The FERET database and evaluation procedure for face recognition algorithms, *Image Vision Comput.* 16 (5) (1998) 295–306.
- [19] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (10) (2000) 1090–1104.
- [20] B. Scholkopf, A. Smola, K. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Comput.* 10 (1999) 1299–1319.
- [21] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Comput.* 12 (2000) 2385–2404.
- [22] C. Blake, E. Keogh, C.J. Merz, uci repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, (<http://www.ics.uci.edu/mllearn>), 1998.

About the Author—JIE WANG received both the B.Eng. and M.Sci. degree in electronic engineering from Zhejiang University, P.R. China, in 1998 and 2001, respectively. Currently, she is pursuing the Ph.D. degree in the Edward S. Rogers, Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada. Her research interests include face detection and recognition, multi-modal biometrics and kernel methods.

About the Author—K.N. PLATANIOTIS received the B.Eng. degree in computer engineering and informatics from University of Patras, Greece, in 1988 and the M.S. and Ph.D. degrees in electrical engineering from Florida Institute of Technology (Florida Tech), Melbourne, in 1992 and 1994, respectively. He is an Associate Professor of Electrical and Computer Engineering at the University of Toronto, Canada. His research interests are in the areas of multimedia systems, biometrics, image and signal processing, communications systems and pattern recognition. Dr. Plataniotis is a registered professional engineer in the province of Ontario.

About the Author—JUWEI LU received the B.Eng. degree in electrical engineering from Nanjing University of Aeronautics and Astronautics, China, in 1994 and the M.Eng. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 1999 and the Ph.D. degree from the Edward S. Rogers, Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada, in 2004. His research interests include multimedia signal processing, face detection and recognition, kernel methods, support vector machines, neural networks, and boosting technologies.

About the Author—A.N. VENETSANOPOULOS received the Bachelors of Electrical and Mechanical Engineering degree from the National Technical University of Athens (NTU), Greece, in 1965, and the M.S., M.Phil., and Ph.D. degrees in Electrical Engineering from Yale University in 1966, 1968 and 1969, respectively. He is a Professor of Electrical and Computer Engineering at the University of Toronto, Canada. Between 2001–2006 he served as the 12th Dean of the Faculty of Applied Science and Engineering of the University of Toronto. Now he is the Vice President, Research and Innovation of Ryerson University. His research interests include multimedia (image compression, database retrieval), digital signal/image processing (multi channel image processing, nonlinear, adaptive and M-D filtering), digital communications (image transmission, image compression), neural networks and fuzzy reasoning in signal/image processing.