Face Recognition Using Kernel Direct Discriminant Analysis Algorithms

Juwei Lu, *Student Member, IEEE*, Konstantinos N. Plataniotis, *Member, IEEE*, and
Anastasios N. Venetsanopoulos, *Fellow, IEEE*

***Abstract*—Techniques that can introduce low-dimensional feature representation with enhanced discriminatory power is of paramount importance in face recognition (FR) systems. It is well known that the distribution of face images, under a perceivable variation in viewpoint, illumination or facial expression, is highly nonlinear and complex. It is, therefore, not surprising that linear techniques, such as those based on principle component analysis (PCA) or linear discriminant analysis (LDA), cannot provide reliable and robust solutions to those FR problems with complex face variations. In this paper, we propose a kernel machine-based discriminant analysis method, which deals with the nonlinearity of the face patterns' distribution. The proposed method also effectively solves the so-called "small sample size" (SSS) problem, which exists in most FR tasks. The new algorithm has been tested, in terms of classification error rate performance, on the multiview UMIST face database. Results indicate that the proposed methodology is able to achieve excellent performance with only a very small set of features being used, and its error rate is approximately 34% and 48% of those of two other commonly used kernel FR approaches, the kernel-PCA (KPCA) and the generalized discriminant analysis (GDA), respectively.**

***Index Terms*—Face recognition (FR), kernel direct discriminant analysis (KDDA), linear discriminant analysis (LDA), principle component analysis (PCA), small sample size problem (SSS), kernel methods.**

I. Introduction

WITHIN the last decade, face recognition (FR) has found a wide range of applications, from identity authentication, access control, and face-based video indexing/browsing, to human-computer interaction/communication. As a result, numerous FR algorithms have been proposed, and surveys in this area can be found in [1]–[5]. Two issues are central to all these algorithms: 1) feature selection for face representation and 2) classification of a new face image based on the chosen feature representation [6]. This work focuses on the issue of feature selection. The main objective is to find techniques that can introduce low-dimensional feature representation of face objects with enhanced discriminatory power. Among various solutions to the problem, the most successful are those appearance-based approaches, which generally operate directly on images or appearances of face objects and process the images as two-dimensional (2-D) holistic patterns, to avoid difficulties associated with three-dimensional (3-D) modeling, and shape or landmark detection [5].

Principle component analysis (PCA) and linear discriminant analysis (LDA) are two classic tools widely used in the appearance-based approaches for data reduction and feature extraction. Many state-of-the-art FR methods, such as Eigenfaces [7] and Fisherfaces [8], are built on these two techniques or their variants. It is generally believed that when it comes to solving problems of pattern classification, LDA-based algorithms outperform PCA-based ones, since the former optimizes the low-dimensional representation of the objects with focus on the most discriminant feature extraction while the latter achieves simply object reconstruction. However, many LDA-based algorithms suffer from the so-called "*small sample size problem*" (SSS) which exists in high-dimensional pattern recognition tasks, where the number of available samples is smaller than the dimensionality of the samples. The traditional solution to the SSS problem is to utilize PCA concepts in conjunction with LDA (PCA+LDA), as it was done for example in Fisherfaces [8]. Recently, more effective solutions, called direct LDA (D-LDA) methods, have been presented [9], [10]. Although successful in many cases, linear methods fail to deliver good performance when face patterns are subject to large variations in viewpoints, which results in a highly nonconvex and complex distribution. The limited success of these methods should be attributed to their linear nature [11]. As a result, it is reasonable to assume that a better solution to this inherent nonlinear problem could be achieved using nonlinear methods, such as the so-called kernel machine techniques [12]–[15].

In this paper, motivated by the success that support vector machines (SVMs) [16]–[18], kernel PCA (KPCA) [19] and generalized discriminant analysis (GDA) [20] have in pattern regression and classification tasks, we propose a new kernel discriminant analysis algorithm for face recognition. The algorithm generalizes the strengths of the recently presented D-LDA and the kernel techniques while at the same time overcomes many of their shortcomings and limitations. Therefore, the proposed algorithm can be seen as an enhanced kernel D-LDA method (hereafter KDDA). Following the SVM paradigm, we first nonlinearly map the original input space to an implicit high-dimensional feature space, where the distribution of face patterns is hoped to be linearized and simplified. Then, a new variant of the D-LDA method is introduced to effectively solve the SSS problem and derive a set of optimal discriminant basis vectors in the feature space.

The rest of this paper is organized as follows. Since KDDA is built on D-LDA and GDA, in Section II, we start the analysis by briefly reviewing the two latter methods. Following that,

Manuscript received December 12, 2001; revised July 1, 2002.

The authors are with the Multimedia Laboratory, Edward S. Rogers, Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: kostas@dsp.toronto.edu).

Digital Object Identifier 10.1109/TNN.2002.806629

KDDA is introduced and analyzed. The relationship of KDDA to D-LDA and GDA is also discussed. In Section III, two sets of experiments are presented to demonstrate the effectiveness of the KDDA algorithm on highly nonlinear highly complex face pattern distributions. KDDA is compared, in terms of the classification error rate performance, to KPCA and GDA on the multiview UMIST face database. Conclusions are summarized in Section IV.

II. Methods

The problem to be solved is formally stated as follows: A set of L training face images $\{\mathbf{z}_i\}_{i=1}^{L}$ is available. Each image is defined as a vector of length $N(=I_w \times I_h)$, i.e., $\mathbf{z}_i \in \mathbb{R}^N$, where $I_w \times I_h$ is the face image size and $\mathbb{R}^N$ denotes a N-dimensional real space. It is further assumed that each image belongs to one of C classes $\{\mathbf{Z}_i\}_{i=1}^{C}$. The objective is to find a transformation φ, based on optimization of certain separability criteria, which produces a mapping $\mathbf{y}_i = \varphi(\mathbf{z}_i)$, with $\mathbf{y}_i \in \mathbb{R}^M$ that leads to an enhanced separability of different face objects.

A. GDA

For solving nonlinear problems, the classic LDA has been generalized to its kernel version, namely GDA [20]. Let $\phi : \mathbf{z} \in \mathbb{R}^N \rightarrow \phi(\mathbf{z}) \in \mathbb{F}$ be a nonlinear mapping from the input space to a high-dimensional feature space $\mathbb{F}$, where different classes of objects are supposed to be linearly separable. The idea behind GDA is to perform a classic LDA in the feature space $\mathbb{F}$ instead of the input space $\mathbb{R}^N$.

Let $\mathbf{S}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{WTH}}$ be the between- and within-class scatter matrices in the feature space $\mathbb{F}$, respectively, expressed as follows:

$$\mathbf{S}_{\mathrm{BTW}} = \frac{1}{L}\sum_{i=1}^{C} C_i(\bar{\phi}_i - \bar{\phi})(\bar{\phi}_i - \bar{\phi})^T \tag{1}$$

$$\mathbf{S}_{\mathrm{WTH}} = \frac{1}{L}\sum_{i=1}^{C}\sum_{j=1}^{C_i} (\phi_{ij} - \bar{\phi}_i)(\phi_{ij} - \bar{\phi}_i)^T \tag{2}$$

where $\phi_{ij} = \phi(\mathbf{z}_{ij}), \bar{\phi}_i = (1/C_i)\sum_{j=1}^{C_i}\phi(\mathbf{z}_{ij})$ is the mean of class $\mathbf{Z}_i, \bar{\phi} = \frac{1}{L}\sum_{i=1}^{C}\sum_{j=1}^{C_i}\phi(\mathbf{z}_{ij})$ is the average of the ensemble, and C_i is the element number in $\mathbf{Z}_i$, which leads to $L = \sum_{i=1}^{C} C_i$. LDA determines a set of optimal discriminant basis vectors, denoted by $\{\psi_k\}_{k=1}^{M}$, so that the ratio of the between- and within-class scatters is maximized [21]. Assuming $\Psi = [\psi_1, \ldots, \psi_M]$, the maximization can be achieved by solving the following eigenvalue problem:

$$\Psi = \arg\max_{\Psi} \frac{|(\Psi^T \mathbf{S}_{\mathrm{BTW}} \Psi)|}{|(\Psi^T \mathbf{S}_{\mathrm{WTH}} \Psi)|}. \tag{3}$$

The feature space $\mathbb{F}$ could be considered as a "linearization space" [22], however, its dimensionality could be arbitrarily large, and possibly infinite. Fortunately, the exact $\phi(\mathbf{z})$ is not needed and the feature space can become implicit by using kernel methods, where dot products in $\mathbb{F}$ are replaced with a kernel function in the input space $\mathbb{R}^N$ so that the nonlinear mapping is performed implicitly in $\mathbb{R}^N$ [23], [24].

In FR tasks, the number of training samples, L, is in most cases much smaller than the dimensionality of $\mathbb{R}^N$ (for LDA) or $\mathbb{F}$ (for GDA) leading to a degenerated scatter matrix $\mathbf{S}_{\mathrm{WTH}}$. Traditional methods, for example GDA and Fisherfaces [8], attempt to solve the so-called SSS problem by using techniques such as pseudo inverse or PCA to remove the null space of $\mathbf{S}_{\mathrm{WTH}}$. However, it has been recently shown that the null space may contain the most significant discriminant information [9], [10].

B. Direct LDA (D-LDA)

Recently, Chen *et al.* [9] and Yang *et al.* [10] proposed the so-called direct LDA (D-LDA) algorithm that attempts to avoid the shortcomings existing in traditional solutions to the SSS problem. The basic idea behind the algorithm is that the null space of $\mathbf{S}_{\mathrm{WTH}}$ may contain significant discriminant information if the projection of $\mathbf{S}_{\mathrm{BTW}}$ is not zero in that direction, and that no significant information will be lost if the null space of $\mathbf{S}_{\mathrm{BTW}}$ is discarded. Assuming, for example, that $\mathcal{A}$ and $\mathcal{B}$ represent the null spaces of $\mathbf{S}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{WTH}}$, respectively, the complement spaces of $\mathcal{A}$ and $\mathcal{B}$ can be written as $\mathcal{A}' = \mathbb{R}^N - \mathcal{A}$ and $\mathcal{B}' = \mathbb{R}^N - \mathcal{B}$. Therefore, the optimal discriminant subspace sought by the D-LDA algorithm is the intersection space $(\mathcal{A}' \cap \mathcal{B})$.

The difference between Chen's method [9] and Yang's method [10] is that Yang's method first diagonalizes $\mathbf{S}_{\mathrm{BTW}}$ to find $\mathcal{A}'$ when seek solution of (3), while Chen's method first diagonalizes $\mathbf{S}_{\mathrm{WTH}}$ to find $\mathcal{B}$. Although there is no significant difference between the two approaches, it may be intractable to calculate $\mathcal{B}$ when the size of $\mathbf{S}_{\mathrm{WTH}}$ is large, which is the case in most FR applications. For example, the size of $\mathbf{S}_{\mathrm{WTH}}$ and $\mathbf{S}_{\mathrm{BTW}}$ amounts to $10\,304 \times 10\,304$ for face images of size 112×92 such as those used in our experiments. Fortunately, the rank of $\mathbf{S}_{\mathrm{BTW}}$ is determined by $\mathrm{rank}(\mathbf{S}_{\mathrm{BTW}}) = \min(N, C-1)$, with C the number of image classes, usually a small value in most of FR tasks, e.g., $C = 20$ in our experiments, resulting in $\mathrm{rank}(\mathbf{S}_{\mathrm{BTW}}) = 19$. $\mathcal{A}'$ can be easily found by solving eigenvectors of a 19×19 matrix rather than the original $10\,304 \times 10\,304$ matrix through the algebraic transformation proposed in [7]. The intersection space $(\mathcal{A}' \cap \mathcal{B})$ can be obtained by solving the null space of projection of $\mathbf{S}_{\mathrm{WTH}}$ into $\mathcal{A}'$, with the projection being a small matrix of size 19×19. For the reasons explained above, we proceed by first diagonalizing the matrix $\mathbf{S}_{\mathrm{BTW}}$ instead of $\mathbf{S}_{\mathrm{WTH}}$ in the derivation of the proposed here algorithm.

C. KDDA

1) Eigen-Analysis of $\mathbf{S}_{\mathrm{BTW}}$ *in the Feature Space:* Following the general D-LDA framework, we start by solving the eigenvalue problem of $\mathbf{S}_{\mathrm{BTW}}$, which can be rewritten here as follows:

$$\begin{aligned}\mathbf{S}_{\mathrm{BTW}} &= \sum_{i=1}^{C}\left(\sqrt{\frac{C_i}{L}}(\bar{\phi}_i - \bar{\phi})\right)\left(\sqrt{\frac{C_i}{L}}(\bar{\phi}_i - \bar{\phi})\right)^T \\ &= \sum_{i=1}^{C} \tilde{\bar{\phi}}_i \tilde{\bar{\phi}}_i^{\,T} = \Phi_b \Phi_b^T \end{aligned} \tag{4}$$

where $\tilde{\bar{\phi}}_i = \sqrt{(C_i/L)}(\bar{\phi}_i - \bar{\phi})$, and $\Phi_b = [\tilde{\bar{\phi}}_1 \ \cdots \ \tilde{\bar{\phi}}_c]$. Since the dimensionality of the feature space $\mathbb{F}$, denoted as N', could be arbitrarily large or possibly infinite, it is intractable to directly compute the eigenvectors of the $(N' \times N')$ matrix $\mathbf{S}_{\mathrm{BTW}}$. Fortunately, the first $m(\leq C-1)$ most significant eigenvectors of $\mathbf{S}_{\mathrm{BTW}}$, which correspond to nonzero eigenvalues, can be indirectly derived from the eigenvectors of the matrix $\Phi_b^T\Phi_b$ (with size $C \times C$) [7].

Computing $\Phi_b^T\Phi_b$, requires dot product evaluation in $\mathbb{F}$. This can be done in a manner similar to the one used in SVM, KPCA, and GDA by utilizing kernel methods. For any $\phi(\mathbf{z}_i), \phi(\mathbf{z}_j) \in \mathbb{F}$, we assume that there exists a kernel function $k(\cdot)$ such that $k(\mathbf{z}_i, \mathbf{z}_j) = \phi(\mathbf{z}_i) \cdot \phi(\mathbf{z}_j)$. The introduction of the kernel function allows us to avoid the explicit evaluation of the mapping. Any function satisfying Mercer's condition can be used as a kernel, and typical kernel functions include polynomial function, radial basis function (RBF) and multilayer perceptrons [17].

Using the kernel function, for two arbitrary classes $\mathbf{Z}_l$ and $\mathbf{Z}_h$, a $C_l \times C_h$ dot product matrix K_{lh} can be defined as

$$K_{lh} = (k_{ij})_{\substack{i=1,\ldots,C_l\\ j=1,\ldots,C_h}}, \quad \text{where} \quad k_{ij} = k(\mathbf{z}_{li}, \mathbf{z}_{hj}) = \phi_{li} \cdot \phi_{hj} \tag{5}$$

For all of C classes $\{\mathbf{Z}_i\}_{i=1}^{C}$, we then define a $L \times L$ kernel matrix $\mathbf{K}$

$$\mathbf{K} = (K_{lh})_{\substack{l=1,\ldots,C\\ h=1,\ldots,C}} \tag{6}$$

which allows us to express $\Phi_b^T\Phi_b$ as follows:

$$\Phi_b^T\Phi_b = \frac{1}{L}\mathbf{B} \cdot \left(\mathbf{A}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{A}_{LC} - \frac{1}{L}\left(\mathbf{A}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{1}_{LC}\right) \right.$$
$$\left. - \frac{1}{L}\left(\mathbf{1}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{A}_{LC}\right) + \frac{1}{L^2}\left(\mathbf{1}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{1}_{LC}\right)\right) \cdot \mathbf{B} \tag{7}$$

where $\mathbf{B} = \mathbf{diag}[\sqrt{C_1} \ \cdots \ \sqrt{C_c}]$, $\mathbf{1}_{LC}$ is a $L \times C$ matrix with terms all equal to: one, $\mathbf{A}_{LC} = \mathbf{diag}[\mathbf{a}_{c_1} \ \cdots \ \mathbf{a}_{c_c}]$ is a $L \times C$ block diagonal matrix, and $\mathbf{a}_{c_i}$ is a $C_i \times 1$ vector with all terms equal to: $(1/C_i)$ (see Appendix I for a detailed derivation of (7).).

Let λ_i and $\mathbf{e}_i(i = 1 \ldots C)$, be the ith eigenvalue and corresponding eigenvector of $\Phi_b^T\Phi_b$, sorted in *decreasing* order of eigenvalues. Since $(\Phi_b\Phi_b^T)(\Phi_b\mathbf{e}_i) = \lambda_i(\Phi_b\mathbf{e}_i), \mathbf{v}_i = \Phi_b\mathbf{e}_i$ is the eigenvector of $\mathbf{S}_{\mathrm{BTW}}$. In order to remove the null space of $\mathbf{S}_{\mathrm{BTW}}$, we only use its first $m(\leq C-1)$ eigenvectors: $\mathbf{V} = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_m] = \Phi_b\mathbf{E}_m$ where $\mathbf{E}_m = [\mathbf{e}_1 \ \cdots \ \mathbf{e}_m]$, whose corresponding eigenvalues are greater than 0. It is not difficult to see that $\mathbf{V}^T\mathbf{S}_{\mathrm{BTW}}\mathbf{V} = \Lambda_b$, with $\Lambda_b = \mathbf{diag}[\lambda_1^2 \ldots \lambda_m^2]$, a $m \times m$ diagonal matrix.

2) Eigen-Analysis of $\mathbf{S}_{\mathrm{WTH}}$ *in the Feature Space:* Let $\mathbf{U} = \mathbf{V}\Lambda_b^{-1/2}$. Projecting $\mathbf{S}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{WTH}}$ into the subspace spanned by $\mathbf{U}$, it can easily be seen that $\mathbf{U}^T\mathbf{S}_{\mathrm{BTW}}\mathbf{U} = \mathbf{I}$ while $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U}$ can be expanded as

$$\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U} = \left(\mathbf{E}_m\Lambda_b^{-1/2}\right)^T \left(\Phi_b^T\mathbf{S}_{\mathrm{WTH}}\Phi_b\right) \left(\mathbf{E}_m\Lambda_b^{-1/2}\right). \tag{8}$$

Using the kernel matrix $\mathbf{K}$, a closed-form expression of $\Phi_b^T\mathbf{S}_{\mathrm{WTH}}\Phi_b$ can be obtained as follows:

$$\Phi_b^T\mathbf{S}_{\mathrm{WTH}}\Phi_b = \frac{1}{L}(\mathbf{J1} - \mathbf{J2}) \tag{9}$$

with $\mathbf{J1}$ and $\mathbf{J2}$ defined in Appendix II along with the detailed derivation of the expression in (9).

We proceed by diagonalizing $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U}$, a tractable matrix with size $m \times m$. Let $\mathbf{p}_i$ be the ith eigenvector of $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U}$, where $i = 1 \ldots m$, sorted in *increasing* order of the corresponding eigenvalue λ_i'. In the set of ordered eigenvectors, those that correspond to the smallest eigenvalues maximize the ratio in (3), and should be considered the most discriminative features. Discarding the eigenvectors with the largest eigenvalues, the $M(\leq m)$ selected eigenvectors are denoted as $\mathbf{P} = [\mathbf{p}_1 \ldots \mathbf{p}_M]$. Defining a matrix $\mathbf{Q} = \mathbf{UP}$, we can obtain $\mathbf{Q}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{Q} = \Lambda_w$, with $\Lambda_w = \mathbf{diag}[\lambda_1' \ldots \lambda_M']$, a $M \times M$ diagonal matrix.

Based on the calculations presented above, a set of optimal discriminant feature vectors can be derived through $\Gamma = \mathbf{Q}\Lambda_w^{-1/2}$. The features form a low-dimensional subspace in $\mathbb{F}$, where the ratio in (3) is maximized. Similar to the D-LDA framework, the subspace obtained contains the intersection space $(\mathcal{A}' \cap \mathcal{B})$ shown in Section II-B. However, it is possible that there exist eigenvalues with $\lambda_i' = 0$ in Λ_w. To alleviate the problem, threshold values were introduced in [10], where any value below the threshold ϵ is promoted to ϵ (a very small value). Obviously, performance heavily depends on the heuristic evaluation of the parameter ϵ.

To robustify the approach, we propose a modified Fisher's criterion to be used instead of the conventional definition in (3) when $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U}$ is singular. The new criterion can be expressed as

$$\Psi = \arg\max_{\Psi} \frac{|(\Psi^T\mathbf{S}_{\mathrm{BTW}}\Psi)|}{|(\Psi^T\mathbf{S}_{\mathrm{BTW}}\Psi) + (\Psi^T\mathbf{S}_{\mathrm{WTH}}\Psi)|.} \tag{10}$$

The modified Fisher's criterion of (10) has been proved to be equivalent to the conventional one (3) in [25]. The expression $\mathbf{U}^T(\mathbf{S}_{\mathrm{BTW}} + \mathbf{S}_{\mathrm{WTH}})\mathbf{U}$ which is used in (10) instead of the $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U}$ can be shown to be nonsingular by the following lemma.

Lemma 1: Suppose $\mathbf{D}$ is a real matrix of size $\mathcal{N} \times \mathcal{N}$, and can be represented by $\mathbf{D} = \Phi\Phi^T$ where Φ is a real matrix of size $\mathcal{N} \times \mathcal{M}$. Then, $(\mathbf{I} + \mathbf{D})$ is positive definite, i.e., $\mathbf{I} + \mathbf{D} > 0$, where $\mathbf{I}$ is a $\mathcal{N} \times \mathcal{N}$ identity matrix.

Proof: Since $\mathbf{D}^T = \mathbf{D}$, $(\mathbf{I} + \mathbf{D})$ is a real symmetric matrix. For any $\mathcal{N} \times 1$ nonzero real vector: $x, x^T(\mathbf{I} + \mathbf{D})x = x^Tx + x^T\mathbf{D}x = x^Tx + (\Phi^Tx)^T(\Phi^Tx) > 0$. According to [26], the matrix $(\mathbf{I} + \mathbf{D})$ that satisfies the above conditions is positive definite. ■

Following a procedure similar to $\mathbf{S}_{\mathrm{BTW}}, \mathbf{S}_{\mathrm{WTH}}$ can be expressed as $\mathbf{S}_{\mathrm{WTH}} = \Phi_w\Phi_w^T$, with $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U} = (\mathbf{U}^T\Phi_w)(\mathbf{U}^T\Phi_w)^T$. Since $\mathbf{U}^T\mathbf{S}_{\mathrm{BTW}}\mathbf{U} = \mathbf{I}$ and $(\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U})$ satisfies the conditions on $\mathbf{D}$ discussed in Lemma 1, $\mathbf{U}^T(\mathbf{S}_{\mathrm{BTW}} + \mathbf{S}_{\mathrm{WTH}})\mathbf{U}$ is positive definite. As a result, $\mathbf{Q}^T(\mathbf{S}_{\mathrm{BTW}} + \mathbf{S}_{\mathrm{WTH}})\mathbf{Q} = \Lambda_w$ is nonsingular.

3) Dimensionality Reduction and Feature Extraction: For any input pattern $\mathbf{z}$, its projection into the set of feature vectors, Γ, derived in Section II-C2, can be calculated by

$$\mathbf{y} = \Gamma^T \phi(\mathbf{z}) = \left(\mathbf{E}_m \cdot \Lambda_b^{-1/2} \cdot \mathbf{P} \cdot \Lambda_w^{-1/2}\right)^T \left(\Phi_b^T \phi(\mathbf{z})\right) \tag{11}$$

where $\Phi_b^T \phi(\mathbf{z}) = [\tilde{\bar{\phi}}_1 \ldots \tilde{\bar{\phi}}_c]^T \phi(\mathbf{z})$. Since

$$\begin{aligned} \tilde{\bar{\phi}}_i^T \phi(\mathbf{z}) &= \left(\sqrt{\frac{C_i}{L}}(\bar{\phi}_i - \bar{\phi})\right)^T \phi(\mathbf{z}) \\ &= \sqrt{\frac{C_i}{L}} \left(\frac{1}{C_i}\sum_{m=1}^{C_i} \phi_{im}^T \phi(\mathbf{z}) - \frac{1}{L}\sum_{p=1}^{C}\sum_{q=1}^{C_p} \phi_{pq}^T \phi(\mathbf{z})\right) \end{aligned} \tag{12}$$

we have

$$\Phi_b^T \phi(\mathbf{z}) = \frac{1}{\sqrt{L}} \mathbf{B} \cdot \left(\mathbf{A}_{LC}^T \cdot \gamma(\phi(\mathbf{z})) - \frac{1}{L} \mathbf{1}_{LC}^T \cdot \gamma(\phi(\mathbf{z}))\right) \tag{13}$$

where $\gamma(\phi(\mathbf{z})) = [\phi_{11}^T \phi(\mathbf{z}) \ \phi_{12}^T \phi(\mathbf{z}) \ \cdots \ \phi_{cc_c}^T \phi(\mathbf{z})]^T$ is a $L \times 1$ kernel vector.

Combining (11) and (13), we obtain

$$\mathbf{y} = \Theta \cdot \gamma(\phi(\mathbf{z})) \tag{14}$$

where $\Theta = (1/\sqrt{L})(\mathbf{E}_m \cdot \Lambda_b^{-1/2} \cdot \mathbf{P} \cdot \Lambda_w^{-1/2})^T (\mathbf{B} \cdot (\mathbf{A}_{LC}^T - (1/L)\mathbf{1}_{LC}^T))$ is a $M \times L$ matrix which can be calculated offline. Thus, through (14), a low-dimensional representation $\mathbf{y}$ on $\mathbf{z}$ with enhanced discriminant power, suitable for classification tasks, has been introduced.

4) Comments: The KDDA method implements an improved D-LDA in a high-dimensional feature space using a kernel approach. Its main advantages can be summarized as follows.

1) KDDA introduces a nonlinear mapping from the input space to an implicit high-dimensional feature space, where the nonlinear and complex distribution of patterns in the input space is "linearized" and "simplified" so that conventional LDA can be applied. It is not difficult to see that KDDA reduces to D-LDA for $\phi(\mathbf{z}) = \mathbf{z}$. Thus, D-LDA can be viewed as a special case of the proposed KDDA framework.

2) KDDA effectively solves the SSS problem in the high-dimensional feature space by employing an improved D-LDA algorithm. Unlike the original D-LDA method of [10], zero eigenvalues of the within-class scatter matrix are never used as divisors in the improved one. In this way, the optimal discriminant features can be exactly extracted from both of inside and outside of $\mathbf{S}_{\mathrm{WTH}}$'s null space.

3) In GDA, to remove the null space of $\mathbf{S}_{\mathrm{WTH}}$, it is required to compute the pseudo inverse of the kernel matrix $\mathbf{K}$, which could be extremely ill-conditioned when certain kernels or kernel parameters are used. Pseudoinversion is based on inversion of the nonzero eigenvalues. Due to round-off errors, it is not easy to identify the true null eigenvalues. As a result, numerical stability problems often occur [14]. However, it can been seen from the derivation of KDDA that such problems are avoided in KDDA. The improvement can be observed also in experimental results reported in Figs. 4(a) and 5(a).

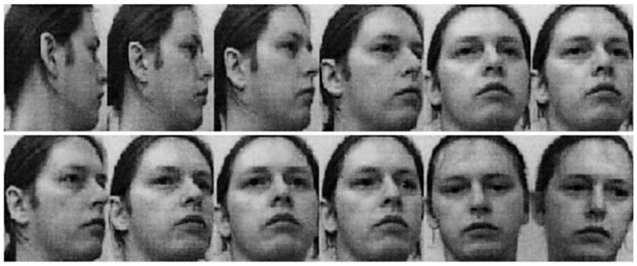


Fig. 1. Some face samples of one subject from the UMIST face database.

The detailed steps for implementing the KDDA method are summarized in Fig. 6.

III. Experimental Results

Two sets of experiments are included in this paper to illustrate the effectiveness of the KDDA algorithm. In all experiments reported here, we utilize the UMIST face database [27], [28], a multiview database, consisting of 575 gray-scale images of 20 subjects, each covering a wide range of poses from profile to frontal views as well as race, gender and appearance. All input images are resized into 112×92, a standardized image size commonly used in FR tasks. The resulting standardized input vectors are of dimensionality $N = 10\,304$. Fig. 1 depicts some sample images of a typical subset in the UMIST database.

A. Distribution of Multiview Face Patterns

The distribution of face patterns is highly nonconvex and complex, especially when the patterns are subject to large variations in viewpoints as is the case with the UMIST database. The first experiment aims to provide insights on how the KDDA algorithm linearizes and simplifies the face pattern distribution.

For the sake of simplicity in visualization, we only use a subset of the database, which contains 170 images of five randomly selected subjects (classes). Four types of feature bases are generalized from the subset by utilizing the PCA, KPCA, D-LDA, and KDDA algorithms, respectively. In the four subspaces produced, two are linear, produced by PCA and D-LDA, and two are nonlinear, produced by KPCA and KDDA. In the sequence, all of images are projected onto the four subspaces. For each image, its projections in the first two most significant feature bases of each subspace are visualized in Figs. 2 and 3.

In Fig. 2, the visualized projections are the first two most significant principal components extracted by PCA and KPCA, and they provide a low-dimensional representation for the samples, which can be used to capture the structure of data. Thus, we can roughly learn the original distribution of face samples from Fig. 2(a), which is nonconvex and complex as we expected based on the analysis presented in the previous sections. In Fig. 2(b), KPCA generalizes PCA to its nonlinear counterpart using a RBF kernel function:

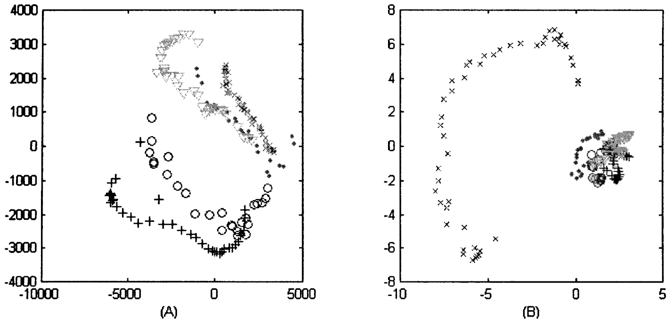


Fig. 2. Distribution of 170 samples of five subjects in PCA- and KPCA-based subspaces. (A) PCA-based subspace ($\subset \mathbb{R}^N$). (B) KPCA-based subspace ($\subset \mathbb{F}$).

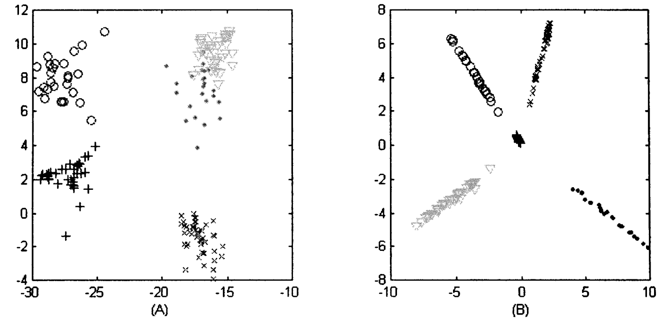


Fig. 3. Distribution of 170 samples of five subjects in D-LDA- and KDDA-based subspaces. (A) D-LDA-based subspace ($\subset \mathbb{R}^N$). (B) KDDA-based subspace ($\subset \mathbb{F}$).

$k(\mathbf{z}_1, \mathbf{z}_2) = \exp((-\|\mathbf{z}_1 - \mathbf{z}_2\|^2)/\sigma^2)$ with $\sigma^2 = 5\mathrm{e}6$. However, it is hard to find any useful improvement for the purpose of pattern classification from Fig. 2(b). It can be concluded, therefore, that the low-dimensional representation obtained by PCA like techniques, achieve simply object reconstruction, and they are not necessarily useful for discrimination and classification tasks [8], [29].

Unlike PCA approaches, LDA optimizes the low-dimensional representation of the objects based on separability criteria. Fig. 3 depicts the first two most discriminant features extracted by utilizing D-LDA and KDDA, respectively. Simple inspection of Figs. 2 and 3 indicates that these features outperform, in terms of discriminant power, those obtained using PCA like methods. However, subject to limitation of linearity, some classes are still nonseparable in the D-LDA-based subspace as shown in Fig. 3(a). In contrast to this, we can see the linearization property of the KDDA-based subspace, as depicted in Fig. 3(b), where all of classes are well linearly separable when a RBF kernel with $\sigma^2 = 5\mathrm{e}6$ is used.

B. Comparison With KPCA and GDA

The second experiment compares the classification error rate performance of the KDDA algorithm to two other commonly used kernel FR algorithms, KPCA and GDA. The FR procedure is completed in two stages:

1) Feature extraction. The overall database is randomly partitioned into two subsets: the training set and test set. The training set is composed of 120 images: Six images per person are randomly chosen. The remaining 455 images are used to form the test set. There is no overlapping between the two. After training is over, both sets are projected into the feature spaces derived from the KPCA, GDA and KDDA methods.
2) Classification. This is implemented by feeding feature vectors obtained in Step 1) into a nearest neighbor classifier. It should be noted at this point that, since the focus in this paper is on feature extraction, a simple classifier is always prefered so that the FR performance is not mainly contributed by the classifier but the feature selection algorithms. We anticipate that the classification accuracy of all the three methods compared here will improve if a more sophisticated classifier such as SVM is used instead of the nearest neighbor. However, such an experiment is beyond the scope of this paper. To enhance the accuracy of performance evaluation, the classification error rates reported in this work are averaged over eight runs. Each run is based on a random partition of the database into the training and test sets. Following the framework introduced in [30], [6], [31], the average error rate, denoted as E_{ave}, is given as follows:

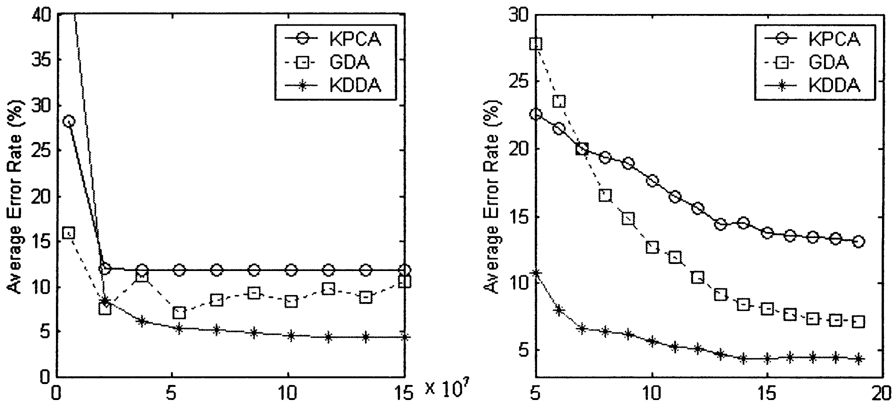
$$E_{\mathrm{ave}} = \left(\sum_{i=1}^{r} t_{\mathrm{mis}}^{i}\right) \Big/ (r \cdot t) \tag{15}$$

where r is the number of runs, t_{mis}^{i} is the number of misclassifications for the ith run, and t is the number of total test samples of each run.

To evaluate the overall performance of the three methods, two typical kernel functions: namely the RBF and the polynomial function, and a wide range of parameter values are tested. Sensitivity analysis is performed with respect to the kernel parameters and the number of used feature vectors M. Figs. 4 and 5 depict the average error rates (E_{ave}) of the three methods compared when the RBF and polynomial kernels are used.

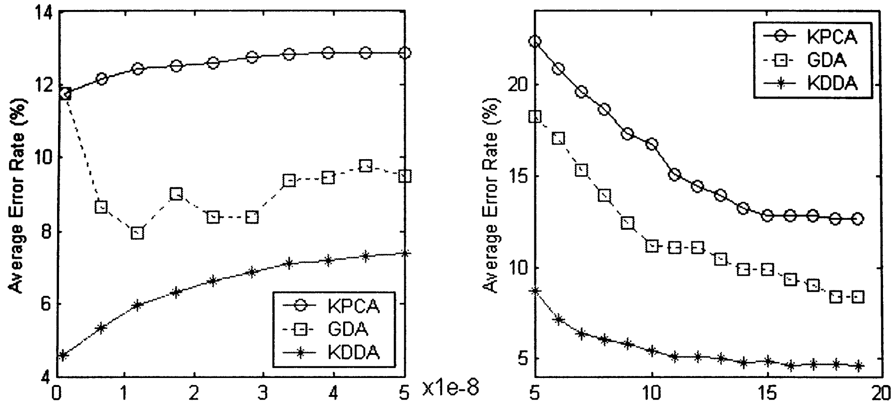
The only kernel parameter for RBF is the scale value σ^2. Fig. 4(a) shows the error rates as functions of σ^2 within the range from $0.5\mathrm{e}7$ to $1.5\mathrm{e}8$, when the optimal number of feature vectors, $M = M_{\mathrm{opt}}$, is used. The optimal feature number is a result of the existence of the peaking effect in the feature selection procedure. It is well known that the classification error initially declines with the addition of new features, attains a minimum, and then starts to increase [32]. The optimal number can be found by searching the number of used feature vectors that results in the minimal summation of the error rates over the variation range of σ^2. In Fig. 4(a), $M_{\mathrm{opt}} = 99$ is the value used for KPCA, while $M_{\mathrm{opt}} = 19$ is used for GDA and KDDA. Fig. 4(b) depicts the error rates as functions of M within the range from 5 to 19, when optimal $\sigma^2 = \sigma_{\mathrm{opt}}^2$ is used. Similar to $M_{\mathrm{opt}}, \sigma_{\mathrm{opt}}^2$ is defined as the scale parameter that results in the minimal summation of the error rates over the variation range of M for the experiment discussed here. In Fig. 4(b), a value $\sigma_{\mathrm{opt}}^2 = 1.5\mathrm{e}8$ is found for KPCA, $\sigma_{\mathrm{opt}}^2 = 5.3333\mathrm{e}7$ for GDA and $\sigma_{\mathrm{opt}}^2 = 1.3389\mathrm{e}7$ for KDDA.

As such, the average error rates of the three methods with polynomial kernel $(k(\mathbf{z}_1, \mathbf{z}_2) = (a \cdot (\mathbf{z}_1 \cdot \mathbf{z}_2) + b)^d)$ are shown in Fig. 5. For the sake of simplicity, we only test the influence of a, while $b = 1$ and $d = 3$ are fixed. Fig. 5(a) depicts the



(a) (b)

Fig. 4. Comparison of error rates based on RBF kernel function. (a) Error rates as functions of σ^2. (b) Error rate as functions of M.



(a) (b)

Fig. 5. Comparison of error rates based on Polynomial kernel function. (a) error rates as functions of a. (b) Error rate as functions of M.

TABLE I
AVERAGE PERCENTAGES OF THE ERROR RATE OF KDDA OVER THOSE OF OTHERS

Kernel	RBF	Polynomial	(RBF+Polynomial)/2
KDDA/KPCA	33.669%	35.081%	34.375%
KDDA/GDA	47.866%	47.664%	47.765%

error rates as functions of a within the range from $1\text{e}-9$ to $5\text{e}-8$, where $M_{\text{opt}} = 100$ for KPCA, $M_{\text{opt}} = 19$ for GDA and KDDA. Fig. 5(b) shows the error rates as functions of M within the range from 5 to 19 with $a_{\text{opt}} = 1\text{e}-9$ for KPCA, $a_{\text{opt}} = 2.822\text{e}-8$ for GDA and $a_{\text{opt}} = 1\text{e}-9$ for KDDA, determined similarly to σ^2_{opt} and M_{opt}.

Let α_M and β_M be the average error rates of KDDA and any one of other two methods respectively, where $M = [5 \ \cdots \ 19]$. From Figs. 4(b) and 5(b), we can obtain an interesting quantity comparison: the average percentages of the error rate of KDDA over those of other methods by $\sum_{M=5}^{19} (\alpha_M/\beta_M)$. The results are tabulated in Table I. The average error rate of KDDA to KPCA and GDA are only about 34.375% and 47.765% respectively. It should be also noted that Figs. 4(a) and 5(a) reveal the numerical stability problems existing in practical implementations of GDA. Comparing the GDA performance to that of KDDA we can easily see that the later is more stable and predictable, resulting in a cost effective determination of parameter values during the training phase.

IV. CONCLUSION

A new FR method has been introduced in this paper. The proposed method combines kernel-based methodologies with discriminant analysis techniques. The kernel function is utilized to map the original face patterns to a high-dimensional feature space, where the highly nonconvex and complex distribution of face patterns is linearized and simplified, so that linear discriminant techniques can be used for feature extraction. The small sample size problem caused by high dimensionality of mapped patterns, is addressed by an improved D-LDA technique which exactly finds the optimal discriminant subspace of the feature space without any loss of significant discriminant information. Experimental results indicate that the performance of the KDDA algorithm is overall superior to those obtained by the KPCA or GDA approaches. In conclusion, the KDDA algorithm is a general pattern recognition method for nonlinearly feature extraction from high-dimensional input patterns without suffering from the SSS problem. We expect that in addition to face recognition, KDDA will provide excellent performance in applications where classification tasks are routinely performed, such as content-based image indexing and retrieval as well as video and audio classification.

```
Input: A set of training face images $\{\mathbf{z}_i\}_{i=1}^{L}$, each of images is represented
  as a $L$-dimensional vector.
Output: A low-dimensional representation $\mathbf{y}$ of $\mathbf{z}$ with enhanced
  discriminatory power.
Algorithm:
  Step 1. Calculate kernel matrix $\mathbf{K}$ using Eq.6.
  Step 2. Calculate $\Phi_b^T\Phi_b$ using Eq.7, and find $\mathbf{E}_m$ and $\Lambda_b$ from $\Phi_b^T\Phi_b$
    in the way shown in section(II-C.1).
  Step 3. Calculate $\mathbf{U}^T\mathbf{S}_{WTH}\mathbf{U}$ using Eq.8 and Eq.9, and
    if $|\mathbf{U}^T\mathbf{S}_{WTH}\mathbf{U}| \neq 0$ then
      /* using the conventional criterion in Eq.3 when $\mathbf{U}^T\mathbf{S}_{WTH}\mathbf{U}$ is nonsingular. */
      Calculate $\mathbf{P}$ and $\Lambda_w$ from $\mathbf{U}^T\mathbf{S}_{WTH}\mathbf{U}$ as shown in section(II-C.2);
    else /* using the modified criterion in Eq.10 when $\mathbf{U}^T\mathbf{S}_{WTH}\mathbf{U}$ is singular. */
      Calculate $\mathbf{P}$ and $\Lambda_w$ from $\mathbf{U}^T(\mathbf{S}_{BTW}+\mathbf{S}_{WTH})\mathbf{U}$ as shown in section(II-C.2);
  Step 4. Calculate $\Theta$ in Eq.14.
  Step 5. For input pattern $\mathbf{z}$, calculate its kernel matrix $\gamma(\phi(\mathbf{z}))$ in Eq.13.
  Step 6. The optimal discriminant feature representation of $\mathbf{z}$ can be obtained
    by $\mathbf{y} = \Theta \cdot \gamma(\phi(\mathbf{z}))$ based on Eq.14.
```

Fig. 6. KDDA pseudocode implementation.

Appendix I
Computation of $(\Phi_b^T\Phi_b)$

Expanding $\Phi_b^T\Phi_b$, we have

$$\Phi_b^T\Phi_b = [\tilde{\bar{\phi}}_1 \ \cdots \ \tilde{\bar{\phi}}_C]^T[\tilde{\bar{\phi}}_1 \ \cdots \ \tilde{\bar{\phi}}_C] = \left(\tilde{\bar{\phi}}_i^{\,T}\tilde{\bar{\phi}}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} \tag{16}$$

where

$$\tilde{\bar{\phi}}_i^{\,T}\tilde{\bar{\phi}}_j = \frac{\sqrt{C_iC_j}}{N}\left(\bar{\phi}_i^T\bar{\phi}_j - \bar{\phi}_i^T\bar{\phi} - \bar{\phi}^T\bar{\phi}_j + \bar{\phi}^T\bar{\phi}\right) \tag{17}$$

We develop each term of (17) according to the kernel matrix $\mathbf{K}$ as follows:

-

$$\bar{\phi}^T\bar{\phi} = \left(\frac{1}{L}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\phi_{lk}\right)^T\left(\frac{1}{L}\sum_{h=1}^{C}\sum_{m=1}^{C_h}\phi_{hm}\right) = \frac{1}{L^2}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\sum_{h=1}^{C}\sum_{m=1}^{C_h}(k_{km})_{lh}$$

$$\Rightarrow \left(\bar{\phi}^T\bar{\phi}\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} = \frac{1}{L^2}\left(\mathbf{1}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right);$$

$$\bar{\phi}^T\bar{\phi}_j = \left(\frac{1}{L}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\phi_{lk}\right)^T\left(\frac{1}{C_j}\sum_{m=1}^{C_j}\phi_{jm}\right) = \frac{1}{LC_j}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\sum_{m=1}^{C_j}(k_{km})_{lj}$$

$$\Rightarrow \left(\bar{\phi}^T\bar{\phi}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} = \frac{1}{L}\left(\mathbf{1}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{A}_{LC}\right);$$

-

$$\bar{\phi}_i^T\bar{\phi} = \left(\frac{1}{C_i}\sum_{m=1}^{C_i}\phi_{im}\right)^T\left(\frac{1}{L}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\phi_{lk}\right) = \frac{1}{LC_i}\sum_{m=1}^{C_i}\sum_{l=1}^{C}\sum_{k=1}^{C_l}(k_{mk})_{il}$$

$$\Rightarrow \left(\bar{\phi}_i^T\bar{\phi}\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} = \frac{1}{L}\left(\mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right);$$

-

$$\bar{\phi}_i^T\bar{\phi}_j = \left(\frac{1}{C_i}\sum_{m=1}^{C_i}\phi_{im}\right)^T\left(\frac{1}{C_j}\sum_{n=1}^{C_j}\phi_{jn}\right) = \frac{1}{C_iC_j}\sum_{m=1}^{C_i}\sum_{n=1}^{C_j}(k_{mn})_{ij}$$

$$\Rightarrow \left(\bar{\phi}_i^T\bar{\phi}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} = \left(\mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{A}_{LC}\right).$$

Applying the above derivations into (17), we obtain the (7).

Appendix II
Computation of $(\Phi_b^T\mathbf{S}_{\mathrm{WTH}}\Phi_b)$

Expanding $\Phi_b^T\mathbf{S}_{\mathrm{WTH}}\Phi_b$, we have

$$\begin{aligned}\Phi_b^T\mathbf{S}_{\mathrm{WTH}}\Phi_b &= [\tilde{\bar{\phi}}_1 \ \cdots \ \tilde{\bar{\phi}}_C]^T\mathbf{S}_{\mathrm{WTH}}[\tilde{\bar{\phi}}_1 \ \cdots \ \tilde{\bar{\phi}}_C]\\ &= \left(\tilde{\bar{\phi}}_i^{\,T}\mathbf{S}_{\mathrm{WTH}}\tilde{\bar{\phi}}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}\end{aligned} \tag{18}$$

where

$$\begin{aligned}\tilde{\bar{\phi}}_i^{\,T}\mathbf{S}_{\mathrm{WTH}}\tilde{\bar{\phi}}_j &= \frac{1}{L}\tilde{\bar{\phi}}_i^{\,T}\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}(\phi_{lk}-\bar{\phi}_l)(\phi_{lk}-\bar{\phi}_l)^T\right)\tilde{\bar{\phi}}_j\\ &= \frac{1}{L}\tilde{\bar{\phi}}_i^{\,T}\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\phi_{lk}\phi_{lk}^T - \sum_{l=1}^{C}\bar{\phi}_l\left(\sum_{k=1}^{C_l}\phi_{lk}^T\right)\right.\\ &\quad\left. - \sum_{l=1}^{C}\left(\sum_{k=1}^{C_l}\phi_{lk}\right)\bar{\phi}_l^T + \sum_{l=1}^{C}C_l\bar{\phi}_l\bar{\phi}_l^T\right)\tilde{\bar{\phi}}_j\\ &= \frac{1}{L}\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\tilde{\bar{\phi}}_i^{\,T}\phi_{lk}\phi_{lk}^T\tilde{\bar{\phi}}_j - \sum_{l=1}^{C}C_l\tilde{\bar{\phi}}_i^{\,T}\bar{\phi}_l\bar{\phi}_l^T\tilde{\bar{\phi}}_j\right)\end{aligned} \tag{19}$$

First, expand the term $\sum_{l=1}^{C}\sum_{k=1}^{C_l}\tilde{\bar{\phi}}_i^{\,T}\phi_{lk}\phi_{lk}^T\tilde{\bar{\phi}}_j$ in (19), and have

$$\begin{aligned}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\tilde{\bar{\phi}}_i^T\phi_{lk}\phi_{lk}^T\tilde{\bar{\phi}}_j &= \frac{\sqrt{C_iC_j}}{L}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\left(\bar{\phi}_i^T\phi_{lk}\phi_{lk}^T\bar{\phi}_j\right.\\ &\quad\left.-\bar{\phi}_i^T\phi_{lk}\phi_{lk}^T\bar{\phi} - \bar{\phi}^T\phi_{lk}\phi_{lk}^T\bar{\phi}_j + \bar{\phi}^T\phi_{lk}\phi_{lk}^T\bar{\phi}\right)\end{aligned} \tag{20}$$

We develop each term of (20) according to the kernel matrix $\mathbf{K}$ as follows:

- $$\begin{aligned}&\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}_i^T\phi_{lk}\phi_{lk}^T\bar{\phi}_j\\ &\quad= \frac{1}{C_iC_j}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\left(\sum_{m=1}^{C_i}\phi_{im}^T\phi_{lk}\right)\left(\sum_{n=1}^{C_j}\phi_{lk}^T\phi_{jn}\right)\\ &\quad= \frac{1}{C_iC_j}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\sum_{m=1}^{C_i}\sum_{n=1}^{C_j}(k_{mk})_{il}(k_{kn})_{lj}\\ &\Rightarrow\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}_i^T\phi_{lk}\phi_{lk}^T\bar{\phi}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}\\ &\quad= \left(\mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{A}_{LC}\right);\end{aligned}$$

- $$\begin{aligned}&\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}_i^T\phi_{lk}\phi_{lk}^T\bar{\phi}\\ &\quad= \frac{1}{LC_i}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\left(\sum_{n=1}^{C_i}\phi_{in}^T\phi_{lk}\right)\left(\sum_{h=1}^{C}\sum_{m=1}^{C_h}\phi_{lk}^T\phi_{hm}\right)\\ &\quad= \frac{1}{LC_i}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\sum_{n=1}^{C_i}\sum_{h=1}^{C}\sum_{m=1}^{C_h}(k_{nk})_{il}(k_{km})_{lh}\\ &\Rightarrow\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}_i^T\phi_{lk}\phi_{lk}^T\bar{\phi}\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}\\ &\quad= \frac{1}{L}\left(\mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right);\end{aligned}$$

- $$\begin{aligned}&\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}^T\phi_{lk}\phi_{lk}^T\bar{\phi}_j\\ &\quad= \frac{1}{LC_j}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\left(\sum_{h=1}^{C}\sum_{m=1}^{C_h}\phi_{hm}^T\phi_{lk}\right)\left(\sum_{n=1}^{C_j}\phi_{lk}^T\phi_{jn}\right)\\ &\quad= \frac{1}{LC_j}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\sum_{h=1}^{C}\sum_{m=1}^{C_h}\sum_{n=1}^{C_j}(k_{mk})_{hl}(k_{kn})_{lj}\\ &\Rightarrow\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}^T\phi_{lk}\phi_{lk}^T\bar{\phi}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}\\ &\quad= \frac{1}{L}\left(\mathbf{1}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{A}_{LC}\right);\end{aligned}$$

- $$\begin{aligned}&\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}^T\phi_{lk}\phi_{lk}^T\bar{\phi}\\ &\quad= \frac{1}{L^2}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\left(\sum_{h=1}^{C}\sum_{m=1}^{C_h}\phi_{hm}^T\phi_{lk}\right)\left(\sum_{p=1}^{C}\sum_{q=1}^{C_p}\phi_{lk}^T\phi_{pq}\right)\\ &\quad= \frac{1}{L^2}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\sum_{h=1}^{C}\sum_{m=1}^{C_h}\sum_{p=1}^{C}\sum_{q=1}^{C_p}(k_{mk})_{hl}(k_{kq})_{lp}\\ &\Rightarrow\left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\bar{\phi}^T\phi_{lk}\phi_{lk}^T\bar{\phi}\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}\\ &\quad= \frac{1}{L^2}\left(\mathbf{1}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right).\end{aligned}$$

Defining $\mathbf{J1} = \left(\sum_{l=1}^{C}\sum_{k=1}^{C_l}\tilde{\bar{\phi}}_i^{T}\phi_{lk}\phi_{lk}^T\tilde{\bar{\phi}}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}$, we conclude

$$\begin{aligned}\mathbf{J1} = \frac{1}{L}\mathbf{B}\cdot&\left(\mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{A}_{LC} - \frac{1}{L}\left(A_{Nc}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right)\right.\\ &- \frac{1}{L}\left(\mathbf{1}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{A}_{LC}\right)\\ &\left.+ \frac{1}{L^2}\left(\mathbf{1}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right)\right)\cdot\mathbf{B}\end{aligned} \tag{21}$$

Expanding the term $\sum_{l=1}^{C}C_l\tilde{\bar{\phi}}_i^{\,T}\bar{\phi}_l\bar{\phi}_l^T\tilde{\bar{\phi}}_j$ in (19), we obtain

$$\begin{aligned}\sum_{l=1}^{C}\sum_{k=1}^{C_l}\tilde{\bar{\phi}}_i^T\bar{\phi}_l\bar{\phi}_l^T\tilde{\bar{\phi}}_j &= \frac{\sqrt{C_iC_j}}{L}\sum_{l=1}^{C}C_l\left(\bar{\phi}_i^T\bar{\phi}_l\bar{\phi}_l^T\bar{\phi}_j\right.\\ &\quad\left.-\bar{\phi}_i^T\bar{\phi}_l\bar{\phi}_l^T\bar{\phi} - \bar{\phi}^T\bar{\phi}_l\bar{\phi}_l^T\bar{\phi}_j + \bar{\phi}^T\bar{\phi}_l\bar{\phi}_l^T\bar{\phi}\right)\end{aligned} \tag{22}$$

Using the kernel matrix $\mathbf{K}$, the terms in (22) can be developed as follows:

- $$\left(\sum_{l=1}^{C}C_l\bar{\phi}_i^T\bar{\phi}_l\bar{\phi}_l^T\bar{\phi}_j\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} = \mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{W}\cdot\mathbf{K}\cdot\mathbf{A}_{LC};$$

- $$\left(\sum_{l=1}^{C}C_l\bar{\phi}_i^T\bar{\phi}_l\bar{\phi}_l^T\bar{\phi}\right)_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}} = \frac{1}{L}\left(\mathbf{A}_{LC}^T\cdot\mathbf{K}\cdot\mathbf{W}\cdot\mathbf{K}\cdot\mathbf{1}_{LC}\right);$$

- $$\left(\sum_{l=1}^{C} C_l \bar{\phi}^T \bar{\phi}_l \bar{\phi}_l^T \bar{\phi}_j\right)_{\substack{i=1,\dots,C\\ j=1,\dots,C}} = \frac{1}{L}\left(\mathbf{1}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{W} \cdot \mathbf{K} \cdot \mathbf{A}_{LC}\right),;$$
- $$\left(\sum_{l=1}^{C} C_l \bar{\phi}^T \bar{\phi}_l \bar{\phi}_l^T \bar{\phi}\right)_{\substack{i=1,\dots,C\\ j=1,\dots,C}} = \frac{1}{L^2}\left(\mathbf{1}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{W} \cdot \mathbf{K} \cdot \mathbf{1}_{LC}\right);$$

where $\mathbf{W} = \mathbf{diag}[\mathbf{w}_1 \ \cdots \ \mathbf{w}_c]$ is a $L \times L$ block diagonal matrix, and $\mathbf{w}_i$ is a $C_i \times C_i$ matrix with terms all equal to: $1/C_i$.

Defining $\mathbf{J2} = (\sum_{l=1}^{C}\sum_{k=1}^{C_l} \tilde{\bar{\phi}}_i^T \bar{\phi}_l \bar{\phi}_l^T \tilde{\bar{\phi}}_j)_{\substack{i=1,\dots,C\\ j=1,\dots,C}}$, and using the above derivations, we conclude that

$$\begin{aligned}\mathbf{J2} = \frac{1}{L}\mathbf{B} \cdot \Big(&\mathbf{A}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{W} \cdot \mathbf{K} \cdot \mathbf{A}_{LC}\\ &- \frac{1}{L}\left(\mathbf{A}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{W} \cdot \mathbf{K} \cdot \mathbf{1}_{LC}\right)\\ &- \frac{1}{L}\left(\mathbf{1}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{W} \cdot \mathbf{K} \cdot \mathbf{A}_{LC}\right)\\ &+ \frac{1}{L^2}\left(\mathbf{1}_{LC}^T \cdot \mathbf{K} \cdot \mathbf{W} \cdot \mathbf{K} \cdot \mathbf{1}_{LC}\right)\Big) \cdot \mathbf{B}.\end{aligned} \tag{23}$$

Thus

$$\Phi_b^T \mathbf{S}_{\mathrm{WTH}} \Phi_b = \left(\tilde{\bar{\phi}}_i^{\,T} \mathbf{S}_{\mathrm{WTH}} \tilde{\bar{\phi}}_j\right)_{\substack{i=1,\dots,C\\ j=1,\dots,C}} = \frac{1}{L}(\mathbf{J1} - \mathbf{J2}). \tag{24}$$

Acknowledgment

The authors would like to thank Dr. D. Graham and Dr. N. Allinson for providing the UMIST face database.

References

[1] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognit.*, vol. 25, pp. 65–77, 1992.

[2] D. Valentin, J. O. Toole Herve Abdi Alice, and G. W. Cottrell, "Connectionist models of face processing: A survey," *Pattern Recognit.*, vol. 27, no. 9, pp. 1209–1230, 1994.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, pp. 705–740, 1995.

[4] S. Gong, S. J. McKenna, and A. Psarrou, *Dynamic Vision From Images to Face Recognition*. Singapore: Imperial College Press, World Scientific, May 2000.

[5] M. Turk, "A random walk through eigenspace," *IEICE Trans. Inform. Syst.*, vol. E84-D, no. 12, pp. 1586–1695, Dec. 2001.

[6] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Networks*, vol. 10, pp. 439–443, Mar. 1999.

[7] M. A. Turk and A. P. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 711–720, July 1997.

[9] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, pp. 1713–1726, 2000.

[10] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data with application to face recognition," *Pattern Recognit.*, vol. 34, pp. 2067–2070, 2001.

[11] M. Bichsel and A. P. Pentland, "Human face recognition and the face image set's topology," *CVGIP: Image Understand.*, vol. 59, pp. 254–261, 1994.

[12] B. Schölkopf, C. Burges, and A. J. Smola, *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[13] (2000). [Online]. Available: http://www.kernel-machines.org

[14] A. Ruiz and P. E. López de Teruel, "Nonlinear kernel-based statistical pattern analysis," *IEEE Trans. Neural Networks*, vol. 12, pp. 16–32, Jan. 2001.

[15] 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, pp. 181–201, Mar. 2001.

[16] C. Cortes and V. N. Vapnik, "Support vector networks," *Machine Learn.*, vol. 20, pp. 273–297, 1995.

[17] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[18] B. Schölkopf, *Support Vector Learning*, Munich, Germany: Oldenbourg-Verlag, 1997.

[19] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1999.

[20] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.

[21] R. A. Fisher, "The use of multiple measures in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[22] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoér, "Theoretical foundations of the potential function method in pattern recognition learning," *Automat. Remote Contr.*, vol. 25, pp. 821–837, 1964.

[23] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[24] B. Scholkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Trans. Neural Networks*, vol. 10, pp. 1000–1017, Sept. 1999.

[25] 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. J. Pattern Recog. Artif. Intell.*, vol. 6, pp. 817–829, 1992.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, MA: Cambridge Univ. Press, 1992.

[27] , D. Graham and N. Allinson. (1998). [Online]. Available: http://images.ee.umist.ac.uk/danny/database.html

[28] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences*, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, Eds., 1998, vol. 163, pp. 446–456.

[29] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 831–836, Aug. 1996.

[30] S. Lawrence, C. Lee Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural network approach," *IEEE Trans. Neural Networks*, vol. 8, pp. 98–113, Jan. 1997.

[31] M. Joo Er, S. Wu, J. Lu, and H. L. Toh, "Face recognition with radial basis function (RBF) neural networks," *IEEE Trans. Neural Networks*, vol. 13, pp. 697–710, May 2002.

[32] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 252–264, May 1991.

Juwei Lu (S'00) 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. Currently, he is pursuing the Ph.D. degree in the Edward S. Rogers, Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada.

From July 1999 to January 2001, he was with the Center for Signal Processing, Singapore, as a Research Engineer. His research interests include multimedia signal processing, face detection and recognition, kernel methods, support vector machines, neural networks, and boosting technologies.

Konstantinos N. Plataniotis (S'88–M'95) received the B.Eng. degree in computer engineering and informatics from University of Patras, 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 was with the Computer Technology Institute (CTI), Patras, Greece from 1989 to 1991. He was a Postdoctoral Fellow at the Digital Signal and Image Processing Laboratory, Department of Electrical and Computer Engineering University of Toronto, Toronto, ON, Canada from November 1994 to May 1997. From September 1997 to June 1999, he was an Assistant Professor with the School of Computer Science at Ryerson University, Toronto. He is now an Assistant Professor with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto, a Nortel Institute for Telecommunications Associate, and an Adjunct Professor in the Department of Mathematics, Physics and Computer Science at Ryerson University. His research interests include multimedia signal processing, intelligent and adaptive systems, and wireless communication systems.

Dr. Plataniotis is a member of the Technical Chamber of Greece.

Anastasios N. Venetsanopoulos (S'66–M'69–SM'79–F'88) received the diploma degree in engineering from the National Technical University of Athens (NTU), Athens, Greece, in 1965, and the M.S., M.Phil., and Ph.D. degrees in electrical engineering from Yale University, New Haven, CT, in 1966, 1968, and 1969 respectively.

He joined the Department of Electrical and Computer Engineering of the University of Toronto, Toronto, ON, Canada, in September 1968, as a Lecturer and he was promoted to Assistant Professor in 1970, Associate Professor in 1973, and Professor in 1981. He has served as Chair of the Communications Group and Associate Chair of the Department of Electrical Engineering. Between July 1997 and June 2001, he was Associate Chair: Graduate Studies of the Department of Electrical and Computer Engineering, and was Acting Chair during spring term 1998–1999. In 1999, a Chair in Multimedia was established in the ECE Department, made possible by a donation of $1.25 M from Bell Canada, matched by $1.0 M of university funds. He assumed the position as Inaugural Chairholder in July 1999, and two additional Assistant Professor positions became available in the same area. Since July 2001, he has served as the 12th Dean of the Faculty of Applied Science and Engineering of the University of Toronto. He was on research leave at the Imperial College of Science and Technology, the National Technical University of Athens, the Swiss Federal Institute of Technology, the University of Florence and the Federal University of Rio de Janeiro, and has also served as Adjunct Professor at Concordia University. He has served as lecturer in 138 short courses to industry and continuing education programs and as Consultant to numerous organizations; he is a contributor to 29 books, a coauthor of *Nonlinear Filters in Image Processing: Principles Applications* (Boston, MA: Kluwer, 1990), and *Artificial Neural Networks: Learning Algorithms, Performance Evaluation and Applications*, (Boston, MA: Kluwer, 1993), *Fuzzy Reasoning in Information Decision and Control Systems*, (Boston, MA: Kluwer, 1994), and *Color Image Processing and Applications* (New York: Springer-Verlag, 2000), and has published over 700 papers in refereed journals and conference proceedings on digital signal and image processing and digital communications.

Prof. Venetsanopoulos has served as Chair on numerous boards, councils, and technical conference committees of the IEEE, such as the Toronto Section (1977–1979) and the IEEE Central Canada Council (1980–1982); he was President of the Canadian Society for Electrical Engineering and Vice President of the Engineering Institute of Canada (EIC) (1983–1986). He was a Guest Editor or Associate Editor for several IEEE journals and the Editor of the *Canadian Electrical Engineering Journal* (1981–1983). He is a member of the IEEE Communications, Circuits and Systems, Computer, and Signal Processing Societies, as well as a member of Sigma Xi, the Technical Chamber of Greece, the European Association of Signal Processing, the Association of Professional Engineers of Ontario (APEO) and Greece. He was elected as a Fellow of the IEEE "for contributions to digital signal and image processing." He is also a Fellow of the EIC, and was awarded an Honorary Doctorate from the National Technical University of Athens, in October 1994. In October 1996, he was awarded the "Excellence in Innovation Award" of the Information Technology Research Centre of Ontario and Royal Bank of Canada, "for innovative work in color image processing and its industrial applications". In November 2000, he became Recipient of the "Millennium Medal of IEEE." In April 2001, he became a Fellow of the Canadian Academy of Engineering, and on July 1, 2001, he was appointed as the twelth Dean of the Faculty of Applied Science and Engineering, University of Toronto.