Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image

Tae-Kyun Kim, *Member*, *IEEE*, and Josef Kittler, *Member*, *IEEE*

Abstract—We present a novel method of nonlinear discriminant analysis involving a set of locally linear transformations called "Locally Linear Discriminant Analysis (LLDA)." The underlying idea is that global nonlinear data structures are locally linear and local structures can be linearly aligned. Input vectors are projected into each local feature space by linear transformations found to yield locally linearly transformed classes that maximize the between-class covariance while minimizing the within-class covariance. In face recognition, linear discriminant analysis (LDA) has been widely adopted owing to its efficiency, but it does not capture nonlinear manifolds of faces which exhibit pose variations. Conventional nonlinear classification methods based on kernels such as generalized discriminant analysis (GDA) and support vector machine (SVM) have been developed to overcome the shortcomings of the linear method, but they have the drawback of high computational cost of classification and overfitting. Our method is for multiclass nonlinear discrimination and it is computationally highly efficient as compared to GDA. The method does not suffer from overfitting by virtue of the linear base structure of the solution. A novel gradient-based learning algorithm is proposed for finding the optimal set of local linear bases. The optimization does not exhibit a local-maxima problem. The transformation functions facilitate robust face recognition in a low-dimensional subspace, under pose variations, using a single model image. The classification results are given for both synthetic and real face data.

Index Terms—Linear discriminant analysis, generalized discriminant analysis, support vector machine, dimensionality reduction, face recognition, feature extraction, pose invariance, subspace representation.

1 Introduction

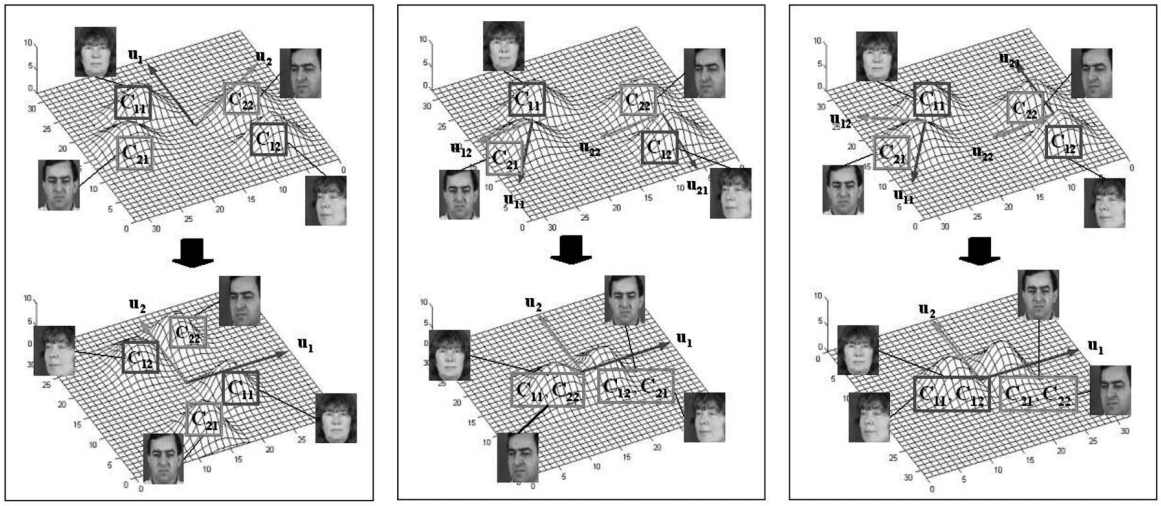
The effectiveness of pattern classification methods can be seriously compromised by various factors which often affect sensory information about an object. Frequently, observations from a single object class are multimodally distributed and samples of objects from different classes in the original data space are more closely located to each other than to those of the same class. The data set of face images taken from a certain number of different viewing angles is a typical example of such problems. It is because the appearance change of face images due to pose changes is usually larger than that caused by different identities. The face manifold is generally known to be continuous with respect to continuous pose changes, as shown in [23]. Although we propose a method for multimodally distributed face classes, the method to be developed may be useful generally, as a continuous pose set can be divided into many subsets of multimodal distributions.

Linear Discriminant Analysis (LDA) [8], [20], [21] is a powerful method for face recognition yielding an effective representation that linearly transforms the original data space into a low-dimensional feature space where the data is as well separated as possible under the assumption that the data classes are Gaussian with equal covariance structure. However, the method fails to solve nonlinear problems, as illustrated in Fig. 1a, because LDA only considers a single linear transformation in a global coordinate system. The multiple linear system [7], [16], [25], which adopts several independent local transformations, attempts to overcome the shortcomings of LDA, but it fails to learn any global data structure, as shown in Fig. 1b. In the LDA mixture model [7], it is assumed that single class objects are distributed normally with an identity covariance matrix structure. It just focuses on maximizing the discriminability of the local structures and it does not make any effort to achieve consistency of the local representations for any single object class. In the upper image of Fig. 1b, the two data sets $\mathbf{C}_{11}$ and $\mathbf{C}_{12}$ corresponding to the different modalities of a class are unfortunately positioned in different directions of the corresponding local components, $\mathbf{u}_{11}$ and $\mathbf{u}_{21}$, therefore, having different representations in a global coordinates as illustrated below. The view-based method for face recognition proposed by Pentland et al. [25] would experience the same difficulty in these circumstances. Following their idea, we could divide images into different pose groups and then train LDA separately for each group. However, because these LDA bases do not encode any relationships of the different pose groups, it is not guaranteed that this "view-based LDA" would yield a consistent representation of different pose images of a single identity. In many conventional face recognition systems [7], [18], [20], [21], [25] which adopt a linear machine such as the LDA or LDA mixture model, as many gallery samples as possible are required so as to capture all the modes of the class distributions.

Support vector machine (SVM) based on kernels has been successfully applied for nonlinear classification problems

- *T.-K. Kim is with the Computing Laboratory, Samsung Advanced Institute of Technology, San 14-1, Nongseo-ri, Kiheung-eup, Yongin, Kyungki-do, Korea, 449-712. E-mail: ktk22@hanmail.net.*
- *J. Kittler is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK. E-mail: j.kittler@eim.surrey.ac.uk.*

Manuscript received 19 Nov. 2003; revised 4 July 2004; accepted 29 July 2004; published online 14 Jan. 2005.
Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0369-1103.



(a) (b) (c)

Fig. 1. (a) LDA, (b) LDA mixture, and (c) LLDA for the nonlinear classification problem: Each upper image shows the simulated data distributions and the components found. In the lower images, the transformed class distributions in the global ouput coordinate system are drawn. The data is generated by $\mathbf{C}_{11} = \{X \sim N(21.6, 2), Y \sim N(21.6, 1)\}$, $\mathbf{C}_{12} = \{X \sim N(7.5, 2), Y \sim N(7.5, 0.8)\}$, $\mathbf{C}_{21} = \{X \sim N(26, 2), Y \sim N(16, 2)\}$, and $\mathbf{C}_{22} = X \sim N(8, 2), Y \sim N(16, 1.2)$, where $N(a, b)$ is a normal variable. Two hundred data points with mean a and standard deviation b are drawn for each mode. $\mathbf{C}_{ij}$ is the jth cluster of the ith class, $\mathbf{u}_{ij}$ is the jth component of the ith cluster, and $\mathbf{u}_i$ denotes the ith component of the output coordinate system.

such as face detection [29], [30]. However, this is inefficient for multiclass recognition and inappropriate when a single sample per class is available to build a class model. Although generalized discriminant analysis (GDA) [2], [14], [22], [31] is suitable for multiclass face recognition problems whereby the original data is mapped into a high-dimensional feature space via a kernel function, it generally has the drawback of high-computational cost in classification and overfitting. In applications such as classification of large data sets on the Internet or video, the computational complexity is particularly important. The global structure of nonlinear manifolds was represented by a locally linear structure in [5], [11]. These methods perform unsupervised learning for locally linear dimensionality reduction, but not a supervised learning for discrimination.

In this paper, several locally linear transformations are concurrently sought so that the class structures manifest by the locally transformed data are well separated in the output space. The proposed method is called "Locally Linear Discriminant Analysis (LLDA)." A preliminary study of the method was reported in [19]. The underlying idea of the proposed approach is that global nonlinear data structures are locally linear and local structures can be linearly aligned. Single class objects, even if multimodally distributed, are transformed into a cluster that is as small as possible, with a maximum distance to the different class objects, by a set of locally linear functions, as illustrated in Fig. 1c. If images of a face class in different poses have similar representations in the trained global subspace, it is much easier to recognize a novel view image even when a single model image is provided.

The advocated method maximizes the separability of classes locally while promoting consistency between the multiple local representations of single class objects. Compared with the conventional nonlinear methods based on kernels, the proposed method is much more computationally efficient because it only involves linear transformations. By virtue of its linear base structure, the proposed method also reduces overfitting normally exhibited by conventional nonlinear methods. The transformation functions learned from the face images of two different views are visualized in Fig. 2a. The functions can be exploited as the bases of a low-dimensional subspace for robust face recognition. The basis functions of each cluster are specific to a particular facial pose. We note two interesting points in this figure compared with the LDA basis images, which are separately trained for different pose groups, view-based LDA. First, the bases of each cluster are similar to those of classical LDA and this ensures that face images of different identities at the same pose are discriminative. Second, the corresponding components of the two different clusters, for example, $\mathbf{u}_{k1}$ and $\mathbf{u}_{l1}$, are aligned to each other. They are characterized by a certain rotation and scaling with similar intensity variation. In consequence, face images of the same identity at different poses have quasi-invariant representation, as shown in Figs. 2a and 2b. For conciseness, only four face classes are plotted in the subspaces of Principal Component Analysis (PCA) [24], view-based LDA, and LLDA in Fig. 2b. Each class has the four samples of two different poses and two different time sessions.

The paper is organized as follows: The next section briefly reviews the conventional methods for linear and nonlinear discriminant analysis. The proposed LLDA method is formulated in Section 3 and a solution of the optimization problem involved is presented in Section 4. Section 5 further simplies the proposed method by replacing the Gaussian mixture model with the case that combines K-means clustering. Section 6 is devoted to the analysis of the computational complexity. Section 7.1 presents the results of experiments performed to demonstrate the beneficial properties of the proposed method on synthetic data. In Section 7.2, the method is applied to face recognition problem. Conclusions are drawn in Section 8.

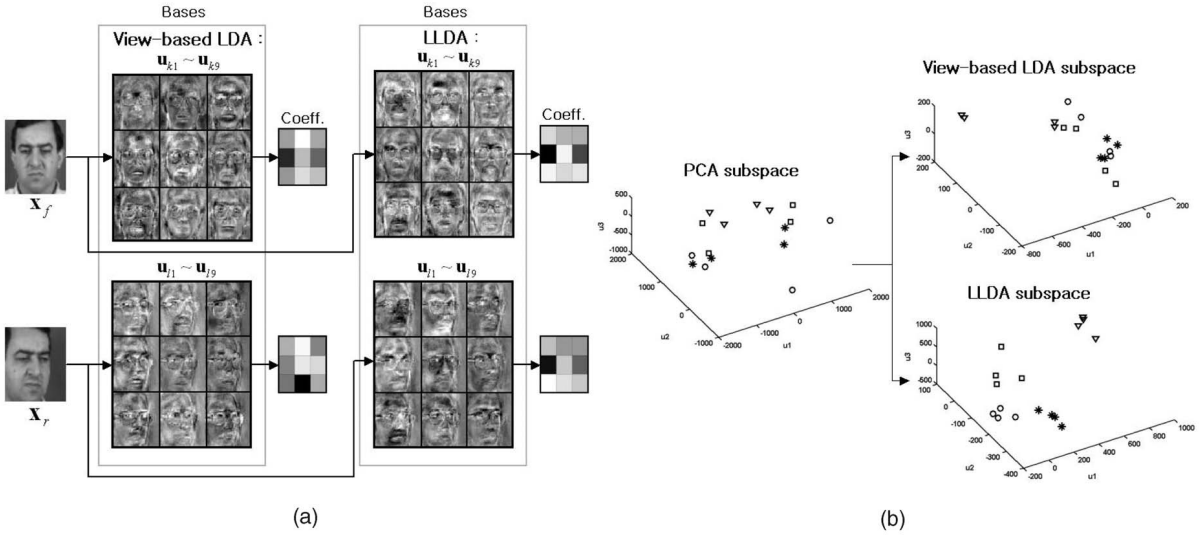


Fig. 2. Representations of LLDA. (a) A flow diagram of LLDA and view-based LDA, where $\mathbf{u}_{ij}$ denotes the jth component of the ith cluster. (b) Plots of some face data in the first three dimensions of PCA, view-based LDA, and LLDA. Different classes are marked as different symbols.

2 Review of Conventional Linear and Nonlinear Discriminant Methods

2.1 Linear Discriminant Analysis

LDA is a class specific method in the sense that it represents data to make it useful for classification [8]. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ be a data set of given N-dimensional vectors of face images. Each data point belongs to one of C object classes $\{\mathbf{X}_1, \ldots, \mathbf{X}_c, \ldots, \mathbf{X}_C\}$. The between-class scatter matrix and the within-class scatter matrix are defined as

$$\mathbf{B} = \sum_{c=1}^{C} M_c(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^{\mathbf{T}},$$

$$\mathbf{W} = \sum_{c=1}^{C} \sum_{\mathbf{x}\in\mathbf{X}_c} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^{\mathbf{T}},$$

where $\mathbf{m}_c$ denotes the class mean and $\mathbf{m}$ is the global mean of the entire sample. The number of vectors in class $\mathbf{X}_c$ is denoted by M_c. LDA finds a matrix, $\mathbf{U}$, maximizing the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix as

$$\mathbf{U}_{opt} = \arg\max_{\mathbf{U}} \frac{\left|\mathbf{U}^{\mathbf{T}}\mathbf{B}\mathbf{U}\right|}{\left|\mathbf{U}^{\mathbf{T}}\mathbf{W}\mathbf{U}\right|} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N].$$

The solution $\{\mathbf{u}_i | i = 1, 2, \ldots, N\}$ is a set of generalized eigenvectors of $\mathbf{B}$ and $\mathbf{W}$, i.e., $\mathbf{B}\mathbf{u}_i = \lambda_i \mathbf{W}\mathbf{u}_i$. Usually, PCA is performed first to avoid a singularity of the within-class scatter matrix commonly encountered in face recognition [20], [21].

2.2 Generalized Discriminant Analysis

The GDA [2] is a method designed for nonlinear classification based on a kernel function Φ which transforms the original space $\boldsymbol{X}$ to a new high-dimensional feature space $\boldsymbol{Z} : \Phi : \boldsymbol{X} \rightarrow \boldsymbol{Z}$. The within-class (or total) scatter and between-class scatter matrix of the nonlinearly mapped data is

$$\mathbf{B}^{\Phi} = \sum_{c=1}^{C} M_c \mathbf{m}_c^{\Phi} \left(\mathbf{m}_c^{\Phi}\right)^{\mathbf{T}}, \ \mathbf{W}^{\Phi} = \sum_{c=1}^{C} \sum_{\mathbf{x}\in\mathbf{X}_c} \Phi(\mathbf{x})\Phi(\mathbf{x})^{\mathbf{T}},$$

where $\mathbf{m}_c^{\Phi}$ is the mean of class $\mathbf{X}_c$ in Z and M_c is the number of samples belonging to $\mathbf{X}_c$. The aim of the GDA is to find such projection matrix $\mathbf{U}^{\Phi}$ that maximizes the ratio

$$\mathbf{U}_{opt}^{\Phi} = \arg\max_{\mathbf{U}^{\Phi}} \frac{\left|(\mathbf{U}^{\Phi})^{\mathbf{T}}\mathbf{B}^{\Phi}\mathbf{U}^{\Phi}\right|}{\left|(\mathbf{U}^{\Phi})^{\mathbf{T}}\mathbf{W}^{\Phi}\mathbf{U}^{\Phi}\right|} = [\mathbf{u}_1^{\Phi}, \ldots, \mathbf{u}_N^{\Phi}].$$

The vectors, $\mathbf{u}^{\Phi}$, can be found as the solution of the generalized eigenvalue problem i.e., $\mathbf{B}^{\Phi}\mathbf{u}_i^{\Phi} = \lambda_i \mathbf{W}^{\Phi}\mathbf{u}_i^{\Phi}$. The training vectors are supposed to be centered (zero mean, unit variance) in the feature space $\boldsymbol{Z}$. From the theory of reproducing kernels, any solution $\mathbf{u}^{\Phi} \in Z$ must lie in the span of all training samples in $\boldsymbol{Z}$, i.e.,

$$\mathbf{u}^{\Phi} = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \alpha_{ci} \Phi(\mathbf{x}_{ci}),$$

where α_{ci} are some real weights and $\mathbf{x}_{ci}$ is the ith sample of class c. The solution is obtained by solving

$$\lambda = \frac{\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{K}\mathbf{D}\mathbf{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{K}\mathbf{K}\boldsymbol{\alpha}},$$

where $\alpha = (\alpha_c), c = 1, \ldots, C$ is a vector of weights with $\alpha_c = (\alpha_{ci}), i = 1, \ldots, M_c$. The kernel matrix $\mathbf{K}(M \times M)$ is composed of the dot products of nonlinearly mapped data, i.e.,

$$\mathbf{K} = (\mathbf{K}_{kl})_{k=1\ldots,C, l=1,\ldots,C},$$

where $\mathbf{K}_{kl} = (k(\mathbf{x}_{ki}, \mathbf{x}_{lj}))_{i=1,\ldots,M_k, j=1,\ldots M_l}$. The matrix $\mathbf{D}$ $(M \times M)$ is a block diagonal matrix such that

$$\mathbf{D} = (\mathbf{D}_c)_{c=1,\ldots,C},$$

where the cth matrix $\mathbf{D}_c$ on the diagonal has all elements equal to $1/M_c$. Solving the eigenvalue problem yields the coefficient

vectors α that define the projection vectors $\mathbf{u}^{\Phi} \in Z$. A projection of a testing vector $\mathbf{x}_{test}$ is computed as

$$(\mathbf{u}^{\Phi})^{\mathbf{T}}\Phi(\mathbf{x}_{test}) = \sum_{c=1}^{C}\sum_{i=1}^{M_c}\alpha_{ci}k(\mathbf{x}_{ci}, \mathbf{x}_{test}).$$

3 Locally Linear Discriminant Analysis (LLDA)

The proposed method, LLDA, is applicable to multiclass nonlinear classification problems by using a set of locally linear transformations. Similarly to the notation adopted in Section 2, consider a data set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ of N-dimensional vectors of face images and C classes $\{\mathbf{X}_1, \ldots, \mathbf{X}_c, \ldots, \mathbf{X}_C\}$. The input vectors are clustered into K subsets denoted by $k, k = 1, \ldots, K$ and each subset k represents a cluster to which a different transformation function is applied. A cluster is defined by K-means clustering or Gaussian mixture modeling of the input vectors. The number of clusters K is chosen to maximize an objective function defined on the training set. Because K usually is a small positive integer, we can make the best choice of K empirically. Assuming that the multimodality of the face data distribution is caused by the different poses, it is pertinent to select K as the number of pose groups. However, general model order selection for a high-dimensional data set remains an open problem. The basic LLDA approach draws on the notion of "soft clustering" in which each data point belongs to each of the clusters with a posterior probability $P(k|\mathbf{x})$. The algorithm that is combined with "hard" K-means clustering will be discussed in Section 5.

We define the locally linear transformation $\mathbf{U}_k = [\mathbf{u}_{k1}, \mathbf{u}_{k2}, \ldots, \mathbf{u}_{kN}], k = 1, \ldots, K$ such that

$$\mathbf{y}_i = \sum_{k=1}^{K} P(k|\mathbf{x}_i)\mathbf{U}_k^{\mathbf{T}}(\mathbf{x}_i - \boldsymbol{\mu}_k), \tag{3.1}$$

where N is the dimension of the transformed space. The mean vector of the kth cluster $\boldsymbol{\mu}_k$ is described by

$$\boldsymbol{\mu}_k = \left(\sum_{i=1}^{M} P(k|\mathbf{x}_i)\mathbf{x}_i\right) \Big/ \left(\sum_{i=1}^{M} P(k|\mathbf{x}_i)\right). \tag{3.2}$$

The locally linear transformation matrices $\mathbf{U}_k$ are concurrently found so as to maximize the criterion function, J. Two objective functions are considered,

$$J_1 = \log\left(|\tilde{\mathbf{B}}|/|\tilde{\mathbf{W}}|\right) \text{ and } J_2 = (1-\alpha)|\tilde{\mathbf{B}}| - \alpha \cdot |\tilde{\mathbf{W}}|, \tag{3.3}$$

where $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{W}}$ are the between-class and within-class scatter matrices in the locally linear transformed feature space, respectively. The constant α takes values from the interval [0 1]. The objective functions maximize the between-class scatter while minimizing the within-class scatter in the locally transformed feature space. One of the differences between the two defined objective functions is manifest in the efficiency of "learning." The log objective function J_1 has the benefit of not requiring a free parameter α, but it is more costly computationally. The function J_2 can efficiently be optimised iteratively, once α is selected. This is exemplified in the subsequent section. In terms of their performance, the two approaches are similar as reported in the experimental Section 7.1.

The global mean $\tilde{\mathbf{m}}$ of all the transformed samples is

$$\tilde{\mathbf{m}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{y}_i = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K}P(k|\mathbf{x}_i)\mathbf{U}_k^{\mathbf{T}}(\mathbf{x}_i - \boldsymbol{\mu}_k), \tag{3.4}$$

where M is the total number of the samples. By substituting for $\boldsymbol{\mu}_k$ from (3.2), we get $\tilde{\mathbf{m}} = \vec{\mathbf{0}}$. The sample mean for class c which consists of M_c samples is given by

$$\tilde{\mathbf{m}}_c = \frac{1}{M_c}\sum_{\mathbf{x}\in \mathrm{X}_c}\mathbf{y} = \sum_{k=1}^{K}\mathbf{U}_k^{\mathbf{T}}\mathbf{m}_{ck}, \tag{3.5}$$

where

$$\mathbf{m}_{ck} = \frac{1}{M_c}\sum_{\mathbf{x}\in \mathrm{X}_c}P(k|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_k).$$

The term $\mathbf{m}_{ck}$ denotes the sample mean of a class c in the kth cluster. Because the transformation is defined with respect to the cluster mean $\boldsymbol{\mu}_k$, the mean of all transformed data of each cluster becomes zero. Using (3.4) and (3.5), the transformed between-class scatter matrix is given as:

$$\begin{aligned}\tilde{\mathbf{B}} &= \sum_{c=1}^{C} M_c(\tilde{\mathbf{m}}_c - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_c - \tilde{\mathbf{m}})^{\mathbf{T}} \\ &= \sum_{c=1}^{C} M_c\left(\sum_{k=1}^{K}\mathbf{U}_k^{\mathbf{T}}\mathbf{m}_{ck}\right)\left(\sum_{k=1}^{K}\mathbf{U}_k^{\mathbf{T}}\mathbf{m}_{ck}\right)^{\mathbf{T}} \\ &= \sum_{k=1}^{K}\mathbf{U}_k^{\mathbf{T}}\mathbf{B}_k\mathbf{U}_k + \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^{\mathbf{T}}\mathbf{B}_{ij}\mathbf{U}_j + \left(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^{\mathbf{T}}\mathbf{B}_{ij}\mathbf{U}_j\right)^{\mathbf{T}},\end{aligned} \tag{3.6}$$

where

$$\mathbf{B}_k = \sum_{c=1}^{C} M_c\mathbf{m}_{ck}\mathbf{m}_{ck}^{\mathbf{T}} \text{ and } \mathbf{B}_{ij} = \sum_{c=1}^{C} M_c\mathbf{m}_{ci}\mathbf{m}_{cj}^{\mathbf{T}}.$$

The between-class scatter matrix consists of the scatter matrices associated with the respective clusters and the correlation matrix of the data samples belonging to two different clusters. The correlation matrix encodes the relationships of the two local structures. Similarly, the within-class scatter is defined by

$$\begin{aligned}\tilde{\mathbf{W}} &= \sum_{c=1}^{C}\sum_{x\in \mathrm{X}_c}(\mathbf{y} - \tilde{\mathbf{m}}_c)(\mathbf{y} - \tilde{\mathbf{m}}_c)^{\mathbf{T}} \\ &= \sum_{k=1}^{K}\mathbf{U}_k^{\mathbf{T}}\mathbf{W}_k\mathbf{U}_k + \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^{\mathbf{T}}\mathbf{W}_{ij}\mathbf{U}_j \\ &\quad + \left(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^{\mathbf{T}}\mathbf{W}_{ij}\mathbf{U}_j\right)^{\mathbf{T}},\end{aligned} \tag{3.7}$$

where

$$\mathbf{W}_k = \sum_{c=1}^{C}\sum_{\mathbf{x}\in \mathrm{X}_c}(P(k|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_k) - \mathbf{m}_{ck})(P(k|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_k) - \mathbf{m}_{ck})^{\mathbf{T}}$$

$$\mathbf{W}_{ij} = \sum_{c=1}^{C}\sum_{\mathbf{x}\in \mathrm{X}_c}(P(i|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i) - \mathbf{m}_{ci})(P(j|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_j) - \mathbf{m}_{cj})^{\mathbf{T}}.$$

Matrix $\mathbf{W}_k$ describes a local cluster and $\mathbf{W}_{ij}$ is the cross-term of two local clusters. Please note that the proposed algorithm without the cross terms $\mathbf{B}_{ij}$ and $\mathbf{W}_{ij}$ would adhere to the same concept as the LDA mixture model by focusing just on the local separability. Moreover, the defined criterion with $K = 1$ is identical to that of the conventional LDA.

4 Gradient-Based Solution for LLDA

In this section, we provide an efficient iterative optimization method based on a gradient learning algorithm for an optimal set of locally linear transformation functions. While it is hard to find good parameters of a kernel function for new data in the conventional GDA, the proposed learning only has parameters which reduce or eliminate overfitting. The discriminant based on such a piecewise linear structure has the benefit of optimizing a convex function with respect to the set of basis vectors of the local coordinates, yielding a unique maximum.

The method is based on a one-basis vector solution for $\mathbf{u}_{k1}, k = 1, \ldots, K$. Other methods based on incremental one-basis at a time solution can be found in [1], [33], [34] for discriminant or independent component analysis criteria. The proposed gradient method yields a global maximum solution by virtue of the criterion function being second-order convex with respect to all the variables $\mathbf{u}_{k1}, k = 1, \ldots, K$. We need to run the one-basis algorithm several times to obtain a multi-dimensional solution $\mathbf{U}_k = [\mathbf{u}_{k1}, \mathbf{u}_{k2}, \ldots, \mathbf{u}_{kN}],\ k = 1, \ldots, K$. The vector orthogonalization is perfomed to prevent different vectors from converging to the same maxima in every iteration. We seek the vectors $\mathbf{u}$ which maximize the criterion function under the constraint of being unit norm vectors:

$$\begin{aligned} &\text{Max } J_1 \text{ or } J_2, \\ &\text{for } \|\mathbf{u}_{kn}\| = 1, k = 1, \ldots, K \text{ and } n = 1, \ldots, N. \end{aligned} \tag{4.1}$$

This constrained optimization problem is solved by the method of projections on the constraint set [1]. A vector normalization imposing a unit norm is executed after every update of the vector. The learning rules are as follows: Do the following steps with an index n starting from 1 to N for $\mathbf{u}_{kn}, k = 1, \ldots, K$.

1. Randomly initialize K unit vectors $\mathbf{u}_{kn}$.
2. Calculate the gradient of the objective function with respect to the variables $\mathbf{u}_k$ by

$$\begin{aligned} \frac{\partial J_1}{\partial \mathbf{u}_{kn}} &= \big(2\tilde{\mathbf{B}}^{-1}\mathbf{B}_k - 2\tilde{\mathbf{W}}^{-1}\mathbf{W}_k\big)\mathbf{u}_{kn} \\ &\quad + \sum_{i=1, i\neq k}^{K} \big(2\tilde{\mathbf{B}}^{-1}\mathbf{B}_{ki} - 2\tilde{\mathbf{W}}^{-1}\mathbf{W}_{ki}\big)\mathbf{u}_{in} \text{ or} \\ \frac{\partial J_2}{\partial \mathbf{u}_{kn}} &= (2(1-\alpha)\mathbf{B}_k - 2\alpha\mathbf{W}_k)\mathbf{u}_{kn} \\ &\quad + \sum_{i=1, i\neq k}^{K} (2(1-\alpha)\mathbf{B}_{ki} - 2\alpha\mathbf{W}_{ki})\mathbf{u}_{in}. \end{aligned} \tag{4.2}$$

3. Update with an appropriate stepsize η as

$$\Delta\mathbf{u}_{kn} \leftarrow \eta \frac{\partial J}{\partial \mathbf{u}_{kn}}. \tag{4.3}$$

4. Carry out the deflationary orthogonalization by

$$\mathbf{u}_{kn} \leftarrow \mathbf{u}_{kn} - \sum_{i=1}^{n-1} \big(\mathbf{u}_{kn}^T \mathbf{u}_{ki}\big)\mathbf{u}_{ki}. \tag{4.4}$$

5. Normalize the vectors $\mathbf{u}_{kn}$ by

$$\mathbf{u}_{kn} \leftarrow \mathbf{u}_{kn} / \|\mathbf{u}_{kn}\|. \tag{4.5}$$

Repeat processes 2 through 5 until the algorithm converges to a stable point, set $n := n + 1$, and then go to Step 1.

Note that the two objective functions have different learning costs. When calculating the gradients of J_2 in (4.2), all the matrices are previously given but the two matrices $\tilde{\mathbf{B}}^{-1}, \tilde{\mathbf{W}}^{-1}$, here scalar values in the one-basis solution, in the learning of J_1 should be iteratively updated. For the synthetic data example given in Fig. 1, the optimization of J_1 takes about 15 times longer than that of J_2. While the learning of J_1 has a benefit of avoiding a free parameter α, J_2 has a simpler optimization cost when the parameter α is fixed. By changing α, one can control the importance of the variance of the between-class to that of the within-class data distributions. The orthogonalization (4.4) ensures that the proposed discriminant is defined by orthonormal basis vectors in each local coordinate system. The orthonormalization of the bases yields more robust performance in the presence of estimation error (please refer to [33], [34] for the details). The benefits of orthonormal bases in discriminant analysis over the classical LDA have also been explained in these studies. Although we do not provide a proof of convergence or uniqueness of the gradient-based iterative learning method, its convergence to a global maximum can be expected by virtue of the criterion being a second-order convex function with respect to a basis vector, $\mathbf{u}_{kn}$, of each local coordinate system, and the joint set of the basis vectors $\mathbf{u}_{kn}, k = 1, \ldots, K$, as explained in [3], [17]. Fig. 3 shows the convergence characteristics of the learning process for the synthetic data presented in Fig. 1. The constant α was explored in steps of 0.1 and 0.1 was found to maximize the value of J_2. The value of J_2 according to the angles of basis vectors has a unique global maximum. It is also noted that the gradient optimization method of the objective function quickly converges regardless of constant α. The learning using the objective function J_1 also stably approaches a unique maximum.

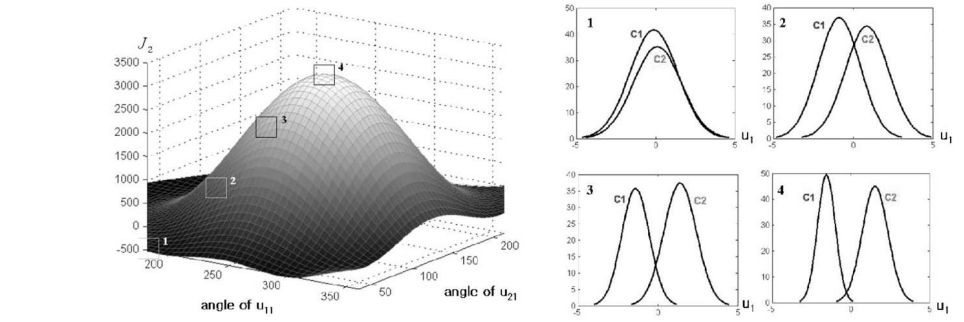
A solution to the constrained optimization problem can also be obtained by using the method of Lagrangian multipliers as

$$L = (1-\alpha)\big|\tilde{\mathbf{B}}\big| - \alpha\big|\tilde{\mathbf{W}}\big| - \sum_{k=1}^{K} \Lambda_k(\mathbf{U}_k^{\mathbf{T}}\mathbf{U}_k - \mathbf{I}), \tag{4.6}$$

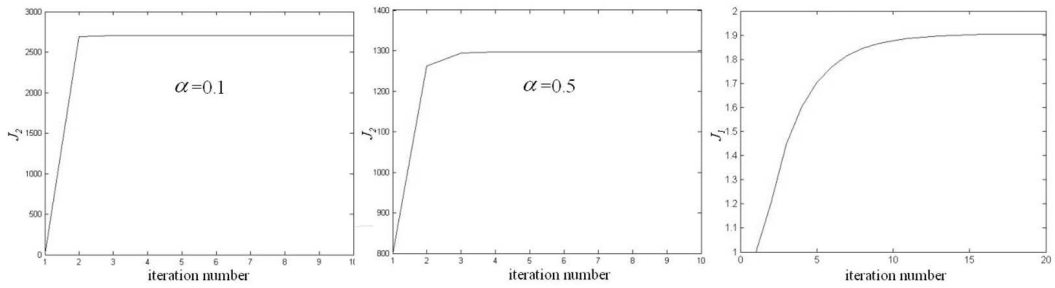
where $\mathbf{I}$ is the identity matrix and the diagonal matrix of eigen values is

$$\Lambda_k = \begin{bmatrix} \lambda_{k1} & & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & & \lambda_{kN} \end{bmatrix}.$$

The gradient of the Lagrangian function with respect to the basis vectors is



(a)



(b)

Fig. 3. An example of learning for the data distribution shown in Fig. 1, where K is set to 2 and step size η is fixed to 0.1. (a) Value of the criterion J_2 (left) as a function of orientation of $\mathbf{u}_{11}, \mathbf{u}_{21}$ with $\alpha = 0.1$. The distributions of the two classes $\mathbf{C1} = \mathbf{C}_{11}\mathrm{U}\mathbf{C}_{12}$, $\mathbf{C2} = \mathbf{C}_{21}\mathrm{U}\mathbf{C}_{22}$, which are in the space defined by the first major component $\mathbf{u}_1$,are drawn (right) as a series while J_2 is maximized. (b) Convergence graphs of J_2 with $\alpha = 0.1, 0.5$, and J_1.

$$\frac{\partial L}{\partial \mathbf{u}_{kn}} = (2(1-\alpha)\mathbf{B}_k - 2\alpha\mathbf{W}_k - 2\lambda_{kn}\mathbf{I})\mathbf{u}_{kn} + \sum_{i=1, i\neq k}^{K} (2(1-\alpha)\mathbf{B}_{ki} - 2\alpha\mathbf{W}_{ki})\mathbf{u}_{in} = 0. \quad (4.7)$$

The solution can be found by numerical optimization of the Lagrangian function. However, in practice, a numerical optimization can only be used in low-dimensional data spaces. As a reference, we utilized the numerical optimization "solve" function in Matlab to solve the two-dimensional problem shown in Fig. 1. The constraint optimization took 600 times longer than the gradient-based optimization of J_2. The two proposed methods of gradient-based learning are much favored for their efficiency.

5 LLDA with K-Means Clustering

Let us revisit the basic model derived in Section 3 by considering the special case involving a discrete posterior probability. K-means clustering divides a data set into disjoint subsets. If the data point $\mathbf{x}$ belongs to the k^*th cluster, $P(k*|\mathbf{x}) = 1$ and $P(k|\mathbf{x}) = 0$ for all the other ks. The mean vector of the kth cluster $\boldsymbol{\mu}_k$ in (3.2) can be rewritten by

$$\boldsymbol{\mu}_k = \left(\sum_{\mathbf{x}} P(k|\mathbf{x})\mathbf{x}\right) / \left(\sum_{\mathbf{x}} P(k|\mathbf{x})\right) = \left(\sum_{\mathbf{x}\in k} \mathbf{x}\right) / M'_k, \quad (5.1)$$

where M'_k is the sample number of the cluster k. The defined transformation in (3.1) becomes

$$\mathbf{y} = \mathbf{U}_k^{\mathbf{T}}(\mathbf{x} - \boldsymbol{\mu}_k) \text{ for } \mathbf{x} \in k. \quad (5.2)$$

The definition of the global mean (3.4) and the class mean (3.5) changes as follows:

$$\tilde{\mathbf{m}} = \frac{1}{M}\sum_{k=1}^{K} \mathbf{U}_k^{\mathbf{T}} \sum_{\mathbf{x}\in k} (\mathbf{x} - \boldsymbol{\mu}_k) = \vec{0}. \quad (5.3)$$

$$\tilde{\mathbf{m}}_c = \sum_{k=1}^{K} \mathbf{U}_k^{\mathbf{T}} \mathbf{m}_{ck}, \text{where } \mathbf{m}_{ck} = \frac{1}{M_c} \sum_{\mathbf{x}\in \mathrm{X}_c \cap k} (\mathbf{x} - \boldsymbol{\mu}_k).$$

The transformed between-class matrix (3.6) and the within-class scatter matrix (3.7) can similarly be expressed by changing the notation from $P(k|\mathbf{x})$ to $\mathbf{x} \in k$. The learning algorithm in Section 4 finds the optimal set of locally linear transformation $\mathbf{U}_k,\ k = 1, .., K$.

When a new pattern $\mathbf{x}_{test}$ is presented, it is first assigned to one of the clusters by

$$\mathbf{x}_{test} \in k* = \min_{\arg\ k} \|\mathbf{x}_{test} - \boldsymbol{\mu}_k\| \quad (5.4)$$

and transformed by using the corresponding function.

$$\mathbf{y}_{test} = \mathbf{U}_{k*}^{\mathbf{T}}(\mathbf{x}_{test} - \boldsymbol{\mu}_{k*}). \quad (5.5)$$

6 Computational Complexity

The complexity of the algorithms depends on the computational costs associated with extracting the features and with matching. For the linear subspace methods such as PCA and LDA, the cost of feature extraction is determined by the dimensionality N of the input vector, $\mathbf{x}$, and the number of components of the subspace S. The cost of extracting features using linear methods is approximately proportional to $N \times S$.

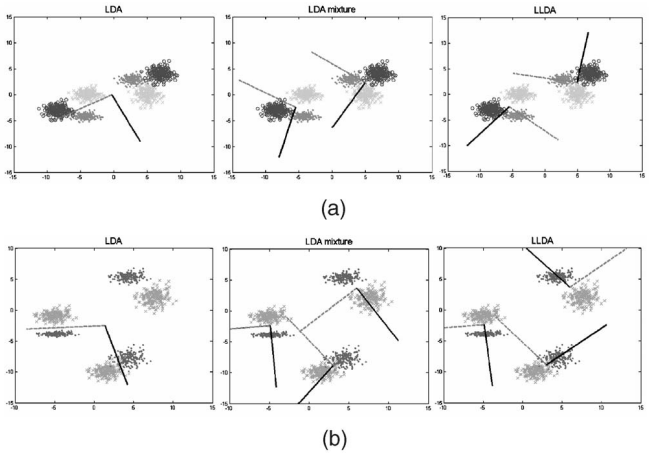


Fig. 4. Simulated data distributions and the components found. Black solid lines represent the first major components and gray dashed lines the second components. (a) For Set 1. (b) For Set 2.

In nonlinear subspace methods like the GDA, the nth component of the projection of vector $\mathbf{x}$ is computed as

$$y_n = \sum_{i=1}^{M} \alpha_{ni} k(\mathbf{x}_i, \mathbf{x}), \tag{6.1}$$

where M is the total number of training patterns, α_{ni} is a real weight, and k denotes a kernel function. The cost of extracting features of the GDA is about $N \times S \times M$. The proposed method, LLDA, has a similar cost as that of PCA or LDA, depending on the preceding clustering algorithm. When a hard clustering such as K-means is applied, the cost of extracting features is $N \times (S + K)$, where the additional term $N \times K$ is for assigning a cluster to the input. When a soft clustering is applied, the cost is multipled by the number of clusters, i.e., $N \times S \times K$. Note that, usually, $K << M$.

When the data points are represented as the S-dimensional feature vectors and C gallery samples are given for the C class categories, the matching cost for recognition is $C \times S$. This applies to all, the linear, nonlinear, and the proposed subspace methods.

7 Experiments

7.1 Results on Synthetic Data

Two sets of two-dimensional synthetic data were experimented with. Set 1 has three classes which have two distinct modes in their distributions generated, respectively, by

$$\mathbf{X}_1 = \{X \sim N(7, 0.9), Y \sim N(4.1, 0.8)\}\mathrm{U}\{X \sim N(-8.4, 0.9), Y \sim N(-3, 0.7)\},$$

$$\mathbf{X}_2 = \{X \sim N(5, 0.9), Y \sim N(0.1, 1)\}\mathrm{U}\{X \sim N(-4, 0.9), Y \sim N(0.1, 0.6)\},$$

$$\mathbf{X}_3 = \{X \sim N(2.9, 0.9), Y \sim N(2.9, 0.5)\}\mathrm{U}\{X \sim N(-4.2, 0.9), Y \sim N(-4.2, 0.4)\},$$

where $N(a, b)$ is a normal variable which has a mean a and standard deviation b. Two hundred data points were drawn from each Gaussian mode. Set 2 has two classes which have three distinct peaks in the distributions generated by

$$\mathbf{X}_1 = \{X \sim N(4.4, 1), Y \sim N(5.4, 0.5)\}\mathrm{U}\{X \sim N(-4.7, 1), Y \sim N(-3.9, 0.2)\}\mathrm{U}\{X \sim N(4.4, 1), Y \sim N(-7.8, 0.8)\}$$

and

$$\mathbf{X}_2 = \{X \sim N(7.6, 1), Y \sim N(2.1, 0.9)\}\mathrm{U}\{X \sim N(-5, 1), Y \sim N(-0.9, 0.6)\}\mathrm{U}\{X \sim N(1.6, 1), Y \sim N(-9.9, 0.7)\}.$$

TABLE 1
Classification Results (Number of Errors)

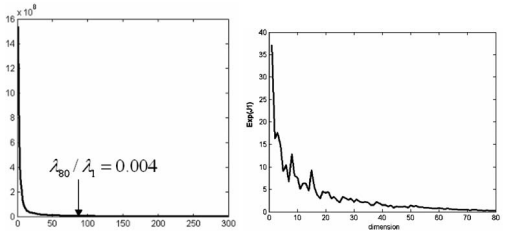
	E.D.	N.C.	M.D.	Cost
Set 1 (400 samples / class)				
LDA	266±115	266±115	81±61	1
LDA mixture	254±27	255±23	169±45	$1+\omega$
GDA	4.3±1.1	4.3±1.1	4.4±0.5	270
LLDA J_1+km	7.6±3.5	7.6±3.5	7±3.4	$1+\omega$
LLDA J_2+km	7.6±3.5	8±3.6	7.3±3.7	$1+\omega$
LLDA J_1+GMM	7.6±3.5	8±3.6	7.3±3.7	$2+\omega$
Lagran. J_2	7.6±3.2	8±2.6	7.3±2.8	$1+\omega$
Set 2 (600 samples / class)				
LDA	308±129	308±129	207±272	1
LDA mixture	205±1.4	205±1.4	206±7	$1+\omega$
GDA	4±1.4	4±1.4	4±0	278
LLDA J_1+km	9.5±3.5	9.5±3.5	7.5±3.5	$1+\omega$
LLDA J_2+km	8±1.4	8±1.4	7±2.8	$1+\omega$

ω indicates the computational cost of deciding which cluster a new pattern belongs to. It is usually less than 1. "LLDA $J_1 + \mathrm{km}$" is the LLDA of the objective function J_1 with K-means clustering algorithm. "LLDA $J_1 + \mathrm{GMM}$" indicates the LLDA of the objective function J_1 with Gaussian mixture modeling. "Lagrangian J_2" denotes a numerical solution of the Lagrangian formulation.

Conventional LDA, mixture of LDA, and GDA with the radial basis function (RBF) as a kernel are compared with LLDA in terms of classification error. Euclidean distance (E.D.), normalized correlation (N.C.), and Mahalanobis distance (M.D.) were utilized as similarity functions for the nearest neighbor (N.N.) classification. It is noted that all the transformed data points were compared with the sample mean of each class (3.5).

In the method of LLDA, the number of clusters, K, was selected to maximize the value of the objective function. For the example of the data of Set 1, the peak values of J_1 changed with K as follows: -7.14, 2.97, 0.85 for $K = 1, 2, 3$, respectively, so the number $K = 2$ was chosen. This is much simpler than the parameter selection of RBF as a kernel function in GDA because the standard deviation of RBF is hard to initialize and it is a real (noninteger) value. The axes of LDA, LDA mixture, LLDA are drawn in Fig. 4. Table 1 shows the average number of classification errors with their standard deviation and the relative costs of feature extraction. It is apparent that the proposed discriminant can well solve the nonlinear classification problem on which the conventional linear methods fail and it is much more profitable in terms of computational efficiency as compared to GDA. The feature extraction complexity of the proposed method is about 1/270 of that of GDA in this example. Although the accuracy of GDA was slightly better, it is noted that the kernel parameter of RBF in GDA was exhaustively searched to find the best performance for the given data. In contrast, the proposed algorithm based on the log objective function has only a small integer K to be adjusted and the

Fig. 5. Some normalized data samples. The leftmost image is the gallery image.



(a) (b)

Fig. 6. (a) Eigenvalues of the face data. (b) Plot of J1 as a function of dimensionality.

learning process is also much faster. Additionally, note that, when the class distributions have a single mode, LLDA with $K = 1$ yields a successful separation by behaving like the conventional LDA. LLDA with $K = 1$ is identical to the conventional LDA with the exception of the orthonormal constraint imposed on the axes by LLDA.

7.2 View-Invariant Face Recognition with One Sample Image

The proposed algorithm has been validated on the problem of free pose face recognition in the scenario when only a single frontal image of each class is available as a gallery image. To recognize a novel view face, some prior experiences of face view changes are required. Conventional discriminative subspace methods such as LDA and GDA can be applied to learn a robust representation from any prototype face set which exhibits different poses. GDA has a benefit of capturing any nonlinear manifolds of face pose changes. Then, the learned subspace representation can be applied to new test identities. In contrast, SVM, which performs binary classification and requires a considerable number of training samples for each class, is completely inappropriate for this scenario.

There are a number of conventional techniques that have been developed for view-invariant face recognition [4], [6], [10], [12], [13], [15], [25], [26], [28]. In spite of the successes of some approaches [6], [10], [12], [13], [26], they have an important drawback of requiring dense correspondences of facial features for image normalization or more than one model image. The step of correspondence solving or detection of abundant salient facial features, which is needed for separating the shape and texture components of face images in these methods, is usually difficult itself. Errors in correspondences seriously degrade the performance of the subsequent recognition methods, as shown in [12]. In our experiments, the proposed algorithm, LLDA, is compared with PCA, LDA, and GDA as the benchmark subspace methods that have been successfully applied to face recognition in the past and FaceIt (v.5.0), the commercial face recognition system from Identix. FaceIt ranked top overall in the Face Recognition Vendor Test 2000 and 2002 [27], [32].

TABLE 2
Face Recognition Rates (%)

	PCA		LDA		GDA		LLDA		FaceIt	
	Ev	Te	Ev	Te	Ev	Te	Ev	Te	Ev	Te
R1	13	4	55	43	66	49	66	56	73	**64**
L1	8	8	55	45	77	57	73	**64**	66	52
U1	28	16	53	43	73	52	71	**66**	46	36
D1	33	29	68	55	84	**66**	75	60	37	24
F2	75	70	73	63	82	71	75	66	95	**83**
R2	8	3	42	22	46	29	40	35	46	**36**
L2	4	4	33	27	44	36	48	**47**	46	30
U2	17	15	28	28	35	35	40	**44**	24	23
D2	20	10	31	32	42	32	35	**40**	33	9
avg	23	18	49	40	61	47	58	**53**	51	39

Database. We used the XM2VTS data set annotated with pose labels of the face. The face database consists of 2,950 facial images of 295 people with five pose variations and two different time sessions which have five months time elapse. The data set consists of five different pose groups (F,R,L,U,D) which are captured at frontal view, about ± 30 horizontal rotations and ± 20 vertical rotations. The two images of a pose group "F" captured at different times are denoted by F1 and F2. This may be the largest public database that contains images of faces taken from different view points. The images were normalized to 46*56 pixel resolution with a fixed eye position and some normalized data samples are shown in Fig. 5. The face set is partitioned into the three subsets: 1,250 images of 125 people, 450 images of 45 people, and 1,250 face images of 125 people for the training (Tr), evaluation (Ev), and test (Te), respectively. Please note that the three sets have different face identities. For the test of the commercial FaceIt system, the original images were applied to the system with the manual eye positions.

Protocol and Setting. The training set is utilized to learn the subspace representation of the conventional PCA/LDA/GDA methods and LLDA with K-means. For efficiency of learning, all of the algorithms were applied to the first 80 eigenfeatures of the face images. Fig. 6 shows the plots of eigenvalues and J_1 of LLDA as a function of dimensionality. The evaluation set is utilized to adjust the kernel parameter of GDA (an RBF kernel with an adjustable width) and the dimensionality of the output vectors for all the methods. The parameters are properly quantized and all combinations of the discrete values of the quantized parameters are examined to get the best recognition rate on the evaluation set. In LLDA, the number of clusters corresponded to the number of the pose groups and K-means algorithm was applied. The log objective function J_1 was utilized to learn the set of transformation functions and the learning rate was controlled to have faster convergence. Typically, the learning took 2 or 3 minutes in Pentium IV 2GHz PC.

In the test, the frontal face images of the test set, which are the leftmost images in Fig. 5, are registered as a gallery and all the other images of the test set are exploited as queries. All the test images are projected into the learned subspace and Nearest-Neighbor-based classification is performed based on the projection coefficients. Recognition rates in (percent) are measured. In LLDA, test face images were assigned to one of the clusters by (5.4) and projected into the corresponding subspace by (5.5).

Results. Table 2 presents the recognition rates on the evaluation and test set and Fig. 7 shows the performance

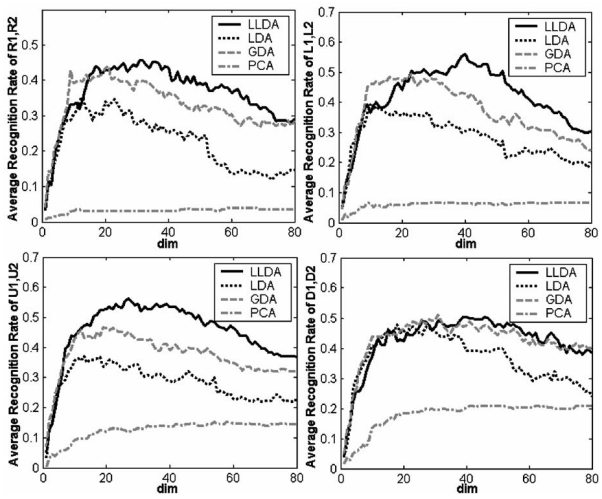


Fig. 7. The test performance curves as a function of dimensionality.

curves of the test set as a function of dimensionality. The recognition rate of the evaluation and test set was much enhanced by the proposed algorithm. FaceIt exhibited the best recognition performance for the frontal images F2, but quite low recognition rates for the rotated faces especially involving up/down rotations. More results showing the effects of the elapsed time and the size of test population are given in Fig. 8.

In LLDA, the number of clusters was chosen as the number of the pose groups, as previously mentioned, by assuming that the multimodality of the face class distributions is caused by the different poses. In each cluster, classes are assumed to be linearly separable. Although this assumption is not necessarily true, as other factors such as time elapse can make a class distributed multimodally and not linearly separable, we found that LLDA performed much better as compared with LDA/GDA/FaceIt. A performance degradation as a function of time was observed for all the methods, but a relative performance gain exhibited by LLDA was still preserved, as shown in Fig. 8. As mentioned above, the results of the test set were obtained by utilizing the output dimensionality found to be the best for the evaluation set. The establishment of a proper evaluation set is important because the test results are sensitive to the output dimensionality, as shown in Fig. 7. This may be because the pose variation is so large that the methods find only a few meaningful axes. We can see that the evaluation set used has proven adequate for solving this peaking problem as the recognition results on the test set using the best dimensionality indicated by the evaluation set in Table 2 agreed with the best results of the graph in Fig. 7. GDA had the tendency to highly overfit on the training set so that a separate evaluation set was needed to suppress this behavior.

Regarding the complexity of the feature extraction, PCA, LDA, and the LLDA are approximately identical and GDA about 40 times worse than the linear methods. Please note that the complexity of GDA depends on the size of the training set. The proposed method is not expensive in terms of computational costs and provides more robust and accurate performance for all the dimensionalities as compared with the other methods.

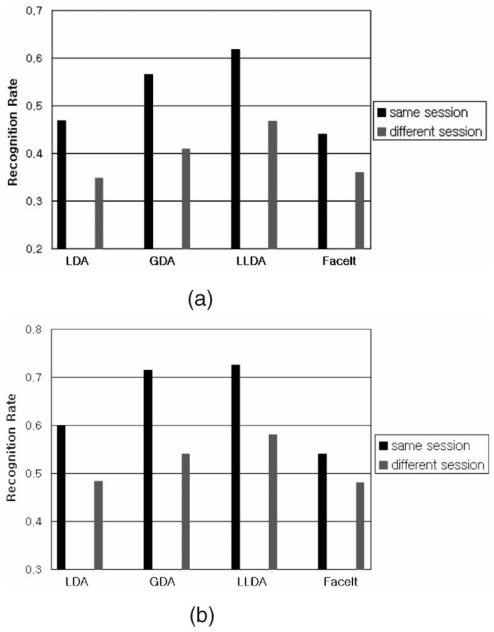


Fig. 8. Recognition rates under aging for different sizes of test population. (a) Recognition rates on the test set consisting of 125 identities. (b) Recognition rates on the test set consisting of randomly chosen 50 identities.

8 Conclusion

A novel discriminant analysis method which can classify a nonlinear structure has been proposed for face recognition. A face data set that exhibits large pose variations has nonlinear manifolds and is not linearly separable. A set of local linear transformations is found so that the locally linearly transformed classes maximize the between-class covariance and minimize the within-class covariance in a single global space. The proposed learning method for finding the optimal set of locally linear bases does not suffer from the local-maxima problem and stably converges to a global maximum point. The proposed discriminant provides a set of discriminant features for the view-invariant face recognition with a given single model image and it is highly efficient computationally as compared with the nonlinear discriminant analysis based on the kernel approach. By virtue of the linear base structure of the solution, the method reduces overfitting. We intend to improve the performance of the proposed approach by exploiting more facial feature correspondences for an image regularization step in the future. The current performance was obtained with the images registered with a fixed eye position and this can be seen as a poor basis of the image normalization for the method. More elaborate regularization is expected to promote face class structures are well separated by a set of local linear transformations, similar to having the results of [4].

Acknowledgments

The authors would like to thank Hyun-Chul Kim at Pohang University of Science and Technology for his helpful discussions and comments and Seok-Cheol Kee at Samsung AIT for the project support. Thanks also to the anonymous reviewers for their constructive comments.

REFERENCES

[1] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, 2001.

[2] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," *Neural Computation,* vol. 12, pp. 2385-2404, 2000.

[3] D.D. Lee and H. Sebastian Seung, "Algorithms for Non-Negative Matrix Factorization," *Advances in Neural Information Processing Systems,* vol. 13, pp. 556-562, 2001.

[4] R. Gross, I. Matthews, and S. Baker, "Appearance-Based Face Recognition and Light-Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 4, pp. 449-465, Apr. 2004.

[5] S.T. Roweis and L.K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, pp. 2323-2326, 2000.

[6] T. Vetter and T. Poggio, "Linear Object Classes and Image Synthesis From a Single Example Image," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 733-742, July 1997.

[7] H.-C. Kim, D. Kim, and S.-Y. Bang, "Face Recognition Using LDA Mixture Model," *Proc. Int'l Conf. Pattern Recognition,* vol. 2, pp. 486-489, 2002.

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

[9] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Varialbe Lighting and Pose," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 6, pp. 643-660, June 2001.

[10] K. Okada and C. von der Malsburg, "Analysis and Synthesis of Human Faces with Pose Variations by a Parametric Piecewise Linear Subspace Method," *Proc. Computer Vision and Pattern Recognition,* pp. 761-768, 2001.

[11] X. He, S. Yan, Y. Hu, and H. Zhang, "Learning a Locality Preserving Subspace for Visual Recognition," *Proc. Int'l Conf. Computer Vision,* pp. 385-392, 2003.

[12] V. Blanz, S. Romdhani, and T. Vetter, "Face Identification Across Different Poses and Illuminations with a 3D Morphable Model," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 192-197, May 2002.

[13] Y. Li, S. Gong, and H. Liddell, "Constructing Facial Identity Surfaces in a Nonlinear Discriminating Space," *Proc. Computer Vision and Pattern Recognition,* vol. 2, pp. 258-263, 2001.

[14] Q. Liu, R. Huang, H. Lu, and S. Ma, "Face Recognition Using Kernel-Based Fisher Discriminant Analysis," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 205-211, 2002.

[15] D.B. Graham and N.M. Allinson, "Automatic Face Representation and Classification," *Proc. British Machine Vision Conf.,* pp. 64-73, 1998.

[16] M.E. Tipping and C.M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation,* vol. 11, pp. 443-482, 1999.

[17] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature,* vol. 401, pp. 788-791, 1999.

[18] T.-K. Kim, H. Kim, W. Hwang, S.C. Kee, and J.H. Lee, "Component-Based LDA Face Descriptor for Image Retrieval," *Proc. British Machine Vision Conf,* pp. 507-526, 2002.

[19] T.-K. Kim, J. Kittler, H.-C. Kim, and S.-C. Kee, "Discriminant Analysis by Multiple Locally Linear Transformations," *Proc. British Machine Vision Conf.,* pp. 123-132, 2003.

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

[21] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical Performance Analysis of Linear Discriminant Classifiers," *Proc. Computer Vision and Pattern Recognition,* pp. 164-169, June 1998.

[22] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "Fisher Discriminant Analysis with Kernels," *Proc. IEEE Workshop Neural Networks for Signal Processing,* pp. 41-48, 1999.

[23] S. Gong, S. McKenna, and J. Collins, "An Investigation into Face Pose Distributions," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 265-270, Oct. 1996.

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

[25] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. Computer Vision and Pattern Recognition,* pp. 84-91, 1994.

[26] B. Heisele, P. Ho, and T. Poggio, "Face Recognition with Support Vector Machines: Global versus Component-Based Approach," *Proc. Int'l Conf. Computer Vision,* vol. 2, pp. 688-694, 2001.

[27] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone, "FRVT 2002: Evaluation Report," Mar. 2003, http://www.frvt.org/FRVT2002/.

[28] T.-K. Kim, H. Kim, W. Hwang, S.-C. Kee, and J. Kittler, "Independent Component Analysis in a Facial Local Residue Space," *Proc. Computer Vision and Pattern Recognition,* vol. 1, pp. 579-586, 2003.

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

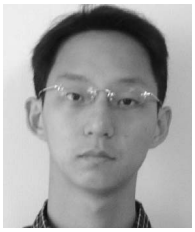
[30] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," *Proc. Computer Vision and Pattern Recognition,* pp. 130-136, June 1997.

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

[32] D.M. Blackburn, M. Bone, and P.J. Phillips, "Facial Recognition Vendor Test 2000: Evaluation Report," 2000.

[33] T. Okada and S. Tomita, "An Optimal Orthonormal System for Discriminant Analysis," *J. Pattern Recognition,* vol. 18, pp. 139-144, 1985.

[34] W. Zhao, "Discriminant Component Analysis for Face Recognition," *Proc. Int'l Conf. Pattern Recognition,* vol. 2, pp. 818-821, 2000.



Tae-Kyun Kim received the BSc and MSc degrees from the Department of Electrical Engineering and Computer Science at the Korea Advanced Institute of Science and Technology (KAIST) in 1998 and 2000, respectively. He has worked as a research staff member at Samsung Advanced Institute of Technology, Korea, since 2000. This is also his period of obligatory military service. In January 2005, he will begin working on the PhD degree at the University of Cambridge. His research interests include computer vision, statistical pattern classification, and machine learning. He reviews for the *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* and is the Korea delegate of MPEG-7. The joint proposal of face descriptor of Samsung and NEC, for which he developed the main algorithms, was accepted as the international standard of ISO/IEC JTC1/SC29/WG11. He is a member of the IEEE.

Josef Kittler is a professor of machine intelligence and director of the Centre for Vision, Speech, and Signal Processing at the University of Surrey. He has worked on various theoretical aspects of pattern recognition and image analysis and on many applications, including personal identity authentication, automatic inspection, target detection, detection of microcalcifications in digital mammograms, video coding and retrieval, remote sensing, robot vision, speech recognition, and document processing. He coauthored the book *Pattern Recognition: A Statistical Approach* (Prentice Hall) and has published more than 500 papers. He is a member of the editorial boards of *Image and Vision Computing*, *Pattern Recognition Letters*, *Pattern Recognition and Artificial Intelligence*, *Pattern Analysis and Applications*, and *Machine Vision and Applications*. He is a member of the IEEE.