Discriminant Analysis by Locally Linear Transformations

Tae-Kyun Kim[1,2], Josef Kittler[2], Hyun-Chul Kim[3], and Seok Cheol Kee[1]

[1] : Samsung Advanced Institute of Technology, KOREA

[2] : Center for Vision, Speech and Signal Processing, University of Surrey,U.K.

[3] : Pohang University of Science and Technology, KOREA

taekyun@sait.samsung.co.kr

Abstract

We present a novel discriminant analysis learning method which is applicable to non-linear data structures. The method can deal with pattern classification problems which have a multi-modal distribution for each class and samples of other classes may be closer to a class than those of the class itself. Conventional linear discriminant analysis (LDA) and LDA mixture model can not solve this linearly non-separable problem. Several local linear transformations are considered to yield locally transformed classes that maximize the between-class covariance and minimize the within-class covariance. The method invloves a novel gradient based algorithm for finding the optimal set of local linear bases. It does not have a local-maxima problem and stably converges to the global maximum point. The method is computationally efficienct as compared to the previous non-linear discriminant analysis based on the kernel approach. The method does not suffer from an overfitting problem by virtue of the linear base structure of the solution. The classification results are given for both simulated data and real face data.

1   Introduction

Pattern classification methods have suffered from various factors which dramatically affect sensory information about an object. It often happens that a single object is multi-modally distributed and samples of other objects are more closely located to the object in the original data space than those of the same clan. Efficient feature extraction is needed when extracting discriminative features under large changes of input data and involving dimension reduction of high dimensional input data like an image. Efficient classifiers associated with the extracted features are also needed for successful classification, considering both the computational cost as well as accuracy.

Linear discriminant analysis (LDA) is an effective representation method that linearly transforms the original data space into a low dimensional feature space where the data is well separated in terms of 2nd order statistics [8]. However, this method fails to solve non-linear problems as illustrated in Figure 1 (a). In many conventional recognition systems which adopt a linear machine like LDA, many gallery samples which consist of at least one sample per one local group can be registered to enhance

recognition. The LDA mixture model [7] which considers the transformation of several local frames independently also fails to separate multi-modally disstributed classes because it does not encode the relationship of local LDAs. This is shown in Figure 1 (b). In this paper, several locally linear transformations are concurrently sought for so that the class structures manifest us the locally transformed data are well separated. The objective function for this problem has a similar form to that of classical LDA,which is to maximize the between-class scatter minimizing the within-class scatter in the locally transformed space. The main idea is to decompose a non-linear classification problem into a set of locally linear ones as illustrated in Figure 1 (c). It was proven in [5] that a non-linear data structure can be represented by a locally linear structure. The discriminant based on such a piecewise linear structure has the benefit of optimising a convex function with respect to a set of basis vectors of the local frames having a unique maximum. Compared with the generalized discriminant analysis (GDA) [2] whereby the original data is mapped into a high-dimensional feature space with a kernel function, the

(a) LDA

(b) LDA mixture model
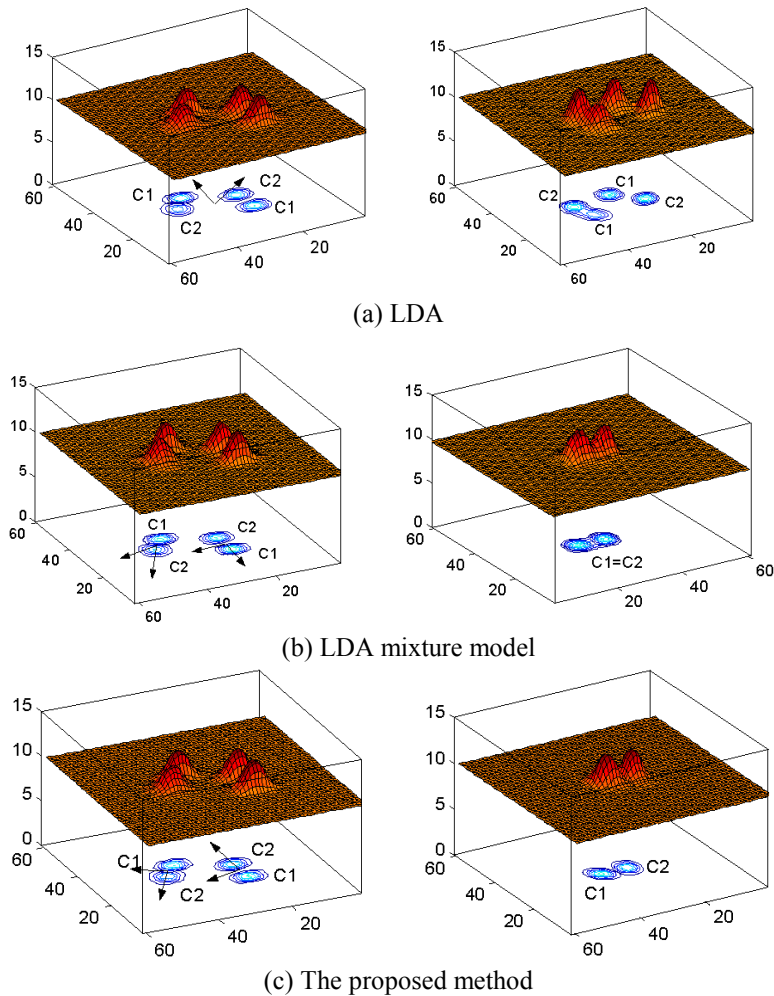

(c) The proposed method

Figure 1: Comparisons for the non-linear classification problem. Left pictures show the original class distribution and components. Transformed class distributions for the components are drawn in right pictures.

proposed method is much more computationally efficient because it only involves linear transformations. The importance of efficiency of feature extraction and matching has been increased for classification of large data sets. The proposed method also reduces overfitting normally exhibited by conventional non-linear methods by virtue of a linear base structure.

Classification results are given for both simulated data and real face data. A large pose change of a face is considered to make a bigger difference between two images of the same faces face images than that of different faces with the same view. To recognize a face taken from a new view, the view-based approach [6] has been adopted. The discriminant is verified for the problem of novel-view face classification considering the face view space as the local space of the proposed method.

The paper is organized as follows: The next section introduces the problem formulation for the case of two local frames. Section 3 develops a solution to the mapping optimisation problem starting from the case of two local frames and then generalizing to multiple frames. One-basis vector algorithm is presented first, followed by a multiple basis vector solution. In section 4 the gradient method is further elaborated. The last section is devoted to the experiments with simulated and real face data.

2   Problem Formulation

Let $\mathbf{x}$ be a data vector which is an element of a subset $\mathbf{C}_i$ of the set of input vectors $\mathbf{X}$. $\mathbf{C}_i$ denots a class and $\mathbf{N_c}$ is the number of classes. The input vectors are also divided into several subsets $\mathbf{L}_i$. Each subset represets a local group which has a different transformation function. Let the number of local groups $\mathbf{N_L}$ be two initially. That is $\mathbf{X} = \bigcup_{i=1}^{N_c} \mathbf{C}_i = \bigcup_{i=1}^{N_L} \mathbf{L}_i$. The local group can be defined in various ways. Any clustering or mixture modeling of the input vectors can be applied to define the group of neighboring data vectors . For simplicity, the data vector $\mathbf{x}$ is now considered as a zero-mean vector such that $E\{\mathbf{x} \,|\, \mathbf{x} \in \mathbf{L}_i\} = 0,\ for\ \mathbf{x} \in \mathbf{L}_i$. A global mean vector $\mathbf{m}$ is defined as

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}}\mathbf{x} = \frac{1}{n}\left(\sum_{\mathbf{x} \in \mathbf{L}_1}\mathbf{x} + \sum_{\mathbf{x} \in \mathbf{L}_2}\mathbf{x}\right) = \mathbf{m}_{\mathbf{L}_1} + \mathbf{m}_{\mathbf{L}_2} \tag{2.1}$$

where n is the total number of the input vectors. A mean vector of a class i which consists of n_i samples is given by

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x} \in \mathbf{C}_i}\mathbf{x} = \frac{1}{n_i}\left(\sum_{\mathbf{x} \in \mathbf{C}_i \cap \mathbf{L}_1}\mathbf{x} + \sum_{\mathbf{x} \in \mathbf{C}_i \cap \mathbf{L}_2}\mathbf{x}\right) = \mathbf{m}_{i,\mathbf{L}_1} + \mathbf{m}_{i,\mathbf{L}_2}. \tag{2.2}$$

The between-class scatter is then represented as follows:

$$\begin{aligned}
\mathbf{S_B} &= \sum_{i=1}^{N_c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^{\mathbf{T}} \\
&= \sum_{i=1}^{N_c} n_i(\mathbf{m}_{i,\mathbf{L}_1} - \mathbf{m}_{\mathbf{L}_1})(\mathbf{m}_{i,\mathbf{L}_1} - \mathbf{m}_{\mathbf{L}_1})^{\mathbf{T}} + \sum_{i=1}^{N_c} n_i(\mathbf{m}_{i,\mathbf{L}_2} - \mathbf{m}_{\mathbf{L}_2})(\mathbf{m}_{i,\mathbf{L}_2} - \mathbf{m}_{\mathbf{L}_2})^{\mathbf{T}} + \\
&\quad \sum_{i=1}^{N_c} n_i(\mathbf{m}_{i,\mathbf{L}_1} - \mathbf{m}_{\mathbf{L}_1})(\mathbf{m}_{i,\mathbf{L}_2} - \mathbf{m}_{\mathbf{L}_2})^{\mathbf{T}} + \sum_{i=1}^{N_c} n_i(\mathbf{m}_{i,\mathbf{L}_2} - \mathbf{m}_{\mathbf{L}_2})(\mathbf{m}_{i,\mathbf{L}_1} - \mathbf{m}_{\mathbf{L}_1})^{\mathbf{T}} \\
&= \mathbf{S_{B,L_1}} + \mathbf{S_{B,L_2}} + \mathbf{R_B} + \mathbf{R_B}^{\mathbf{T}}
\end{aligned} \tag{2.3}$$

Similarly, the within-class scatter is defined by

$$S_W = \sum_{i=1}^{N_c} \sum_{x \in C_i} (x - m_i)(x - m_i)^T$$

$$= \sum_{i=1}^{N_c} \sum_{x \in C_i \cap L_1} (x - m_i)(x - m_i)^T + \sum_{i=1}^{N_c} \sum_{x \in C_i \cap L_2} (x - m_i)(x - m_i)^T$$

$$= \sum_{i=1}^{N_c} \left(\sum_{x \in C_i \cap L_1} (x - m_{i,L_1})(x - m_{i,L_1})^T + \sum_{x \in C_i \cap L_2} m_{i,L_1} m_{i,L_1}{}^T \right) \tag{2.4}$$

$$+ \sum_{i=1}^{N_c} \sum_{x \in C_i \cap L_1} \left(-(x - m_{i,L_1}) m_{i,L_2}{}^T - m_{i,L_2} (x - m_{i,L_1})^T \right)$$

$$+ \sum_{i=1}^{N_c} \left(\sum_{x \in C_i \cap L_2} (x - m_{i,L_2})(x - m_{i,L_2})^T + \sum_{x \in C_i \cap L_1} m_{i,L_2} m_{i,L_2}{}^T \right)$$

$$+ \sum_{i=1}^{N_c} \sum_{x \in C_i \cap L_2} \left(-(x - m_{i,L_2}) m_{i,L_1}{}^T - m_{i,L_1} (x - m_{i,L_2})^T \right)$$

$$= S_{W,L_1} + (R_{W,12} + R_{W,12}^T) + S_{W,L_2} + (R_{W,21} + R_{W,21}^T)$$

We define the locally linear transformation $W_i = [w_{i1}, ..., w_{in}]$, $i = 1, ..., N_L$ such that

$$y_1 = W_1^T x \quad for\ x \in L_1$$
$$y_2 = W_2^T x \quad for\ x \in L_2 \tag{2.5}$$

to maximize the between-class variance and minimize the within-class variance in the locally transformed data space. The objective function to be maximized is

$$J = tr\widetilde{S}_B - k \cdot tr\widetilde{S}_w \tag{2.6}$$

, where $\widetilde{S}_B$ and $\widetilde{S}_w$ are the transformed versions of S_B and S_w respectively. k is a constant which can be adjusted. This criterion which is based on the between-class scatter and the within-class scatter is conceptually similar to that of the conventional LDA. This kind of criterion helps to find the solution for the distance based separation problem in terms of 2nd order statistics. The locally linear transformation matrices W_1 and W_2 are found so as to maximize the criterion funciton, J.

3   Gradient based Learning Algorithm

3.1   The Case of two local frames
3.1.1   One-basis Algorithm

The solution of the above equation (2.6) may explicitly be obtained by using a Lagrangian formulation and some basic calculus. Even if the solution exists, representation of the solution in a vector is not simple. In any case, even for problems which have a closed from solution, frequently an iterative solution is perfomed from the programming point of view. For the some reason, we shall adopt an iterative optimisation approach to find a solution of (2.6). The most appropriate candidate is a gradient based learning algorithm. The gradient method has a global maximum by virtue of 2nd-order convex criterion function with respect to both variables w_{11} and w_{21}. The algorithm can be derived as follows.

The transformed global mean vector: $\widetilde{\mathbf{m}} = \mathbf{w}_{11}^t \mathbf{m}_{\mathbf{L_1}} + \mathbf{w}_{21}^t \mathbf{m}_{\mathbf{L_2}}$ (3.1)

The mean vector of the transformed class i : $\widetilde{\mathbf{m}}_i = \mathbf{w}_{11}^t \mathbf{m}_{i,\mathbf{L_1}} + \mathbf{w}_{21}^t \mathbf{m}_{i,\mathbf{L_2}}$ (3.2)

The transformed between-class scatter matrix which represents the between-class variance is given by

$$\widetilde{\mathbf{S}}_{\mathbf{B}} = \sum_{i=1}^{N_c} n_i \mathbf{w}_{11}^t (\mathbf{m}_{i,\mathbf{L_1}} - \mathbf{m}_{\mathbf{L_1}})(\mathbf{m}_{i,\mathbf{L_1}} - \mathbf{m}_{\mathbf{L_1}})^{\mathbf{T}} \mathbf{w}_{11} + \sum_{i=1}^{N_c} n_i \mathbf{w}_{21}^t (\mathbf{m}_{i,\mathbf{L_2}} - \mathbf{m}_{\mathbf{L_2}})(\mathbf{m}_{i,\mathbf{L_2}} - \mathbf{m}_{\mathbf{L_2}})^{\mathbf{T}} \mathbf{w}_{21} +$$

$$\sum_{i=1}^{N_c} n_i 2\mathbf{w}_{11}^t (\mathbf{m}_{i,\mathbf{L_1}} - \mathbf{m}_{\mathbf{L_1}})(\mathbf{m}_{i,\mathbf{L_2}} - \mathbf{m}_{\mathbf{L_2}})^{\mathbf{T}} \mathbf{w}_{21}$$

$$= \mathbf{w}_{11}^t \mathbf{S}_{\mathbf{B},\mathbf{L_1}} \mathbf{w}_{11} + \mathbf{w}_{21}^t \mathbf{S}_{\mathbf{B},\mathbf{L_2}} \mathbf{w}_{21} + 2\mathbf{w}_{11}^t \mathbf{R}_{\mathbf{B}} \mathbf{w}_{21}$$

(3.3)

Simiarly, the within-class scatter matrix is transformed such that

$$\widetilde{\mathbf{S}}_{\mathbf{W}} = \mathbf{w}_{11}^t \mathbf{S}_{\mathbf{W},\mathbf{L_1}} \mathbf{w}_{11} + \mathbf{w}_{21}^t \mathbf{S}_{\mathbf{W},\mathbf{L_2}} \mathbf{w}_{21} + 2\mathbf{w}_{11}^t \mathbf{R}_{\mathbf{W},\mathbf{12}} \mathbf{w}_{21} + 2\mathbf{w}_{21}^t \mathbf{R}_{\mathbf{W},\mathbf{21}} \mathbf{w}_{11} \quad (3.4)$$

We seek the vectors $\mathbf{w}_{11}, \mathbf{w}_{21}$ which make the criterion function to be maximized under the constraint of unit norm vectors. This contstrained optimization problem is solved by the method of projections on the constraint set [1]. The learning rules are as follows :

$$\text{Max } J = \widetilde{\mathbf{S}}_B - k\widetilde{\mathbf{S}}_w, \text{ for } \|\mathbf{w}_{11}\| = 1, \ \|\mathbf{w}_{21}\| = 1$$

$$\frac{\partial J}{\partial \mathbf{w}_{11}} = \left(2\mathbf{S}_{B,L_1} - 2k\mathbf{S}_{W,L_1}\right)\mathbf{w}_{11} + \left(2\mathbf{R}_B - 2k\mathbf{R}_{W,12} - 2k\mathbf{R}_{W,21}^T\right)\mathbf{w}_{21}$$

$$\frac{\partial J}{\partial \mathbf{w}_{21}} = \left(2\mathbf{R}_B^T - 2k\mathbf{R}_{W,12}^T - 2k\mathbf{R}_{W,21}\right)\mathbf{w}_{11} + \left(2\mathbf{S}_{B,L_2} - 2k\mathbf{S}_{W,L_2}\right)\mathbf{w}_{21}$$

$$\Delta\mathbf{w}_{11} \leftarrow \eta \frac{\partial J}{\partial \mathbf{w}_{11}}, \ \Delta\mathbf{w}_{21} \leftarrow \eta \frac{\partial J}{\partial \mathbf{w}_{21}}$$

$$\mathbf{w}_{11} \leftarrow \mathbf{w}_{11}/\|\mathbf{w}_{11}\|, \ \mathbf{w}_{21} \leftarrow \mathbf{w}_{21}/\|\mathbf{w}_{21}\|$$

(3.5)

where η denotes an appropriate stepsize.

3.1.2   Multiple Solutions

In the previous section we described how to get one-basis vector of the transformations $\mathbf{W}_1$ and $\mathbf{W}_2$. To find multiple solutions of $\mathbf{W}_{1j}$ and $\mathbf{W}_{2j}$ efficiently, deflationary orthogonalization [1] is considered. We need to run the one-basis algorithm several times for vectors $\mathbf{W}_{12},...,\mathbf{W}_{1p}$ and $\mathbf{W}_{22},...,\mathbf{W}_{2p}$. After every iteration, orthogonalization of the vectors is perfomed to prevent different vectors from converging to the same maxima. The learning is achieved by

$$\Delta\mathbf{w}_{1p} \leftarrow \eta \frac{\partial J}{\partial \mathbf{w}_{1p}}, \quad \mathbf{w}_{1p} \leftarrow \mathbf{w}_{1p} - \sum_{j=1}^{p-1} (\mathbf{w}_{1p}^T \mathbf{w}_{1j})\mathbf{w}_{1j}$$

$$\mathbf{w}_{1p} \leftarrow \mathbf{w}_{1p}/\|\mathbf{w}_{1p}\|$$

(3.6)

Simiarly, $\mathbf{w}_{2p}$ is found efficiently. This orthogonalization ensures that the proposed discriminant is defined by orthonomal basis vectors in a local frame.

3.2 The General Case (*L* local frames)

The learning algorithm is extended to the case which has an arbitrary number of local frames. The lower dimensional representation is locally obtained as $\mathbf{y}_i = \mathbf{W}_i^t \mathbf{x}$ *for* $\mathbf{x} \in \mathbf{L}_i$. The transformed global mean vector and the mean vector of a class *i* are given by $\widetilde{\mathbf{m}} = \sum_{i=1}^{L} \mathbf{W}_i^t \mathbf{m}_{L_i}$ and $\widetilde{\mathbf{m}}_i = \sum_{j=1}^{L} \mathbf{W}_j^t \mathbf{m}_{i,L_j}$ respectively. The transformed scatter matrices, are defined by

$$\widetilde{\mathbf{S}}_{\mathbf{B}} = \sum_{i=1}^{L} \mathbf{W}_i^t \mathbf{S}_{\mathbf{B},L_i} \mathbf{W}_i + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} 2\mathbf{W}_i^t \mathbf{R}_{\mathbf{B},ij} \mathbf{W}_j \tag{3.7}$$

$$\widetilde{\mathbf{S}}_W = \sum_{i=1}^{L} \mathbf{W}_i^t \mathbf{S}_{W,L_i} \mathbf{W}_i + \sum_{i=1}^{L} \sum_{j=1,j\neq i}^{L} 2\mathbf{W}_i^t \mathbf{R}_{W,ij} \mathbf{W}_j + \sum_{i=1}^{L} \sum_{j=1,j\neq i}^{L} \sum_{k=1,k\neq i,j}^{L} \mathbf{W}_j^t \mathbf{T}_{W,ijk} \mathbf{W}_k \tag{3.8}$$

where

$$\mathbf{S}_{\mathbf{B},L_j} = \sum_{i=1}^{c} n_i (\mathbf{m}_{i,L_j} - \mathbf{m}_{L_j})(\mathbf{m}_{i,L_j} - \mathbf{m}_{L_j})^T \tag{3.9}$$

$$\mathbf{R}_{\mathbf{B},L_{jk}} = \sum_{i=1}^{c} n_i (\mathbf{m}_{i,L_j} - \mathbf{m}_{L_j})(\mathbf{m}_{i,L_k} - \mathbf{m}_{L_k})^T \tag{3.10}$$

$$\mathbf{S}_{W,L_j} = \sum_{i=1}^{c} \left(\sum_{\mathbf{x} \in C_i, L_j} (\mathbf{x} - \mathbf{m}_{i,L_j})(\mathbf{x} - \mathbf{m}_{i,L_j})^T + (n_i - n_{i,L_j})\mathbf{m}_{i,L_j}\mathbf{m}_{i,L_j}^T \right) \tag{3.11}$$

$$\mathbf{R}_{W,jk} = \sum_{i=1}^{c} \left(\sum_{\mathbf{x} \in C_i, L_j} -(\mathbf{x} - \mathbf{m}_{i,L_j})\mathbf{m}_{i,L_k}^T \right) \tag{3.12}$$

$$\mathbf{T}_{W,jkl} = \sum_{i=1}^{c} \left(\sum_{\mathbf{x} \in C_i, L_j} \mathbf{m}_{i,L_k}\mathbf{m}_{i,L_l}^T \right) . \tag{3.13}$$

The constrined optimization problem is given by

$$\text{Max } J = tr\widetilde{\mathbf{S}}_B - k \cdot tr\widetilde{\mathbf{S}}_w, \text{ for } \|\mathbf{w}_{il}\| = 1 . \tag{3.14}$$

The gradient of the objective funtion *J* with respect to a vector $\mathbf{w}_{il}$ is

$$\frac{\partial J}{\partial \mathbf{w}_{il}} = (2\mathbf{S}_{B,L_i} - 2k\mathbf{S}_{W,L_i})\mathbf{w}_{il} + \sum_{j=1,j\neq i}^{L} 2\mathbf{R}_{B,ij}\mathbf{w}_{jl} - 2k \sum_{j=1,j\neq i}^{L} (\mathbf{R}_{W,ij} + \mathbf{R}_{W,ji}^T)\mathbf{w}_{jl} \tag{3.15}$$

$$- k \sum_{j=1,j\neq i}^{L} \sum_{k=1,k\neq i,j}^{L} (\mathbf{T}_{W,jik} + \mathbf{T}_{W,jki}^T)\mathbf{w}_{kl}$$

The multiple solutions which are orthonormal to the other vectors in the *i*-th local frame are found by

$$\Delta\mathbf{w}_{ip} \leftarrow \eta \frac{\partial J}{\partial \mathbf{w}_{ip}}, \quad \mathbf{w}_{ip} \leftarrow \mathbf{w}_{ip} - \sum_{j=1}^{p-1} (\mathbf{w}_{ip}^T \mathbf{w}_{ij})\mathbf{w}_{ij}$$

$$\mathbf{w}_{ip} \leftarrow \mathbf{w}_{ip} / \|\mathbf{w}_{ip}\| \tag{3.16}$$

4 Discussion

A solution of the above constrained optimization problem (3.5) is obtained by using Lagrangian multipliers as

$$L = tr\left[\widetilde{\mathbf{S}}_B - k\widetilde{\mathbf{S}}_w - \Lambda_1(\mathbf{W}_1^T\mathbf{W}_1 - \mathbf{I}) - \Lambda_2(\mathbf{W}_2^T\mathbf{W}_2 - \mathbf{I})\right]$$

where $\boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_{i1} \cdot & \mathbf{O} \\ & \ddots \\ \mathbf{O} & \lambda_{ip} \end{bmatrix}$ is the diagonal matrix of eigenvalues and $\mathbf{I}$ is the identity matrix.

The gradient of the above lagrangian function with respect to the basis vectors is

$$\frac{\partial L}{\partial \mathbf{w}_{1l}} = \left(2\mathbf{S}_{B,L_1} - 2k\mathbf{S}_{W,L_1} - 2\lambda_1\mathbf{I}\right)\mathbf{w}_{1l} + \left(2\mathbf{R}_B - 2k\mathbf{R}_W - 2k\mathbf{T}_W^T\right)\mathbf{w}_{2l} = 0$$

$$\frac{\partial L}{\partial \mathbf{w}_{2l}} = \left(2\mathbf{R}_B^T - 2k\mathbf{R}_W^T - 2k\mathbf{T}_W\right)\mathbf{w}_{1l} + \left(2\mathbf{S}_{B,L_2} - 2k\mathbf{S}_{W,L_2} - 2\lambda_2\mathbf{I}\right)\mathbf{w}_{2l} = 0$$

Although we did not provide the convergence proof for the gradient based iterative learning method described in the previous section, the convergence of the proposed method to a global maximum can be expected by virtue of the criterion being the 2nd-order convex function with respect to basis vectors $\mathbf{w}_{1l}$ and $\mathbf{w}_{2l}$ and the two variables jointly. A few examples of learning are given in Figure 2. They show that the objective function has two local maxima corresponding to two sets of basis vectors in opposite directions. Both cases yield the same value of the objective function which is a global maximum. It is also noted that the gradient method converges stably regardless of the constant k.

We imposed the orthonormal condition for each local frame in order to find multiple solutions easily. However, conventional LDA has non-orthogonal axes in a single global frame. The validity of orthonormal condition in local frame should be examined further.
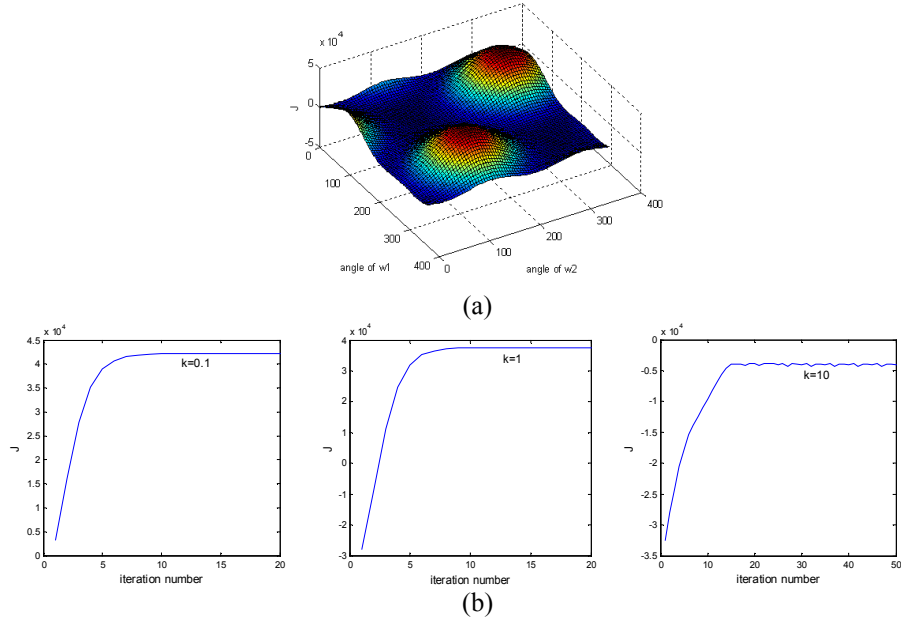

(a)

(b)

Figure 2: A learning example. (a) The value of the objective function (k=0.1) as a function of orientation of $\mathbf{w}_{1l}$ and $\mathbf{w}_{2l}$. (b) The convergence graphs with k=0.1, k=1 and k=10. The data distribution is given in Figure 1.

5   Experiments

5.1   Simulated Data

Two 2D simulated data sets were created and tested to demonstrate the superiority of the proposed learning algorithm. The first set has three classes which has two distinct modalities in their distributions. The second set has two classes which has three distinct peaks in distribution. The data sets are illustrated in Figure 3. The conventional LDA, mixture of LDA, GDA and the proposed discriminant are compared in terms of classfication error. Euclidean distance, normalized cross-correlation and Mahalobis distance were utilized for N-N classification. We assume that the number of the local frames are given. For defining the local groups, various techniques can be utilized. Here, the data set was simply and well separated into several local groups by the k-means clustering algorithm. The recognition results are given in Table 1. It is shown that the proposed discriminant can solve the non-linear classification problem on which the conventional linear method fails and it is much better in terms of computational efficiency as compared to the GDA.

5.2   Face Data

The proposed algorithm has been validated on real face data. The face images which have a large variation of pose have been known to be multi-modally distributed. The previous study [6] have attempted to synthesize and recognize a novel view face image by modeling the face view spaces, which consist of face images within a certain range of view-angles. The face view sapce can be considered as the local group in the proposed learning algorithm. We used the XM2VTS data set which has the pose label of the face and the pose label was utilized to define the local groups. The face database consists of 295*2 facial images normalized to 23*28 pixel resolution with a fixed eye position. We have the frontal and right-rotated view images of each identity. The frontal face was
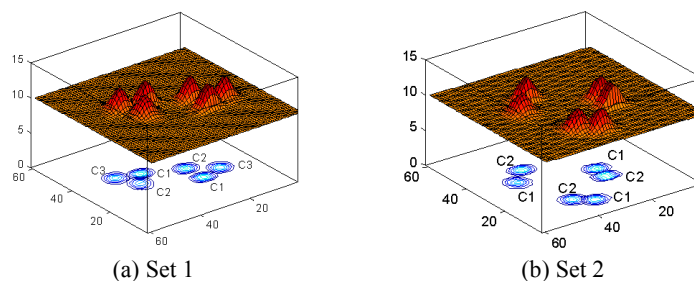

| | (a) Set 1 | (b) Set 2 |

Figure 3: Simulated data distributions

		Euclidean	Cross-corr.	Mahal	Relative F.E. complexity
		Error	Error	Error	
Set 1 (400 samples / class)	**Proposed**	**7.6±3.5**	**8±3.6**	**7.3±3.7**	**1+alpha**
	LDA	266.6±115.4	266.6±115.4	81.3±61.6	1
	LDA mixture	254±27.8	255±23.5	169.6±45.5	1+alpha
	GDA	4.3±1.1	4.3±1.1	4.4±0.5	270
Set 2 (600 samples / class)	**Proposed**	**8±1.4**	**8±1.4**	**7±2.8**	**1+alpha**
	LDA	308.5±129.4	308.5±129.4	207.5±272.2	1
	LDA mixture	205±1.4	205±1.4	206±7	1+alpha
	GDA	4±1.4	4±1.4	4±0	278

* alpha indicates a computational cost for deciding which local group a new pattern belongs to. It is usually less than 1.

Table 1: Classification Results

registered and the rotated face image was considered as a query. For simplicity of the learning, the algorithm was applied to the first 50 eigenfeatures. The eigenvalue plot of the set showed that the first 50 features were enough to describe the images. Figure 4 shows the transformation vectors of PCA and the proposed discriminant. The transformation vectors of the frontal faces and right-rotated faces are visualized in the first and second row respectively. It is noted that the relationship between the frontal eigenfaces and rotated eigenfaces is hard to describe except for the first eigenface. The first eigenfaces show a certain rotation, scaling and translation relationships between the two. On the contrary, all the corresponding transformation vectors shows a certain relationship yielding the same feature of the same face regardless of the view-angles in the locally linear transformation. It also appears to provide a discriminat feature for different faces with the same view-angle like the conventional fisherface method. 3 training and test sets were randomly created for the two cases that have different number of training and test images. The case 1 has 245*2 images of 245 persons for training and 50*2 images of 50 persons for testing. The case 2 has 100*2 images and 195*2 images for training and test respectively. In the proposed algorithm, k was selected heuristically to yield the best performance for the training set. For the GDA, RBF kernel was utilized with the adjustment of the standard deviation of the kernel. It is noted that the GDA is highly overfitted on the training set but the proposed method is robust to the test set. Figure 5 shows the average recognition percentage of novel view face images with a standard deviation.
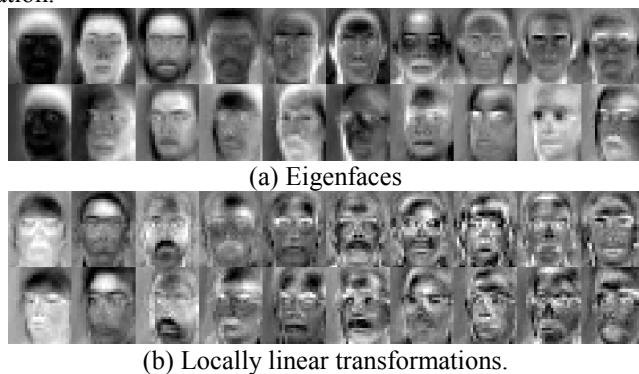

(a) Eigenfaces

(b) Locally linear transformations.
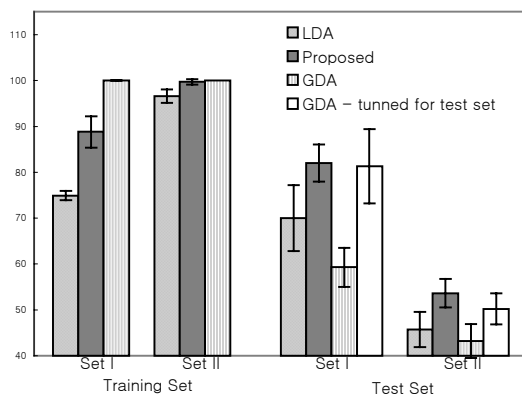Figure 4: Visualization of the transformation vectors.

Figure 5: Recognition results of a novel view face image. 'GDA- tuned for test set' indicates the results were obtained by adjustment of the kernel parameter for the test set.

6 Conclusion

A novel learning method has been described for the discriminant analysis which can classify a non-linear structure. Multiple local linear transformations are considered to yield that the locally transformed classes maximize the between-class covariance and minimize the within-class covariance. The learning method for finding the optimal set of local linear bases does not have a local-maxima problem and stably converge to a global maximum point. The classification results obtained on both simulated data and real face data show that the proposed discriminant provides a set of discriminant features for linearly non-separable data and it is computational efficient as compared with the non-linear discriminant analysis based on the kernel approach. The method does not much suffer from the problem of overfitting by virtue of the linear base structure of the solution. A more effective learning procedure will be sought in the future by finding a closed-form solution and a method to decide constant k.

Acknowledgements

References

[1] Aapo Hyvarinen, Juha Karhunen and Erkki Oja, *Independent Component Analysis*, John Wiley & Sons, Inc. 2001.

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

[3] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization", *Adv. Neural Info. Proc. Syst.* 13, 556-562, 2001.

[4] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature* , 401, 788-791, 1999.

[5] Sam T. Roweis and Lawrence 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. PAMI*, vol. 19, no. 7, pp. 733-742, 1997.

[7] Hyun-Chul Kim, Daijin Kim, Sung-Yang Bang, "Face Recognition Using LDA Mixture Model," *International Conference on Pattern Recognition*, Canada, 2002.

[8] Fukunaga, K. *Introduction to statistical pattern recognition* (2nd 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. on PAMI*, vol. 23, no. 6, pp. 643-660, 2001.

[10] K.Okada, C.Malsburg, "Analysis and synthesis of human faces with pose variations by a parametric piecewise linear subspace method", *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1 , pp. I -761-8 , 2001.

[11] T.-K. Kim, H. Kim, W. Hwang, S. Kee and J. Kittler, "Independent Component Analysis in a Facial Local Residue Space", *IEEE International Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, 2003.

[12] T.-K. Kim, H. Kim, W. Hwang, S. Kee and J. Kittler, "Face Description based on Decomposition and Combining of a Facial Space with LDA, *IEEE International Conference on Image Processing*, Spain, 2003, to appear.

[13] T.-K. Kim, H. Kim, W. Hwang, S. Kee, J. Lee, "Component-based LDA Face Descriptor for Image Retrieval", *British Machine Vision Conference*, Cardiff, UK, Sept. 2-5, 2002.