Enhanced Direct Linear Discriminant Analysis for Feature Extraction on High Dimensional Data

A. K. Qin[1], **S. Y. M. Shi**[1], **P. N. Suganthan**[1] and **M. Loog**[2]

[1]School of EEE, Nanyang Technological University, Singapore 639798
{qinkai, yeminshi}@pmail.ntu.edu.sg, epnsugan@ntu.edu.sg
[2]Image Analysis Group, Department of Innovation, IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark
marco@itu.dk

Abstract

We present an enhanced direct linear discriminant analysis (EDLDA) solution to effectively and efficiently extract discriminatory features from high dimensional data. The EDLDA integrates two types of class-wise weighting terms in estimating the average within-class and between-class scatter matrices in order to relate the resulting Fisher criterion more closely to the minimization of classification error. Furthermore, the extracted discriminant features are weighted by mutual information between features and class labels. Experimental results on four biometric datasets demonstrate the promising performance of the proposed method.

1. Introduction

Fisher's linear discriminant analysis (LDA) is one of the most popular supervised feature extraction techniques. LDA seeks an optimal set of discriminant projection vectors $\mathbf{W}=[\varphi_1,...,\varphi_d]$, to map the original data space onto a lower dimensional feature space, by maximizing the Fisher criterion: $J_F(\mathbf{W})=\left|\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}\right|/\left|\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}\right|$. Here, $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between-class and within-class scatter matrices of the training sample group respectively, and are estimated as follows:

$$\mathbf{S}_b=\sum_{i=1}^{C}P_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^{\mathrm{T}}=\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}P_iP_j(\mathbf{m}_i-\mathbf{m}_j)(\mathbf{m}_i-\mathbf{m}_j)^{\mathrm{T}}$$
$$\mathbf{S}_w=\sum_{i=1}^{C}P_i\mathbf{S}_i \qquad (1)$$

where C, P_i, $\mathbf{m}_i$, $\mathbf{m}$ and $\mathbf{S}_i$ represent the number of pattern classes, *a priori* probability of pattern class ω_i, the mean vector of the samples in class ω_i, the mean vector of all samples and the covariance matrix of samples in class ω_i, respectively. The between-class scatter matrix $\mathbf{S}_b$ can be expressed by both the original definition and its equivalent pairwise decomposition form (Loog, Duin, & Haeb-Umbach, 2001). The total scatter matrix $\mathbf{S}_t$ is the summation of $\mathbf{S}_b$ and $\mathbf{S}_w$, estimated as $\mathbf{S}_t=\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^{\mathrm{T}}$. We can easily verify that:

$$rank(\mathbf{S}_w)\le\min(n,N-C)$$
$$rank(\mathbf{S}_B)\le\min(n,C-1)$$
$$rank(\mathbf{S}_t)\le\min(n,N-1)$$

where n and N denote the dimensionality and the total number of sample vectors, respectively.

However, the LDA may encounter difficulties if applied to problems involving high dimensional data, such as images. Usually, the vectorization of images results in very high dimensional data vectors, e.g. the vector representation of images of size 100×100 leads to a 10000 dimensional feature space. Since the LDA technique operates on the very high dimensional scatter matrices constructed from the high dimensional datasets, computational challenge (sometimes intractable) will be introduced to the eigen-analysis. Moreover, those scatter matrices are always singular in these high dimensional cases unless the number of training samples is more than the number of dimensions, which is usually impossible in applications such as face recognition, palmprint recognition and so on.

Many approaches (Belhumeur, Hespanha, & Kriegman, 1997; Chen, Liao, & Ko et al., 2000; Yu & Yang, 2001; Lu, Plataniotis, & Venetsanopoulos, 2003a; Lu, Plataniotis, & Venetsanopoulos, 2003b; Price & Gee, 2005; Thomaz, Gillies, & Feitosa, 2004) have been proposed to deal with the above difficulties. Belhumeur et al. proposed a two stage PCA+LDA method (Belhumeur, Hespanha, & Kriegman, 1997), which first applied the Principle Component Analysis (PCA) to project the high dimensional data onto an intermediate lower dimensional space where the average within-class scatter matrix $\mathbf{S}_w$ is guaranteed to be non-singular, then the LDA can be executed to further reduce the dimensionality. Although this method has been widely used in many applications, especially in the face recognition field, it has a major problem that the PCA step used to make the $\mathbf{S}_w$ non-

degenerate simultaneously removes the null space of $\mathbf{S}_w$, which usually includes most important discriminatory information (Chen, Liao, & Ko et al., 2000). Chen et al. proposed a null space based LDA (NLDA) technique (Chen, Liao, & Ko et al., 2000), where the sample vectors are firstly projected to the null space of the $\mathbf{S}_w$ and then the optimal set of discriminant vectors is extracted from this null space by maximizing the between-class scatter matrix $\mathbf{S}_b$ by the PCA. The NLDA technique suffers from high computation complexity because the high dimensionality of $\mathbf{S}_w$ places high computational demand when extracting the null space of $\mathbf{S}_w$. Further, NLDA is sensitive to the number of pattern classes and the number of training samples per class since these two numbers determine the dimension of the null space of $\mathbf{S}_w$. The Direct LDA (DLDA) method (Yu & Yang, 2001), proposed by Yu and Yang, employed a computational technique (described in section 2), to operate on scatter matrices of much smaller scale, usually $C \times C$, which dramatically reduce the computational complexity and avoid singularity problem related to the traditional LDA procedure. It is worth noting that DLDA is equivalent to traditional LDA on problems of low dimensionality, where all scatter matrices are non-singular.

The DLDA method has achieved good performance (Yu & Yang, 2001; Lu, Plataniotis, & Venetsanopoulos, 2003a; Lu, Plataniotis, & Venetsanopoulos, 2003b; Price & Gee, 2005) in face recognition tasks. However, as the Fisher criterion used is not directly related to the minimization of classification error, it suffers from two deficiencies: 1) from the equivalent pairwise decomposition expression of the $\mathbf{S}_b$ matrix in Eq. (1), we can observe that the class pairs with large distances between them are overemphasized in eigen-analysis of the pairwise formula (Loog, Duin, & Haeb-Umbach, 2001). Hence, the obtained discriminant projection directions attempt to preserve the distances of already well-separated classes while causing larger overlap between pairs of classes that are not well separated in the original space. Consequently, the discriminant directions to well separate the neighboring classes cannot be obtained if there are some outlier classes that are far away and well separated from some other classes. 2) The estimation of the average within-class scatter matrix $\mathbf{S}_w$ in Eq. (1) assumes that all classes have the same covariance matrix, which is usually violated in reality. Therefore, classes with largely deviating covariance matrices may dominate the eigen-decomposition of $\mathbf{S}_w$. If those dominated covariance matrices coincidently belongs to the outlier classes far away from all the other classes, the resulting projection directions attempt to minimize the spread of the outlier classes while neglecting the minimization of the covariance matrices that indeed impacts the classification error.

In this paper, we propose an enhanced direct linear discriminant analysis (EDLDA) solution integrating two types of class-wise weighting terms: class-pair weights and within-class weights in estimating the between class scatter matrices and average within class in order to relate the resulting Fisher criterion more closely to the minimization of classification error. Furthermore, the extracted discriminant features are weighted by the mutual information between features and class labels. Experimental results on one Palmprint dataset and three face image datasets (Yale, Pix and AR) demonstrate the very good performance of the proposed method.

2. Enhanced Direct LDA

2.1 Direct LDA Solution

Due to low computational complexity, no singularity problem, and promising performance, the DLDA technique has been successfully applied in face recognition. The core idea behind the DLDA is to simultaneously diagonalize the between-class scatter matrix $\mathbf{S}_b$ and the average with-class scatter matrix $\mathbf{S}_w$ while taking the order as first diagonalizing $\mathbf{S}_b$, and then $\mathbf{S}_w$. In the first step of $\mathbf{S}_b$ diagonalization, the null space of $\mathbf{S}_b$ is removed and the training sample vectors are projected from original data space $\mathbf{R}^n$ onto the $C-1$ column spaces of $\mathbf{S}_b$. This requires the computation of the eigenvectors of $\mathbf{S}_b$ corresponding to non-zero eigenvalues. Although the dimensionality of $\mathbf{S}_b$ is very high ($n \times n$), since the matrix $\mathbf{S}_b$ can be expressed as:

$$\mathbf{S}_b = \sum_{i=1}^{C} P_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^{\mathrm{T}} = \mathbf{U}\mathbf{U}^{\mathbf{T}}$$

where $\mathbf{U} = \left[\sqrt{P_1}(\mathbf{m}_1 - m),\ldots,\sqrt{P_c}(\mathbf{m}_c - m)\right]$, the eigenvectors corresponding to non-zero eigenvalues of $\mathbf{S}_b$ can be computed through the eigen-analysis of a generally much smaller[1] $C \times C$ matrix $\mathbf{U}^{\mathbf{T}}\mathbf{U}$ as follows:

$$\left(\mathbf{U}^{\mathbf{T}}\mathbf{U}\right)\mathbf{V} = \mathbf{V}\lambda \rightarrow \mathbf{U}\mathbf{U}^{\mathbf{T}}(\mathbf{U}\mathbf{V}) = (\mathbf{U}\mathbf{V})\lambda$$

where $\mathbf{V}$ and λ are an eigenvector and the corresponding eigenvalue of matrix $\mathbf{U}^{\mathbf{T}}\mathbf{U}$, and the $\mathbf{U}\mathbf{V}$ denotes an eigenvector of $\mathbf{S}_b = \mathbf{U}\mathbf{U}^{\mathbf{T}}$ with the corresponding eigenvalue λ. We take $C-1$ eigenvectors $\mathbf{\Psi}_1 = [\mathbf{U}\mathbf{V}_1,\ldots,\mathbf{U}\mathbf{V}_{C-1}]$ corresponding to non-zeros eigenvalues $\mathbf{\Lambda}_1 = [\lambda_1,\ldots,\lambda_{C-1}]$ of $\mathbf{S}_b$ and project all training sample vectors from $\mathbf{X} \in \mathbf{R}^n$ to $\left(\mathbf{\Lambda}_1^{-1/2}\mathbf{\Psi}_1^{\mathbf{T}}\right)\mathbf{X} \in \mathbf{R}^{C-1}$. After the projection, $\mathbf{S}_b$ is whitened and the second step of $\tilde{\mathbf{S}}_w = \left(\mathbf{\Lambda}_1^{-1/2}\mathbf{\Psi}_1^{\mathbf{T}}\right)\mathbf{S}_w\left(\mathbf{\Psi}_1\mathbf{\Lambda}_1^{-1/2}\right)$ diagonalization proceeds. Since the matrix $\tilde{\mathbf{S}}_w$ is of size $C-1 \times C-1$, the computational complexity of the $\tilde{\mathbf{S}}_w$ eigen-decomposition is low. Thus we take the *m* eigenvectors $\mathbf{\Psi}_2$ corresponding to *m* lowest (possibly zero) eigenvalues $\mathbf{\Lambda}_2$. The projection matrix $\mathbf{\Psi}_1\mathbf{\Lambda}_1^{-1/2}\mathbf{\Psi}_2$ should be normalized along columns to conform to orthonormality.

[1] The number of classes is usually much smaller than the number of training samples and the number of dimensions in high dimensional problems

We note that the between-class scatter matrix (Eq.1) used in the DLDA is constructed by only the class means, which would not be optimal unless the covariance matrices of all classes are equal. Although this condition is actually hard to satisfy in real applications, the DLDA does demonstrate good performance in face recognition tasks. We explain this fact as follows: because the sample vectors in face recognition tasks are characterized by high dimensionality and small size, the covariance matrix of each class cannot be well estimated by any technique. Therefore, why could not we assume that all classes have the same covariance matrix where only a very small number of data vectors have been sampled into the training set? Justifying in this manner, we may regard the usage of only class means to estimate the $\mathbf{S}_b$ be reasonable in the DLDA.

However, as stated in (Loog, Duin, & Haeb-Umbach, 2001), the Fisher criterion employed in the traditional LDA technique is not an optimal approximation of the minimization of the Bayes error especially when outlier classes (described in the introduction section) exist. This shortcoming is inherited by the DLDA without any improvements. In the next section, we propose a Class-weighted Direct LDA solution (CwDLDA) introducing the class-pair and within-class weights in the estimation of $\mathbf{S}_b$ and $\mathbf{S}_w$ to handle this problem.

2.2 Class-wise Weighted Direct LDA Solution

The Fisher criterion employed in the traditional LDA is only a sub-optimal approximation of the Bayes error but it is easy to implement due to the non-iterative eigen-decomposition technique. Two major deficiencies associated with the original Fisher criterion are: 1) it attempts to find the linear transformation to maximize the distance between the classes in the projected low dimensional space, which is not directly related to the minimization of misclassification. Therefore, class pairs with large distances would dominate the eigen-decomposition of $\mathbf{S}_b$ such that the obtained discriminatory directions attempt to preserve the distances of already well-separated classes. Therefore, the discriminant directions to well separate the neighboring classes cannot be obtained if there exist outlier classes well separated from the other classes. 2) The expression of the average within-class scatter matrix $\mathbf{S}_w$ in Eq. 1 has an implicit assumption that all classes have the same covariance matrix. If it is not true, those significantly deviating covariance matrices may dominate the eigen-decomposition process of $\mathbf{S}_w$. Therefore, the LDA may fail when the dominating covariance matrices belong to outlier classes since the obtained transformation attempt to minimize the within-class scatter of the outlier classes while neglecting the minimization of those classes' covariance matrices impacting the classification error.

To tackle the above deficiencies in the DLDA, we devised a Class-weighted Direct LDA (CwDLDA) solution by introducing class-pair weights and the within-class weights in the estimation of the between-class and within-class scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$, respectively.

For each pair of classes in the pairwise decomposed definition of $\mathbf{S}_b$ in Eq. 1, we introduce a class-pair weighting factor, suggested in (Loog, Duin, & Haeb-Umbach, 2001), $w_{ij}^c(d) = \left(1/2d^2\right) erf\left(d/2\sqrt{2}\right)$, where $erf(x) = \left(2/\sqrt{\pi}\right)\int_0^x e^{-t^2}dt$ is related to the pairwise approximated Bayesian accuracy. Therefore, the neighboring class pairs, not well separated in the original feature space, can have increased influence in the computation of $\mathbf{S}_b$ than those already well separated. The $\mathbf{S}_b$ in Eq. 1 is thus modified to:

$$\hat{\mathbf{S}}_b = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} P_i P_j w_{ij}^c(d_{ij})(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}} \tag{2}$$

where $d_{ij} = (\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}}\mathbf{S}_w^{-1}(\mathbf{m}_i - \mathbf{m}_j)$ is the Mahalanobis distance between classes ω_i and ω_j. Since the within-class weighting term is used to eliminate the influences form the possible outlier classes, it can be calculated based on the class-pair weighting factors as: $\boldsymbol{w_i^w} = \sum_{j=1}^{C} \boldsymbol{w_{ij}^c}(\boldsymbol{d_{ij}})$.

Therefore, the outlier classes with small class-pair weighting factors to the other classes also receive small weights in the estimation of the weighted average within-class scatter matrix expressed as:

$$\tilde{\mathbf{S}}_w = \sum_{i=1}^{C} P_i w_i^w \mathbf{S}_i \tag{3}$$

As pointed out and proven in (Price & Gee, 2005), assuming that the class-pair weights are always above zero, the null space of the non-weighted $\mathbf{S}_b$ is equal to the weighted one $\tilde{\mathbf{S}}_b$. Therefore, we could remove the null space of $\tilde{\mathbf{S}}_b$ by projecting all training samples on the column space of $\mathbf{S}_b$, where the projection directions can be obtained by the eigen-decomposition of a $C \times C$ matrix. The weighting term $w_{ij}^c(d)$ and w_i^w are calculated in the transformed $C-1$ dimensional space.

The calculation of the Mahalanobis distance between a pair of classes involving the inversion of the average within-class scatter matrix that is a non-weighted estimation in the $C-1$ dimensional space. To avoid the occasional singularity problem of $\bar{\mathbf{S}}_w \in \mathbf{R}^{C-1\times C-1}$, we employ a recently proposed covariance matrix estimation technique called maximum entropy covariance selection method (Thomaz, Gillies, & Feitosa, 2004) to re-estimate the $\bar{\mathbf{S}}_w$ when it is singular. The procedure of re-estimation of the matrix $\bar{\mathbf{S}}_w$ is shown in Table 1.

Table 1. The Maximum Entropy Covariance Selection Algorithm

1. Find eigenvectors $\mathbf{\Phi}$ and corresponding eigenvalues $\mathbf{\Lambda}$ of $\mathbf{S}_p \in R^{C-1\times C-1}$, where $\mathbf{S}_p = \bar{\mathbf{S}}_w/(N-C)$ is the unbiased estimation $\bar{\mathbf{S}}_w$

2. Calculate the average eigenvalue $\bar{\lambda}$ of $\mathbf{S}_p$

3. Form a new diagonal eigenvalue matrix as : $\Lambda^{new} = diag\left(\max\left(\lambda_1, \bar{\lambda}\right), ..., \max\left(\lambda_{C-1}, \bar{\lambda}\right)\right)$

4. Form the new average within-class scatter matrix as $\bar{\mathbf{S}}_w^{new} = \mathbf{S}_p^{new} \cdot (N-C) = \left(\mathbf{\Phi}\mathbf{\Lambda}^{new}\mathbf{\Phi}^{\mathbf{T}}\right) \cdot (N-C)$

The detailed description of the CwDLDA solution is listed in Table 2 without performing the step 6.

2.3 Enhanced Direct LDA Solution

Compared with the original DLDA algorithm, the Fisher Criterion in the CwDLDA with the weighted estimation of between and within class scatter matrices is more related to the minimization of classification error. However, since it is still a suboptimal solution to approximate the Bayes error criterion, the extracted features may not retain the complete discriminatory power. Hence, we consider assigning each extracted feature $\boldsymbol{f_i} = \left(f_{i1}, f_{i2}, ..., f_{iN}\right) \in \mathbf{R}^N$ $\boldsymbol{i} = 1, ..., \boldsymbol{m}$ with different weights $\kappa_{\boldsymbol{i}}, \boldsymbol{i} = 1, ..., \boldsymbol{m}$. The mutual information is a nonlinear metric to evaluate the correlation between two random variables, which has demonstrated good performance in measuring the salience of features (Kwak & Choi, 2002). Therefore, in our paper, we choose the normalized mutual information $I(f_i, \mathbf{C})$ between f_i and the corresponding class label $\mathbf{C} = \{\boldsymbol{c_i}, \boldsymbol{i} = 1, ..., \boldsymbol{N} \mid \boldsymbol{c_i} \in \{1, ..., \boldsymbol{C}\}\}$ to weight features As suggested in (Kwak & Choi, 2002), we evenly discretize each feature into 10 equal intervals between $[\mu_i - 2*\delta_i, \mu_i - 2*\delta_i]$, where μ_i and δ_i are the mean and standard deviation of f_i respectively, and then count the sample frequency in each interval as the probability. Therefore, $I(f_i, \mathbf{C})$ is calculated as:

$$\boldsymbol{I}(\boldsymbol{f_i}, \mathbf{C}) = -\sum_{s=1}^{C}\sum_{t=1}^{10} \boldsymbol{P}(\boldsymbol{f_{it}}, \boldsymbol{s}) \log(\boldsymbol{P}(\boldsymbol{f_{it}}, \boldsymbol{s}) / \left(\boldsymbol{P}(\boldsymbol{f_{it}})\boldsymbol{P}(\boldsymbol{s})\right)) \qquad (4)$$

where $\boldsymbol{P}(\boldsymbol{f_{it}}, \boldsymbol{s})$ is the joint probability of $\boldsymbol{f_{it}}$ and $\boldsymbol{s}$ and $\boldsymbol{f_{it}}$ is the tth interval of feature $\boldsymbol{f_i}$. Consequently, κ_i can be expressed as:

$$\kappa_{\boldsymbol{i}} = \boldsymbol{I}(\boldsymbol{f_i}, \mathbf{C}) \Big/ \sum_{j=1}^{\boldsymbol{m}} \boldsymbol{I}(\boldsymbol{f_j}, \mathbf{C}), \boldsymbol{i} = 1, ..., \boldsymbol{m} \qquad (5)$$

If mutual information between f_i and $\mathbf{C}$ is large, it means that f_i and $\mathbf{C}$ are closely related, and vice versa.

By combining the above step with the CwDLDA solution, an Enhanced Direct LDA (EDLDA) solution is devised. The outline of the EDLDA method is described in Table 2.

Table 2. The algorithmic description the Enhanced Direct LDA

1. Find a set of $C-1$ orthonormal eigenvectors $\mathbf{\Phi}_1$ and the corresponding non-zero eigenvalues of the between-class scatter matrix $\mathbf{S}_b$ defined in the original data space $\mathbf{R}^n$. Remove the null space of $\mathbf{S}_b$ by projecting all samples onto $\mathbf{R}^{C-1}$, i.e. $\mathbf{x} \in \mathbf{R}^n \rightarrow \mathbf{\Phi}_1^{\mathbf{T}}\mathbf{x} \in \mathbf{R}^{C-1}$

2. Calculate $\bar{\mathbf{S}}_w$ in the reduced space $\mathbf{R}^{C-1}$. If $\bar{\mathbf{S}}_w$ is not singular, $\bar{\mathbf{S}}_w^{-1}$ is computed, otherwise it is re-estimated as $\bar{\mathbf{S}}_w^{new}$ by method in Table 1 and the inverse of $\bar{\mathbf{S}}_w^{new}$ is computed to replace $\bar{\mathbf{S}}_w^{-1}$

3. Calculate, in $\mathbf{R}^{C-1}$, the weighted between-class and within-class scatter matrices $\tilde{\mathbf{S}}_b$ and $\tilde{\mathbf{S}}_w$ using Eqs. 2 and 3

4. Whiten $\tilde{\mathbf{S}}_b$ by: $\mathbf{\Phi}_2^{\mathbf{T}}\tilde{\mathbf{S}}_b\mathbf{\Phi}_2 = \mathbf{\Lambda}_1 \rightarrow \mathbf{\Lambda}_1^{-1/2}\mathbf{\Phi}_2^{\mathbf{T}}\tilde{\mathbf{S}}_b\mathbf{\Phi}_2\mathbf{\Lambda}_1^{-1/2} = \mathrm{I}_{C-1\times C-1}$ and project all samples by transformation matrix $\mathbf{\Lambda}_1^{-1/2}\mathbf{\Phi}_2^{\mathbf{T}}$. Thus, we have $\tilde{\mathbf{S}}_w = \left(\mathbf{\Lambda}_1^{-1/2}\mathbf{\Phi}_2^{\mathbf{T}}\right)\tilde{\mathbf{S}}_w\left(\mathbf{\Phi}_2\mathbf{\Lambda}_1^{-1/2}\right)$

5. Diagonalize $\tilde{\tilde{\mathbf{S}}}_w$ by $\mathbf{\Phi}_3^{\mathbf{T}}\tilde{\tilde{\mathbf{S}}}_w\mathbf{\Phi}_3 = \mathbf{\Lambda}_2$ and take the m eigenvectors $\mathbf{\Phi}_3^{'}$ from $\mathbf{\Phi}_3$ corresponding to the m smallest eigenvalues. Therefore, the total projection transformation is obtained as:
$$\mathbf{x} \in \mathbf{R}^n \rightarrow \mathbf{\Phi}_3^{'\mathbf{T}}\mathbf{\Lambda}_1^{-1/2}\mathbf{\Phi}_2^{\mathbf{T}}\mathbf{\Phi}_1^{\mathbf{T}}\mathbf{x} \in \mathbf{R}^m$$
which transforms data from $\mathbf{R}^n$ to $\mathbf{R}^m$

6. Calculate the weights $\kappa_{\boldsymbol{i}}, \boldsymbol{i} = 1,..\boldsymbol{m}$ for each of the extracted features $\boldsymbol{f_i} \in \mathbf{R}^N, \boldsymbol{i} = 1,..\boldsymbol{m}$ by Eqs. 4 and 5

3. Experimental Results

We applied our proposed CwDLDA and EDLDA methods on 4 real biometric datasets to evaluate their performances compared with the original DLDA. We do not list the comparison with results of other linear dimension reduction techniques since our current work focuses on the enhancement of the DLDA technique itself while the promising performance of the DLDA compared with other techniques has been illustrated in many literatures (Belhumeur, Hespanha, & Kriegman, 1997; Chen, Liao, & Ko et al., 2000; Yu & Yang, 2001; Lu, Plataniotis, & Venetsanopoulos, 2003a; Lu, Plataniotis, & Venetsanopoulos, 2003b; Price & Gee, 2005). It is worth noting that our proposed technique can also be easily generalized to other linear dimension reduction methods. The 4 datasets[1] we used are: 1) NTU palmprint datasets, which contains 10 different right hand palmprint images of 40 persons, for a total of 400 images with the size of 200×200 . 2) Yale face image dataset comprising 165 images with 11 images for each of 15 subjects. All images are centered and cropped to the size 127×99. 3) Pix face image datasets is composed of 300 images with 30 subjects. Each subject has 10 face images of a size 512×512, which are sub-sampled to the size of 100×100 in our experiments. 4) AR face image dataset is a large and complex one. Here we use a subset of AR that contains 1638 face images with 13 images for 133 persons. All images are sub-sampled from the original size 768×576 to a smaller size of 60×40 for our experiments.

[1] The three face image datasets used are publicly available.

In our experiments, we randomly choose 5, 5, 5 and 7 sample images of each person from the dataset to be the training set and the remaining samples constitute the testing set. The number of features extracted is defined as $C-1$ and the 1-nearest neighbor approach is used as the classifier. The experiments are repeated 20 times for each approach and two performance measures are employed. The first one is the average classification accuracy (*CR*) over 20 runs on the testing set. The second one is the average improvements of the classification rate (*AveImp*) of the proposed CwDLDA and EDLDA methods over the DLDA solution in terms of all possible numbers of extracted features from 1 to $C-1$. The results are illustrated in Table 3.

Table 3. Performance Comparison of DLDA, CwDLDA and EDLDA on 4 biometric datasets

Dataset	Perfor. Measure	DLDA (%)	CwDLDA (%)	EDLDA (%)
Palm	*CR*	93.43	95.67	96.40
	AveImp	---	2.47	2.46
Yale	*CR*	74.72	76.39	78.00
	AveImp	---	3.47	3.69
Pix	*CR*	95.77	96.50	97.37
	AveImp	---	1.86	2.15
AR	*CR*	34.84	63.83	75.01
	AveImp	---	15.52	16.22

It is clear that the CwDLDA solution outperforms the DLDA solution on both the average classification accuracy and the average improvement classification rate on all 4 datasets, while the EDLDA approach further improves the CwDLDA. The improvements of the EDLDA and CwDLDA over the DLDA on the AR face image dataset is significant due to the existence of a large number of classes and the complex image sample distribution within this dataset.

4. Conclusions

In this paper, we proposed an enhanced direct LDA (EDLDA) solution incorporating the class-wise weights and the within-class weights into the estimation of the between-class and average within-classes scatter matrices, respectively. It may alleviate the influences from the possible outlier classes such that the resulting Fisher criterion is more closely related to the minimization of classification error. Moreover, the extracted discriminant features are weighted by the normalized mutual information between each feature vector and the corresponding class label vector. Experimental results on 4 biometric datasets (NTU palmprint, Yale, Pix and AR) demonstrate the promising performance of the proposed method. We note that the proposed schemes can also be easily generalized into other linear discriminant methods.

Acknowledgement

Thanks for Dr. Jieping Ye to make its Matlab codes and associated datasets publicly available for academic usage.

References

Loog, M., Duin, R.P.W., Haeb-Umbach, R., "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria", *IEEE Trans. on PAMI*, 23(7): 762-766, 2001.

Belhumeur, P., Hespanha, J., D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. on PAMI*, 19(7): 711-720, 1997.

Chen, L-F, Liao, H-Y M, Ko., M-T, Lin, J-C, Yu, G-J, "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, 33(10) : 1713-1726, 2000.

Yu, H., Yang, J., "A direct LDA algorithm for high-dimensional data – with application to face recognition", *Pattern Recognition*, 34(10): 2067-2070, 2001.

J W Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms", IEEE Trans. on Neural Networks, 14(1): 117-126, 2003.

J W Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using LDA Based Algorithms", *IEEE Trans. on Neural Networks*, 14(1): 195-200, 2003.

Price, J. R., Gee, T. F., "Face recognition using direct, weighted linear discriminant analysis and modular subspaces", *Pattern Recognition*, 38(2): 209-219, 2005.

Thomaz, C.E., Gillies, D.F., Feitosa, R.Q., "A new covariance estimate for Bayesian classifiers in biometric recognition", *IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Image-and Video-Based Biometrics*, 14(2): 214-223, 2004.

Kwak, N., Choi, C-H, "Input feature selection for classification problems," *IEEE Trans. on Neural Networks*, 13(1): 143–159, 2002.