IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, May 14-19, 2006.

CONSTRUCTING DESCRIPTIVE AND DISCRIMINANT FEATURES FOR FACE CLASSIFICATION

Jie Yu
Department of Computer Science
University of Texas at San Antonio

Qi Tian
Department of Computer Science
University of Texas at San Antonio

ABSTRACT

Linear Discriminant Analysis (LDA) has been widely applied in the field of face classification because of its simplicity and efficiency in capturing the most discriminant features. However LDA often fails when facing the small sample set and change in illumination, pose or expression. To overcome those difficulties, Principal Component Analysis (PCA), which recovers the most descriptive/informative features in the dimension-reduced feature space, is often used in the preprocessing stage. Although there is a trend of preferring LDA to PCA in classification, it has been found that PCA may perform better than LDA in some cases, especially when the size of the training set is small. In this paper we propose a parametric framework that can unify PCA and LDA to find both discriminant and descriptive features. To avoid the exhaustive parameter searching, we incorporate a non-linear boosting process to enhance a pool of hybrid classifiers and adaptively combine them into a more accurate one. To evaluate the performance of our boosted hybrid method, we compare it to state-of-the-art LDA variants and the other PCA-LDA techniques on three widely used face image benchmark databases. The experiment results show the superior performance of our novel boosted hybrid discriminant analysis.

1. INTRODUCTION

Face classification is a computer vision application that aims at automatically classifying and retrieving face images according to user interest. It may follow a face detection process, which locates, crops and aligns regions containing humane face images from pictures. The user interest often focuses on face/non-face classification, gender classification or face recognition. The mapping between high-level human interest and low-level visual content is the primary goal that face classification tries to find through a learning process. Although face classification has been successfully applied in many fields of science and engineering, it still faces many challenging problems [1].

Small Sample Set: In most face classification applications, the number of labeled face images is limited due to the cost of human effort. If the size of the labeled training sample set is very small compared to the feature dimensionality or the samples are not representative due to the changes of illumination, pose and/or expression, it is very difficult to estimate the correct data structure and distribution. The learning-based classification may be over trained on the small sample set and overfitting may occur.

High Dimensionality: Face classification could use raw image pixel values or extracted statistical property values to represent the images in feature space. In either case the dimension of feature vector is high, ranging from tens to hundreds. The sample points are usually assumed to be from Gaussian Mixtures distribution. Traditional statistical approaches often break down during classification in the high-dimension space.

2. LDA AND PCA

2.1. Linear Discriminant Analysis

Discriminant analysis is concerned with data in which each observation comes from one of several well-defined classes or populations. The main objective is to construct rules for assigning future observation to one of the classes so as to minimize the probability of misclassification or some similar criteria. Because of its effectiveness and efficiency in classification, many discriminant analysis based approaches have been used in face classification.

Linear Discriminant Analysis (LDA) [2] is one of the most widely used discriminant analysis techniques in classification and dimension reduction. It has played a key role in many science and engineering fields such as image retrieval, face recognition and bioinformatics. Essentially LDA tries to find an optimal map W from the original high dimension space to a low dimension space, which makes the samples from different classes more separate and the ones from same class more clustered. The problem of finding the optimal W can be mathematically represented as the following maximization problem:

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \quad (1)$$

where S_B is the scatter matrix between means of different classes and S_W measures the variance of the samples in the same class.

Although LDA has gained great success in many face classification applications, research indicates that its disadvantages may prevent correct estimation of the underlying distribution of data in the projected subspace [3, 4].

Regularization

It is well known that sample-based plug-in estimates of the scatter matrices based on equation (1) will be severely biased for a small training set. If the number of the feature dimension is large compared with the number of training examples, the problem becomes ill posed, i.e., $|W^T S_W W| = 0$ in equation (1). A compensation or regularization can be simply done by adding quantities to the diagonal of the scatter matrix [5]. Although regularization is widely used to avoid the singularity, it is not theoretically justified.

Effective dimension

In LDA, W maps the original d_1-dimensional data space X to a d_2-dimensional space Δ. The maximum dimension of the projected subspace is $C-1$, where C is the number of the classes [2]. In many applications C is unknown and difficult to estimate. For those cases, C has to be assumed to be 2. Obviously such a

restriction on effective dimension may prevent more precise distribution modeling in the subspace with dimension larger than 1.

2.2. Principal Component Analysis

Principal Component Analysis (PCA) [6] is an unsupervised dimension reduction technique, which tries to represent high-dimension data with some low-dimension data by finding the "Principal Components". The Principal Components are defined as orthogonal to each other and account for the variances along each dimension. Compared with LDA which tries to capture the most discriminant features, PCA can be considered as finding the most descriptive features. Mathematically PCA can be modeled as following maximization problem:

$$W_{opt} = \arg\max_{W} \frac{|W^T S_\Sigma W|}{|W^T \cdot I \cdot W|} \qquad (2)$$

where $S_\Sigma = \sum(\mathbf{x} - m_G)(\mathbf{x} - m_G)^T$ is the covariance matrix and m_G is the grand mean of all samples $\mathbf{x}$ and I is an identity matrix.

3. HYBRID DISCRIMINANT ANALYSIS

3.1. Problem Statement

Compared with LDA, PCA is an unsupervised method and does not utilize the class information. Intuitively for a classification task one would prefer LDA to PCA. However recent research shows that in some cases PCA outperforms LDA [7]. Figure 1 shows a classical example of that situation. [7]

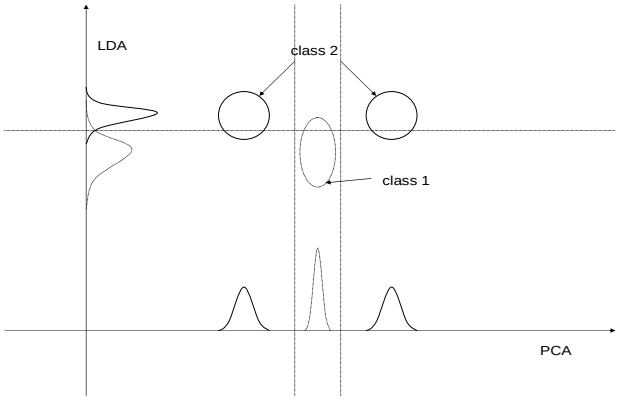


Figure 1. Example of PCA outperforms LDA

LDA assumes that the underlying structure of each class is Gaussian and the samples are evenly distributed. The assumptions are often invalid when the training sample set is small and the dimensionality of the feature space is high. Thus the most discriminant features learned by LDA often overfit on the training set. Since PCA estimate the distribution of all samples, it could provide robust performance in finding the most descriptive features. It is also worth noticing that PCA doesn't have regularization problem and the limitation on effective dimension.

3.2. Hybrid Discriminant Analysis

Motivated by the observations and analysis in Section 3.1, we propose a parametric Hybrid Discriminant Analysis (HDA) as in equation (3), which would combine LDA and PCA in a unified framework and find both discriminant and descriptive feature for a classification task.

$$W_{opt} = \arg\max_{W} \frac{|W^T[(1-\lambda)\cdot S_B + \lambda \cdot S_\Sigma]W|}{|W^T[(1-\eta)\cdot S_W + \eta \cdot I]W|} \qquad (3)$$

where λ, η are two parameters, S_Σ is the covariance matrix of all the training samples, and I is the identity matrix. The range of the parametric pair (λ, η) is from $(0,0)$ to $(1,1)$.

The different combinations of (λ, η) generate a variety of discriminant analysis with different attention on the scatter between classes and the cluster within classes. It is worth noticing that with $(\lambda = 0, \eta = 0)$ we recover LDA and with $(\lambda = 1, \eta = 1)$ we recover PCA. The alternatives to LDA and PCA offer a possible estimation to the underlying distribution of data. Table 1 summarizes three special cases of such a hybrid analysis.

Table 1: **Special cases of HDA**

(λ, η)	HDA	Note
(0, 0)	$W_{opt} = \arg\max_{W} \frac{\|W^T S_B W\|}{\|W^T S_W W\|}$	LDA
(1, 1)	$W_{opt} = \arg\max_{W} \frac{\|W^T S_\Sigma W\|}{\|W^T \cdot I \cdot W\|}$	PCA
$(\frac{1}{2}, \frac{1}{2})$	$W_{opt} = \arg\max_{W} \frac{\|W^T (S_B + S_\Sigma) W\|}{\|W^T (S_W + I) W\|}$	Trade-off

3.3. Boosted Hybrid Discriminant Analysis

As we indicated in Section 3.2, the optimal classifier of HDA could lie beyond PCA and LDA in the parametric space of (λ, η) . We have to search the whole parametric space to find the best pair (λ^*, η^*) . This will result in extra computational complexity. It is also true that the best pair found for one particular dataset could be different from that of another dataset and therefore this cannot lead to a generalization.

Based on the above analysis we adopt the idea of AdaBoost [8] and propose a boosted HDA which combines and enhances a set of HDA classifiers in the parametric space. The basic idea of Boosted HDA lies in the following two folds: 1) The incorrectly classified samples receive larger weight and the estimated distribution is biased to those samples, which forces the classifier to pay more "attention" to those difficult to learn samples. 2) The final prediction is the combination of the prediction from each classifier weighted by its classification performance, that is, the smaller the training error rate, the larger the weight.

Algorithm **Boosted Hybrid Discriminant Analysis**

Given: Training Sample set **X** and label Y

K HDA classifiers with different (λ, η)

T: The total number of runs that the classifiers will be trained for.

Initialization: weight $w_{k,t=1}(x) = 1/|X|$

Boosting

For $t = 1, \ldots, T$

For each classifier $k = 1, \ldots, K$ do

- Train the classifier on weighted samples. Note that $\sum_{x \in X} w_{k,t}(x) = 1$
  - (i) Update weighted mean μ_{all}, μ_p, and μ_n in the following way
    $\mu_{all} = \sum w_{k,t}(x) \cdot x / \sum w_{k,t}(x)$
  - (ii) Update within-class, between-class scatter matrices and co-variance matrix
- Get the probability-rated prediction on each sample $h_{k,t}(x) \in (-1,1)$
- Compute the weights of classifiers based on its classification error rate $\varepsilon_{k,t}$

$$\alpha_{k,t} = \frac{1}{2} \ln(\frac{1-\varepsilon_{k,t}}{\varepsilon_{k,t}})$$

- Update the weight of each sample

$$w_{k,t+1}(x) = w_{k,t}(x) \exp(-\alpha_{k,t} \cdot h_{k,t}(x) \cdot y)$$

End for each classifier

End for *t*

The final prediction $H(x) = sign(\sum_{k,t} \alpha_{k,t} h_{k,t}(x))$

4. EXPERIMENTS AND ANALYSIS

4.1. Comparison to variants of discriminant analysis

In the first experiment, we test the performance of our proposed methods when handling the small sample set problem. The state-of-the-art linear and nonlinear variants of discriminant analysis we test include DEM [9], kernel DEM (KDEM) [3], BDA [4] and kernel BDA (KBDA) [4] with and without regularization. They are tested as comparison to our method. In all the experiments we conducted, our boosted HDA is trained on 36 HDA classifiers with λ and η evenly sampled from 0 to 1 with step size 0.2. Simple Bayesian classifier is used in the dimension-reduced space for classification.

The data sets used in the experiments are the MIT facial image dataset (2358 images) [10] and non-face images (2958 images) from Corel database. All the face and non-face images are scaled down to 16×16 gray images and normalized feature vector of dimension 256 is used to represent each image. The size of the training set is 100, 200, 400, and 800, respectively. Compared with the feature vector dimension of 256, the training sample size is set from relatively small to relatively large. Table 2 gives the experiment results with smallest error rate in bold.

Table 2. Comparison to DEM, BDA, KDEM and KBDA

Error Rate (%)	Size of Training Set			
	100	200	400	800
DEM w/o regulation	51.3	49.43	16.7	11.2
DEM w/ regulation	10.5	19.3	15.0	9.0
BDA w/o regulation	49.2	50.1	50.0	20.85
BDA w/ regulation	34.7	25.4	18.5	19.3
KDEM	6.93	1.93	1.7	**0.5**
KBDA	3.04	2.89	2.58	1.44
HDA (λ^*, η^*)	2.3 (0.4,0.2)	1.9 (0.4,0.2)	1.8 (0.2,0)	1.3 (0.4,0)
Boosted HDA	**1.73**	**1.7**	**1.5**	0.73

Several conclusions can be drawn from result in Table 2: 1) Our proposed methods performs well when the training set size is small compared to the feature dimension. 2) Regularization is very important for sample-based estimations such as DEM and BDA while our HDA and boosted HDA implicitly release the need for regularization. Regularization can significantly improve the classification performance when the training set size is small, e.g., for DEM and BDA in average by 15%~40%. 3) When compared with the regularized DEM and BDA, the HDA performs much better than them. 4) Even when compared with the KDEM and KBDA, the boosted HDA performs better in 3 out of 4 cases. It should be noted only linear transformation is used in our hybrid analysis, but it is more efficient than nonlinear algorithms such as KDEM and KBDA. All these show the robust performance of the hybrid analysis.

4.2. Effective dimension

In the second experiment, we test the HDA versus the projection dimension using the same MIT face databases and COREL databases as in Experiment 1. Since the task is considered as two-class classification problem (face vs. non-face) in traditional LDA, the effective projection dimension is one. The feature dimension and experimental setting are same as the previous except that the size of the training dataset is fixed at 100.

Table 3. The error rate vs. the projection dimension for Boosted HDA

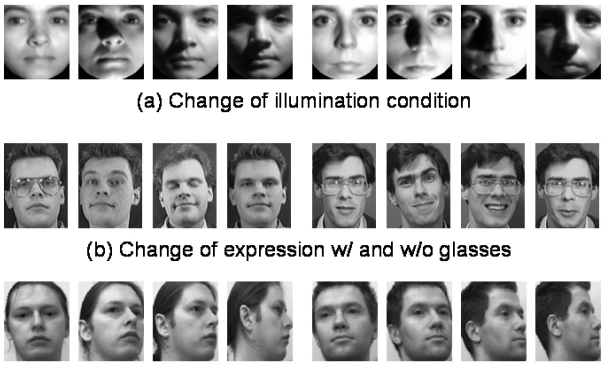
Projection Dimension	1	2	4	6	8
Error Rate(%)	1.73	1.76	1.2	**1.07**	1.1

Table 3 shows the error rate vs. the projection dimension. Clearly, the projection dimension of 6 gives the least error rate and all the higher projection dimension yields smaller error rate than that of *C-1 (C=2)*. This shows that the HDA increases the effective dimension and as a result the classification performance improves since the data structure could be more accurately modeled in a higher dimension space.

4.3 Comparison to state-of-the-art PCA-LDA related techniques

To evaluate how well our boosted Hybrid Discriminant Analysis can discover the discriminant and descriptive features of face images, we test it on three benchmark face image databases with change of illumination, expression and head pose, respectively. Harvard Face Image database consists of grayscale images of 10 persons. Each person has totally 66 images which are classified into 10 sets based on increasingly changed illumination condition [1]. The ATT Face Image database [10] consists of 400 images for 10 persons. The facial images have resolution of 92×112 with different expressions, with or without glasses under almost same illumination condition. The UMIST Face Database [11] consists of 564 images of 20 people, which covers a range of poses from profile to frontal views. We randomly chose one person's face

images as positive and the rest face images of others are considered as negative. In all experiments one third of the images in the database are randomly chosen as training set while the rest are used as test set. Figure 2 gives some example images from the databases.



(a) Change of illumination condition

(b) Change of expression w/ and w/o glasses

(c) Change of head pose

Fig. 2 Example face images

For comparison purpose, five state-of-the-art PCA and LDA related techniques are also tested on the same databases: Eigenface and Fisherface are two of the most widely used techniques in face classification [1]. S. Fidler and A. Leonardis studied the Fisherface method on how to appropriately perform PCA to facilitate LDA and propose a normalized PCA before LDA (*N.PCA-LDA*) [12]. A. Talukder and D. Casasent propose a linear combination of PCA and LDA (*L. PCA-LDA*), which takes LDA and PCA as the only two extreme cases of the combined classifier [13]. M. Wang *et al* proposed a Principal Discriminant Analysis *(PDA*) which tries to combine PCA and LDA by finding an optimal projection matrix in linear combinations of PCA and LDA projection matrix [14].

Table 4. Comparison of HDA, Boosted HDA and State-of-the-art PCA-LDA techniques

Error Rate(%)	Harvard Database			ATT Database	UMIST Database
	Subset 1	Subset 2	Subset 3		
Eigenface	1.2	5.4	25.3	28.1	38.3
Fisherface	0.7	1.4	3.7	19.5	31.2
N.PCA-LDA	0.8	1.1	2.8	17.6	32.9
L.PCA-LDA	0.6	1.3	2.2	16.3	23.6
PDA	0.9	1.2	3.4	17.9	29
HDA	0.4	0.7	2.3	11.3	27.9
Boosted HDA	**0.3**	**0.5**	**1.9**	**7.3**	**18.5**

The results are listed Table 4 with smallest error rate in bold. It is clear that our HDA performs better or comparable to other techniques while the boosted version provide best classification in all tests and more robustness to the changes of illumination, expression and pose than other techniques. Our approaches can be considered as novel compared to the previous work in that: 1) we use two parameters to control the balance between PCA and LDA. Thus our methods can search a parameter space and could find the most discriminant and descriptive features that fits the classification task and data set. And 2) in our boosted method we use AdaBoost to provide robust combination and enhance the classifier iteratively. It also avoids parameter searching which often suffers from a biased training data set.

5. CONCLUSION AND FUTURE WORK

Our novel Hybrid Discriminant Analysis provides a richer set of alternatives to LDA and PCA. As a result, it not only compensates for regularization that is afflicted by all sample-based estimation methods, but also increases the effective dimension of the projected subspace. In order to reduce the searching time, the boosted HDA is also proposed. We found it can provide two desirable properties. First, the boosted HDA can provide a unified solution to find both discriminant and descriptive features for specific application and database. Second, the weighted training schemes in boosting add indirect non-linearity and adaptivity to the linear methods and thus enhance it by iterations.

The experimental tests on benchmark image databases have shown the superior performance of HDA and boosted HDA. In the future, we are interested in continuing this research work in the following direction: 1) using fusion methods to combine HDA classifiers and 2) exploring learning based approaches to find optimal parameter settings for HDA.

6. REFERENCE

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. PAMI*, Vol. 19, No. 7, July 1997

[2] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2001.

[3] Q. Tian, Y. Wu, J. Yu, and T.S. Huang, "Self-supervised learning based on discriminative nonlinear features for image classification," *Pattern Recognition, Special Issue on Image Understanding for Digital Photographs*, Vol. 38, 2005.

[4] X. Zhou, and T.S. Huang, "Small sample learning during multimedia retrieval using biasMap," *Proc. of IEEE Conf. CVPR*, Hawaii, December 2001.

[5] J. Friedman, "Regularized discriminant analysis," *Journal of American Statistical Association*, vol. 84, 1989.

[6] I. T. Jolliffe, *Principal Component Analysis*. 2nd edition, New-York: Springer-Verlag, 2002.

[7] A. M. Martinez, and A. C. Kak, "PCA versus LDA," *IEEE Trans. PAMI*, vol. 23, no. 2, pp. 228-233, February 2001.

[8] Y. Freund, and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780, Sep., 1999.

[9] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant EM algorithm with application to image retrieval," *IEEE Conf. CVPR*, June 2000.

[10] H. A. Rowley and S. Baluja and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. PAMI*, Vol. 20, 1998.

[11] F. Samaria and A. Harter, "Parameterisation of a Stochastic Model for Human Face Identification," *IEEE Workshop on Applications of Computer Vision*, Sarasota FL, December 1994

[12] S. Fidler and A. Leonardis, "Robust LDA classification by subsampling," *IEEE Conf. CVPR Workshop*, June 2003

[13] A. Talukder and D. Casasent, "A general methodology for simultaneous representation and discrimination of multiple object classes," *Optical Engineering, special issue on "Advances in Recognition techniques"*, March 1998

[14] M. Wang, A. Perera-Lluna and R. Gutierrez-Osuna, "Principal Discriminants Analysis for small-sample-size problems: application to chemical sensing," *Proc. of IEEE SENSORS*, Vienna, October 2004