Ensemble LDA for Face Recognition

Hui Kong[1], Xuchun Li[1], Jian-Gang Wang[2], and Chandra Kambhamettu[3]

[1] School of Electrical and Electronic Engineering,
Nanyang Technological University,
50 Nanyang Ave., Singapore 639798
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[3] Department of Computer and Information Science, University of Delaware,
Newark, DE 19716-2712

Abstract. Linear Discriminant Analysis (LDA) is a popular feature extraction technique for face image recognition and retrieval. However, It often suffers from the small sample size problem when dealing with the high dimensional face data. Two-step LDA (PCA+LDA) [1, 2, 3] is a class of conventional approaches to address this problem. But in many cases, these LDA classifiers are overfitted to the training set and discard some useful discriminative information. In this paper, by analyzing the overfitting problem for the two-step LDA approach, a framework of **En**semble **L**inear **D**iscriminant **A**nalysis (E_nLDA) is proposed for face recognition with small number of training samples. In E_nLDA, a Boosting-LDA (B-LDA) and a Random Sub-feature LDA (RS-LDA) schemes are incorporated together to construct the total weak-LDA classifier ensemble. By combining these weak-LDA classifiers using majority voting method, recognition accuracy can be significantly improved. Extensive experiments on two public face databases verify the superiority of the proposed E_nLDA over the state-of-the-art algorithms in recognition accuracy.

1 Introduction

Linear Discriminant Analysis [4] is a well-known scheme for feature extraction and dimension reduction. It has been used widely in many applications such as face recognition [1], image retrieval [2], etc. Classical LDA projects the data onto a lower-dimensional vector space such that the ratio of the between-class scatter to the within-class scatter is maximized, thus achieving maximum discrimination. The optimal projection (transformation) can be readily computed by solving a generalized eigenvalue problem. However, the intrinsic limitation of classical LDA is that its objective function requires the within-class covariance matrix to be nonsingular. For many applications, such as face recognition, all scatter matrices in question can be singular since the data vectors lie in a very high-dimensional space, and in general, the feature dimension far exceeds the number of data samples. This is known as the *Small Sample Size* or singularity problem [4].

D. Zhang and A.K. Jain (Eds.): ICB 2006, LNCS 3832, pp. 166–172, 2005.

In recent years, many approaches have been proposed to deal with this problem. Among these LDA extensions, the two-stage LDA (PCA+LDA) has received a lot of attention, especially for face recognition [1, 2]. Direct-LDA (D-LDA) [5], Null-space based LDA (N-LDA) [3, 6] and Discriminant Common Vector based LDA (DCV) [7] have also been proposed. However, they all discard some useful subspaces for such-and-such reasons that prevent themselves from achieving higher recognition rate. Recently, Wang and Tang [8] presented a random sampling LDA for face recognition with small number of training samples. This paper concludes that both Fisherface and N-LDA encounter respective overfitting problem for different reasons. A random subspace method and a random bagging approach are proposed to solve them. A fusion rule is adopted to combine these random sampling based classifiers. A dual-space LDA approach [9] for face recognition was proposed to simultaneously apply discriminant analysis in the principal and null subspaces of the within-class covariance matrix. The two sets of discriminative features are then combined for recognition.

One common property of the above LDA techniques is that the image matrices must be transformed into the image vectors before feature extraction. More recently, a straightforward strategy was proposed for face recognition and representation, i.e., Two-Dimensional Fisher Discriminant Analysis (2DFDA) [10]. Different from conventional LDA where data are represented as vectors, 2DFDA adopts the matrix-based data representation model. That is, the image matrix does not need to be transformed into a vector beforehand. Instead, the covariance matrix is evaluated directly using the 2D image matrices. In contrast to the $\mathbf{S}_b$ and $\mathbf{S}_w$ of conventional LDA, the covariance matrices obtained by 2DFDA are generally not singular. Therefore, 2DFDA has achieved more promising results than the conventional LDA-based methods.

In this paper, by analyzing the overfitting problem for the two-step LDA approach, a framework of **En**semble **L**inear **D**iscriminant **A**nalysis (E_nLDA) is proposed for face recognition with small number of training samples. In E_nLDA, two different schemes are proposed and coupled together to construct the component weak-LDA classifier ensemble, i.e., a Boosting-LDA (B-LDA) algorithm and a Random Sub-feature LDA (RS-LDA) scheme. In B-LDA, multiple weighted-LDA classifiers are built where the weights of the component weak-LDA classifiers and those of the training samples are updated online based on AdaBoost algorithm. In RS-LDA, the component weak-LDA classifiers are created based on randomly selected PCA sub-features. Thus, the LDA ensemble comprises all the component weak-LDA classifiers created by B-LDA and RS-LDA. By combining these weak-LDA classifiers using majority voting method, recognition accuracy can be significantly improved.

It is well known that, in the two-step LDA methods (e.g., Fisherface), an intermediate PCA step is implemented before the LDA step and then LDA is performed in the PCA subspace. It can easily be seen that there are several drawbacks in the two-step LDA. Firstly, the obtained optimal transformation is a global and single projection matrix. Secondly, the overfitting problem is

usually inevitable when the training set is relatively small compared to the high dimensionality of the feature vector. In addition, the constructed classifier is numerically unstable, and much discriminative information has to be discarded to construct a stable classifier. There are two major reasons that arouse the overfitting problem in the two-step LDA. The first one is the existence of the non-representative training samples (or noise/unimportant data). The second is that although $\mathbf{S}_w$ is nonsingular, $N - c$ dimensionality is still too high for the training set in many cases. When the training set is small (e.g., only two/three training samples available for each subject), $\mathbf{S}_w$ is not well estimated. A slight disturbance of noise on the training set will greatly change the inverse of $\mathbf{S}_w$. Therefore, the LDA classifier is often biased and unstable. In fact, the proper PCA subspace dimension depends on the training set.

2 Ensemble LDA

Ensemble method is one of the major developments in machine learning in the past decade, which finds a highly accurate classifier by combining many moderately accurate component classifiers. Bagging [11], Boosting [12] and Random Subspace [13] methods are the most successful techniques for constructing ensemble classifiers.

To reduce the effect of the overfitting problem in the two-step LDA, we use Ensemble LDA (E_nLDA) to improve LDA based face recognition. Two different schemes are proposed to overcome the two problems that arouse the overfittings. To erase the effect brought by the existence of the nonrepresentative training samples, a boosting-LDA (B-LDA) is proposed to dynamically update the weights of training samples so that more important (more representative) training samples have larger weights and less important (less representative) training samples have smaller weights. With iteration of updated weights for the training samples, a series of weighted component weak-LDA classifiers are constructed. To remove the effect brought by the discrepancy between the size of training set and the length of feature vectors, a random sub-feature LDA (RS-LDA) is proposed to reduce such a discrepancy.

2.1 Boosting-LDA

In this section, the AdaBoost algorithm is incorporated into the B-LDA scheme (Table 1), where the component classifier is the standard Fisherface method. A set of trained weak-LDA classifiers can be obtained via B-LDA algorithm, and the majority voting method is used to combine these weak-LDA classifiers. One point deserving attention is that a so-called nearest class-center classifier instead of nearest neighborhood classifier is used in computing the training and test error. The nearest class-center classifier is similar to the nearest neighborhood classifier except that the metric used is the distance between the test data and the centers of the training data of each class not the one between the test sample and each training sample.

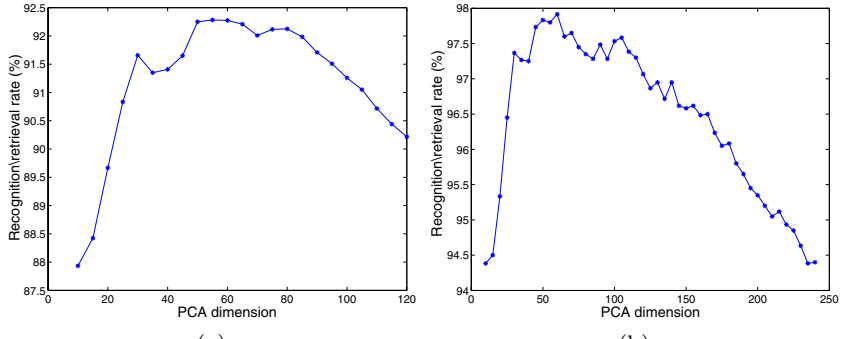
Table 1. Boosting-LDA algorithm

Algorithm: Boosting-LDA
1. Input: a set of training samples with labels $\{(\mathbf{x}_1, y_1)$, $..., (\mathbf{x}_N, y_N)\}$, Fisherface algorithm, the number of cycles T.
2. Initialize: the weight of samples: $w_i^1 = 1/N$, for all $i = 1, ..., N$.
3. Do for $t = 1, ..., T$
(1)Use Fisherface algorithm to train the weak-LDA classifier h_t on the weighted training sample set.
(2)Calculate the training error of $h_t : \epsilon_t = \sum_{i=1}^{N} w_i^t$, $y_i \neq h_t(\mathbf{x}_i)$.
(3)Set weight of weak learner $h_t : \alpha_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$.
(4)Update training samples' weights: $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(\mathbf{x}_i)\}}{C_t}$ where C_t is a normalization constant, and $\sum_{i=1}^{N} w_i^{t+1} = 1$.
4. Output: a series of component weak-LDA classifiers.

2.2 Random Sub-feature LDA

Although the dimension of image space is very high, only part of the full space contains the discriminative information. This subspace is spanned by all the eigenvectors of the total covariance matrix with nonzero eigenvalues. For the covariance matrix computed from N training samples, there are at most $N-1$ eigenvectors with nonzero eigenvalues. On the remaining eigenvectors with zero eigenvalues, all the training samples have zero projections and no discriminative information can be obtained. Therefore, for Random Sub-feature LDA, we first project the high dimension image data to the $N-1$ dimension PCA subspace before random sampling.

In Fisherface, the PCA subspace dimension should be $(N-C)$, however, Fig.1 (a) reports that the optimal result does not appear at the *120th* $(40 \times 4 - 40)$ dimension of PCA subspace when there are 4 training samples for each subject



(a) (b)

Fig. 1. Recognition/retrieval accuracy of Fisherface classifier with different dimension of PCA subspace

Table 2. E_nLDA algorithm

Algorithm: E_nLDA
1. Input: a set of training samples with labels $\{(\mathbf{x}_1, y_1)$, $..., (\mathbf{x}_N, y_N)\}$, Fisherface algorithm, the number of cycles R.
2. Do: Apply PCA to the face training set. All the eigenfaces with zero eigenvalues are removed, and $N-1$ eigenfaces $\mathbf{U}_t = [u_1, u_2, ..., u_{N-1}]$ are retained as candidates to construct the random subspaces.
3. Do for $k = 1, ..., K$: Generate K random subspaces $\{\mathbf{S}_i\}_{i=1}^{K}$. Each random subspace $\mathbf{S}_i$ is spanned by $N_0 + N_1$ dimension. The first N_0 dimensions are fixed as the first N_0 largest eigenfaces in $\mathbf{U}_t$. The remaining N_1 dimensions are randomly selected from the other $N - 1 - N_0$ eigenfaces in $\mathbf{U}_t$
4. Do: Perform B-LDA to produce T weak-LDA classifiers in each iteration of RS-LDA.
5. Output: a set of $K \times T$ component weak-LDA classifiers.

in ORL database. A similar case appears in Fig.1 (b) where the optimal PCA dimension is about *60th* instead of *240th* $(40 \times 7 - 40)$ when there are 7 training samples for each subject.

Therefore, in order to construct a stable LDA classifier, we sample a small subset of features to reduce discrepancy between the size of the training set and the length of the feature vector. Using such a random sampling method, we construct a multiple number of stable LDA classifiers. A more powerful classifier can be constructed by combining these component classifiers. A detailed description of RS-LDA is listed in Table 2.

2.3 Ensemble LDA: Combination of B-LDA and RS-LDA

Ensemble LDA (E_nLDA) can be constructed by combining B-LDA and RS-LDA. This is because that the dimension of the PCA subspace is fixed in B-LDA while the dimension of the PCA subspace is random in RS-LDA. As long as we first perform the random selection of different dimension of PCA subspace, B-LDA can be performed based on the selected PCA subspace to construct T weak-LDA classifiers. That means, if we perform K iterations of random selection (RS-LDA), $K \times T$ weak-LDA classifiers can be constructed. E_nLDA algorithm is listed in Table 2. Similarly, all the obtained component LDA classifiers can be combined via majority voting method for final classification.

3 Experiment Results

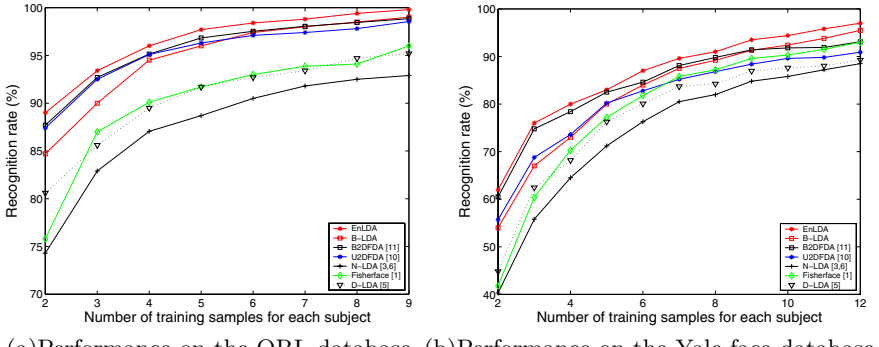
The proposed E_nLDA method is used for face image recognition/retrieval and tested on two well-known face image databases (ORL and Yale face database B). ORL database is used to evaluate the performance of E_nLDA under conditions where the pose, face expression, face scale vary. Yale face database B is used to examine the performance when illumination varies extremely.

3.1 Experiments on the ORL Database

The ORL database (http://www.cam-orl.co.uk) contains images from 40 individuals, each providing 10 different images. All images are grayscale and normalized to a resolution of 46×56 pixels. We test the recognition performance with different training numbers. k ($2 \leq k \leq 9$) images of each subject are randomly selected for training and the remaining 10-k images of each subject for testing. For each number k, 50 runs are performed with different random partition between training set and testing set. For each run, E_nLDA method is performed by training the selected fixed samples and testing on the left images. The dimension, $\{N_0, N_1\}$, for the RS-LDA is $\{15, 15\}$, $\{20, 40\}$, $\{20, 60\}$, $\{20, 80\}$, $\{20, 120\}$, $\{20, 150\}$, $\{20, 180\}$ and $\{20, 210\}$ respectively with the number of training samples for each subject changes from 2 to 9. Fig.2(a) shows the average recognition rate. From Fig.2(a), it can be seen that the performance of E_nLDA is much better than other linear subspace methods, no matter the size of training set.

3.2 Experiments on Yale Face Database B

In our experiment, altogether 640 images for 10 subjects from the Yale face database B are used (64 illumination conditions under the same frontal pose). The image size is 50×60. The recognition performance is tested with different training numbers. k ($2 \leq k \leq 12$) images of each subject are randomly selected for training and the remaining 64-k images of each subject for testing. For each number k, 100 runs are performed with different random partition between training set and testing set. For each run, E_nLDA method is performed by training the selected fixed samples and testing on the left images. The dimension, $\{N_0, N_1\}$, for the RS-LDA is $\{5, 5\}$, $\{5, 15\}$, $\{10, 20\}$, $\{10, 25\}$, $\{10, 30\}$, $\{15, 35\}$, $\{20, 40\}$, $\{30, 40\}$, $\{40, 40\}$ and $\{40, 50\}$ respectively with the number of training samples for each subject changes from 2 to 11. Fig.2(b) shows the average recognition rate. Similarly, From Fig.2(b), it can be seen that E_nLDA is the best of all the algorithms.



(a)Performance on the ORL database (b)Performance on the Yale face database B

Fig. 2. Recognition rate on the ORL database and the Yale face database B

4 Conclusions

In this paper, a framework of **En**semble **L**inear **D**iscriminant **A**nalysis (E_nLDA) is proposed for face recognition with small number of training samples. In E_nLDA, a Boosting-LDA (B-LDA) and a Random Sub-feature LDA (RS-LDA) schemes are coupled together to construct the total weak-LDA classifier ensemble. By combining these weak-LDA classifiers using majority voting method, recognition accuracy can be significantly improved. Extensive experiments on two public face databases verify the superiority of the proposed E_nLDA over the state-of-the-art algorithms in recognition accuracy.

References

1. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. on PAMI **19** (1997) 711–720
2. Swets, D., Weng, J.: Using discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996) 831–836
3. Chen, L., Liao, H., Ko, M., Lin, J., Yu, G.: A new lda-based face recognition system which can solve the samll sample size problem. Pattern Recognition (2000)
4. Fukunnaga: Introduction to Statistical Pattern Recognition. Academic Press, New York (1991)
5. Yu, H., Yang, J.: A direct lda algorithm for high-dimensional data with application to face recognition. Pattern Recognition **34** (2001) 2067–2070
6. Huang, R., Liu, Q., Lu, H., Ma, S.: Solving the small sample size problem of lda. In: Proceedings of International Conference on Pattern Recognition. (2002)
7. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 4–13
8. Wang, X., Tang, X.: Random sampling lda for face recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2004)
9. Wang, X., Tang, X.: Dual-space linear discriminant analysis for face recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2004)
10. Kong, H., Wang, L., Teoh, E., Wang, J., Venkateswarlu, R.: A framework of 2d fisher discriminant analysis: Application to face recognition with small number of training samples. In: to appear in the IEEE International Conference on Computer Vision and Pattern Recognition 2005. (2005)
11. Breima, L.: Bagging predictors. Machine Learning **10** (1996) 123–140
12. Schapire, R., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning **37** (1999) 297–336
13. Ho, T.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence (1998)