BOOSTING FACE RECOGNITION ON A LARGE-SCALE DATABASE

Juwei Lu and K.N. Plataniotis

Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto, Toronto, M5S 3G4, ONTARIO, CANADA

ABSTRACT

Performance of many state-of-the-art face recognition (FR) methods deteriorates rapidly, when large in size databases are considered. In this paper, we propose a novel clustering method based on a linear discriminant analysis methodology which deals with the problem of FR on a large-scale database. Contrary to traditional clustering methods such as K-means, which are based on certain "similarity criteria", the proposed here method uses a novel "separability criterion", to partition a training set from the large database into a set of K smaller and simpler subsets or maximal-separability clusters (MSCs). Based on these MSCs, a novel two-stage hierarchical classification framework is proposed. Under the framework, the complex FR problem on a large database is decomposed into a set of simpler ones, where traditional methods can be successfully applied. Experiments with a database containing 1654 face images of 157 subjects indicate that the error rate performance of a traditional method under the proposed framework is able to be greatly improved without significantly increasing computational complexity.

1. INTRODUCTION

For many of appearance-based face recognition (FR) methods, their success is confined to face databases with only a few hundred images. Their performance deteriorates rapidly, when they are applied to large databases [1]. The main reason is that most of these methods use linear statistical pattern recognition (SPR) methodologies, which normally require face images to follow a convex distribution. This may be approximately met in a small database with limited variations of faces. Nevertheless, since images or appearances of the face patterns may vary significantly due to differences in viewpoint, illumination and facial expression, the distribution dramatically becomes highly non convex and complex as the size of the database increases, so that the feature representation obtained by these linear methods is not capable of generalizing all of the introduced variations.

Briefly, there are two ways to address the above problem: (1) Model the complex distribution by nonlinear techniques. However, the main problem with most nonlinear methods such as Kernel Machine Based Approaches is, that it is quite difficult to find a way to optimize the involved nonlinear parameters, which significantly influence the performance. Also, the overfit is a frequent problem for these methods. Moreover, the computational complexity of the nonlinear methods is normally much higher than that of their linear counterparts. (2) Piecewise learn the complex distribution by a mixture of local linear models. This strategy is based on the principle of "divide and conquer", by which the large database is decomposed into a few smaller ones, in each of which it is hoped that the distribution of the face samples becomes concave and simple enough so that those traditional linear methods can be successfully applied to generalize a local linear distribution. Compared to the nonlinear methods, this kind of approaches is simpler, more effective and computationally attractive. Also, linear models are rather robust against noise and most likely will not overfit [2]. A mixture of view-based PCA subspaces [3], a mixture of Gaussians (for face detection) [4] and a mixture of Factor Analyzers [5], are several examples that have been applied to databases of $O(10^3)$ face images. From the designer's point of view, the central issue to these decomposition based methods is to find an appropriate criterion to partition the large training database. Surprisingly, the existing clustering techniques unanimously adopt certain "similarity criteria", based on which only those samples with certain similar properties are assigned to the same cluster or subset. However, although such criteria may be optimal in the sense of approximating real face distribution for object reconstruction, they may not be good for classification tasks. It is not hard to see that from a classification point of view, the database partition criterion should be aimed to maximize the difference or separability between classes.

In this paper, we propose a novel clustering method optimized for pattern classification. Contrary to the conventional "similarity criterion", we introduce the concept of "separability criterion" in the proposed method. A powerful tool to optimize the separability criterion is Linear Discriminant Analysis (LDA), which finds optimal discriminatory feature representation by maximizing between-class scatter of patterns. In the proposed one, the training set from the original large database is partitioned into a set of K maximal-separability clusters (MSCs) by a LDA-like technique, and the separability between classes is maximized in each MSC. We then propose a novel hierarchical classification framework, which consists of two levels of nearest neighbor classifiers (NNCs) and is able to take advantages of the obtained MSCs. The first level is composed of K NNCs corresponding to the K MSCs, each responsible for one MSC. For a given query, classification is firstly done independently in each NNC, and thus K results can be obtained. Next, a new subset with K classes is formed using the K results, and a new NNC is accordingly generalized for the subset, which results in the final classification decision. As a result, the complex FR problem on a large database is

gradually decomposed into a set of simpler ones, and traditional FR methods can be successfully applied in each MSC with high between-class separability.

2. FACE REPRESENTATION

One of central issues for FR tasks is to decide what features should be used to represent a face. Since in this work our focus is on designing a general classification framework to boost performance of existing methods on large-scale face databases, we utilize the direct LDA (D-LDA) method introduced recently in [6] for face feature extraction. The D-LDA method provides an effective solution to the so-called "small sample size problem" which exists in high-dimensional pattern recognition tasks. Another consideration to select D-LDA is, that LDA-based methods often overfit when they are applied to a large database, and good results in this case have not been reported yet [7]. For completeness, the D-LDA procedure is briefly described here. Before starting, it is important to distinguish two frequently used terms in the paper: class and cluster. Here, class means a set of face images from the same person, while cluster means a set of classes.

Given a training set containing C classes $\{\mathbf{Z}_i\}_{i=1}^{C}$, with each class consisting of a number of face images: $\mathbf{Z}_i = \{\mathbf{z}_{ij}\}_{j=1}^{C_i}$, thus totaling $L = \sum_{i=1}^{C} C_i$ face images are available. Each image is represented as a column vector of length $N(= I_w \times I_h)$, i.e. $\mathbf{z}_{ij} \in \mathbb{R}^N$, where $I_w \times I_h$ is the image size, and $\mathbb{R}^N$ denotes the N-dimensional real space. The D-LDA finds a set of optimal discriminant basis vectors, denoted as $\Psi = [\psi_1, \ldots, \psi_M]$ where $M \ll N$, by optimizing a separability criterion, or equivalently solving the eigenvalue problem: $\Psi = \arg\max_{\Psi} \frac{\left|\Psi^T S_{BTW} \Psi\right|}{\left|\Psi^T S_{WTH} \Psi\right|}$, where S_{BTW} and S_{WTH} are the between- and within-class scatter matrices of the training set respectively. For any input face image $\mathbf{z}$, its D-LDA based representation $\mathbf{y}$ can be obtained by projecting $\mathbf{z}$ into a M-dimensional feature space spanned by Ψ, where the separability of different face objects is enhanced, thus $\mathbf{y} = \Psi^T \mathbf{z}$, where $\mathbf{y} \in \mathbb{R}^M$.

3. CLUSTERING BASED ON THE SEPARABILITY CRITERION (CSC)

Motivated by the LDA algorithm and its successful application in FR tasks [8, 6], we introduce the concept of separability criterion in the proposed CSC method. Contrary to the traditional similarity criterion, the proposed one is from the standpoint of classification, which requires those classes with more different properties to be remained in the same cluster, so that classification becomes easier in the cluster. Similar to the LDA, we optimize the criterion by maximizing a total between-class scatter of all clusters.

Let Ω_k denote the k-th cluster, where $k = [1 \ \cdots \ K]$ with K: the number of clusters. Representing each class $\mathbf{Z}_i$ by its mean: $\bar{\mathbf{z}}_i = \frac{1}{C_i} \sum_{j=1}^{C_i} \mathbf{z}_{ij}$, we can define a total between-class scatter of all clusters as follows,

$$S_t = \sum_{k=1}^{K} \sum_{\bar{\mathbf{z}}_i \in \Omega_k} C_i \cdot (\bar{\mathbf{z}}_i - \mathbf{w}_k)^T (\bar{\mathbf{z}}_i - \mathbf{w}_k) \qquad (1)$$

where $\mathbf{w}_k = (\sum_{\bar{\mathbf{z}}_i \in \Omega_k} C_i \cdot \bar{\mathbf{z}}_i)/(\sum_{\bar{\mathbf{z}}_i \in \Omega_k} C_i)$ is the center of cluster Ω_k. The clustering algorithm works as follows:

Firstly, an initial partition can be formed by randomly assigning $\bar{\mathbf{z}}_i$ where $i = [1 \ \cdots \ C]$ to one of clusters $\{\Omega_k\}_{k=1}^{K}$. Secondly, we find the class mean $\hat{\mathbf{z}}_k \in \Omega_k$ which has minimal Euclidean distance to $\mathbf{w}_k$ by

$$\hat{\mathbf{z}}_k = \arg\min_{\bar{\mathbf{z}}_i \in \Omega_k} \left\{ (\bar{\mathbf{z}}_i - \mathbf{w}_k)^T (\bar{\mathbf{z}}_i - \mathbf{w}_k) \right\} \qquad (2)$$

Then, compute distances of $\hat{\mathbf{z}}_k$ to other cluster centers: $\mathbf{d}_{kh} = (\hat{\mathbf{z}}_k - \mathbf{w}_h)^T (\hat{\mathbf{z}}_k - \mathbf{w}_h)$, find the cluster $\hat{h}$ so that $\hat{h} = \arg\max_{h} \{\mathbf{d}_{kh}\}$ where $h = [1 \ \cdots \ K]$, and reassign the class represented by $\hat{\mathbf{z}}_k$ to cluster $\hat{h}$, i.e. set $\hat{\mathbf{z}}_k \in \Omega_{\hat{h}}$ if $\hat{h} \neq k$. Update the cluster centers $\mathbf{w}_k$ and the total scatter S_t, and repeat the above procedure until S_t stops increasing.

The objective in the proposed CSC method is to maximize S_t by iteratively reassigning those classes whose means have minimal distances to their own cluster centers, so that the separability between classes is enhanced gradually in each cluster.

4. HIERARCHICAL CLASSIFICATION FRAMEWORK

The original training database is partitioned into a set of smaller and simpler subsets or maximal-separability clusters (MSCs) by the CSC method. Based on these MSCs, we then propose a hierarchical classification framework (hereafter HCF), which is able to take advantages of these obtained MSCs.

The HCF consists of two levels of nearest neighbor classifiers (NNCs) as shown in Fig.1, where $(\cdot)_k^{(l)}$ denotes components corresponding to the l-th level and the k-th MSC (i.e. Ω_k), $k = [1 \ \cdots \ K]$. The first level is composed of K NNCs, each corresponding to a MSC and in charge of classification in the MSC. In the learning stage, D-LDA is applied to each MSC to find a $M_k^{(1)}$-dimensional feature space spanned by $\Psi_k^{(1)}$ (for Ω_k), and then the training images $\mathbf{z}_{ij} \in \Omega_k$ are mapped to their corresponding feature space by $\mathbf{y}_{ij} = (\Psi_k^{(1)})^T \mathbf{z}_{ij}$ respectively so that the NNCs can be performed in these feature spaces with enhanced discriminatory power.

In the FR procedure, any input query $\mathbf{z}$ is firstly fed to the first level of K NNCs, where classification is independently performed in each MSC by measuring Euclidean distances between $\mathbf{z}$'s projection $\mathbf{y}_k^{(1)} = (\Psi_k^{(1)})^T \mathbf{z}$ and pre-existing $\mathbf{y}_{ij} \in \Omega_k$ based on the nearest neighbor criterion. Thus, K classification results $\{\theta_k\}_{k=1}^{K}$ are produced, and they generalize a new subset, $\{\mathbf{Z}_{\theta_k}\}_{k=1}^{K}$, to be passed to the next level. The second level only contains one NNC, $(\mathbf{NNC})^{(2)}$, which operates in the derived subset $\{\mathbf{Z}_{\theta_k}\}_{k=1}^{K}$ and gives the final classification result. Here, we use a $\left(\sum_{k=1}^{K} M_k^{(1)}\right)$-dimensional joint feature space spanned by $\Psi^{(2)} = [\Psi_1^{(1)} \ \cdots \ \Psi_K^{(1)}]$ to take advantages of the K feature representation obtained from the first level. The input query $\mathbf{z}$ is projected to the joint feature space by $\mathbf{y}^{(2)} = (\Psi^{(2)})^T \mathbf{z}$ as well as those training samples belonging to classes $\{\theta_k\}_{k=1}^{K}$ by the same mapping operation. The fi-

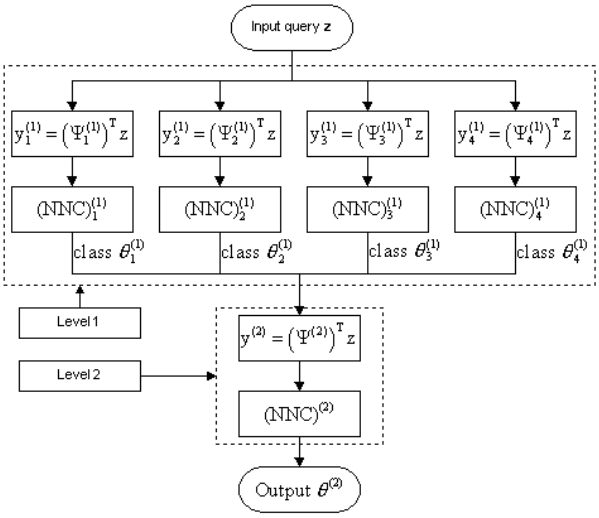

Fig. 1. The architecture of the hierarchical classification framework, where $K = 4$.

nal decision $\theta^{(2)}$ is one of classes $\{\theta_k\}_{k=1}^{K}$, and given by performing the $(\mathbf{NNC})^{(2)}$ in the joint feature space.

Actually, the operation on each pair of $(\mathbf{NNC})_k^{(1)}$ and Ω_k is an independent and standard D-LDA FR procedure similar to the one discussed in [6]. Under the HCF, the D-LDA is performed in each individual MSCs, and their results form a new subset with only K classes. Since K is usually quite small ($K = 4$ in our experiment), the final decision can be easily done in the second level of NNC. As a result, the complex classification task on a large database is gradually "divided and conquered".

It should be noted that the proposed HCF is a general framework that can be used in conjunction with other clustering methodologies. For example, clusters derived by utilizing the K-mean approach can be used instead of those clusters formed through the MSC approach. For the sake of simplicity in the sequence, we call the proposed MSC based HCF as MSC-HCF, and the K-mean based HCF as Kmean-HCF. We shall compare the two methods in the experiments.

5. EXPERIMENTAL RESULTS

5.1. A Large Mixture Face Database

A compound database with 1654 face images of 157 objects or classes is used to assess the performance of the proposed CSC and HCF schemes. The compound face database is composed of the following six databases: (1) The ORL database containing 40 distinct persons with 10 images per person. The images are taken at different times, with varying lighting conditions, facial expressions and facial details (glasses/no-glasses). All persons are in the up-right, frontal position, with tolerance for some side movement. (2) The Bern database containing frontal views of 30 persons, each person having 10 images with slight variations in the head positions, specifically two images right into the camera, two looking to the right, two looking to the left, two downwards, and two upwards. (3) The Yale database containing 15 persons. For each person, 10 of its 11 frontal view images are randomly selected for the compound database. The images are taken under ten different conditions: a normal image under ambient lighting, one with or without glasses, three images taken with different point light sources, and five different facial expression. (4) Five persons selected from the Harvard database with each person having 10 images which are subject to heavy variations in lighting in which the longitudinal and latitudinal angles of light source direction reach up to 90°. (5) The UMIST Face Database, which is a multi-view database, consisting of 575 images of 20 people, each covering a wide range of poses from profile to frontal views. (6) Since most images in the above databases are from Caucasians, a database representing Asians, composed of 179 frontal views of 47 Asian students was added to the compound database. Each person is represented by 3 or 4 images, taken at different facial expression, view points and facial details.

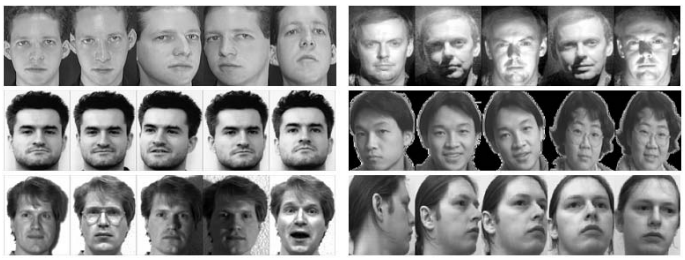

Fig. 2. Some face samples from the compound database. Left: 1st row from ORL, 2nd row from Bern, 3rd row from Yale; Right: 1st row from Harvard, 2nd row from Asian, 3rd row from UMIST.

All of the images are resized into 112×92, thus $N = 10304$, with some cropped to only contain faces before scaling. Fig.2 depicts some sample images contained in the compound database after the simple preprocessing.

5.2. Comparison With Other Methods

To start the FR experiments, the compound database is randomly partitioned into two subsets: a training set and a test set. The training set is composed of 704 images: 5 images per person are randomly chosen from the ORL, Bern, Yale and Harvard databases, 2 images per person are randomly chosen from the Asian database, and 8 images per person are randomly chosen from the UMIST database. The remaining 950 images are used to form the test set. There is no overlapping between the two. In the experiments, each run is performed on such a random partition of the database into two sets (i.e. the partition between the training and test sets is changing at every run).

In each of the runs, we firstly partition the training set into $K = 4$ clusters, and Fig.3 and Table 1 (where S_t is calculated by Equ(1)) depict quite different results obtained by the CSC method and the standard K-means. The total scatter S_t tells us how different between classes in each

Fig. 3. Means of $K = 4$ clusters obtained by the CSC method (Left) and standard K-means (Right).

Table 1. Comparison of the Total between-class scatter S_t.

Methods	1st run	2nd run	3rd run	Average
K-mean	1.0735e10	1.0537e10	1.0644e10	1.0638e10
CSC	2.0404e10	2.0432e10	2.0424e10	2.0420e10

cluster, while the cluster centers let us roughly know how different between clusters. Not surprisingly, due to different clustering criteria, S_t obtained by the CSC method is around two times of that by the K-means as shown in Table 1, while the K-means obtains more compact clusters, each having its own certain common properties such that the difference between the clusters is more obvious compared to those clusters obtained by the CSC method as shown in Fig.3.

Besides the MSC-HCF and Kmean-HCF methods, the standard D-LDA [6] (hereafter Yang-D-LDA) is also applied in the database to benchmark the performance of the formers. In the Yang-D-LDA, D-LDA is directly applied to the training set without the clustering procedure to find a M-dimensional feature space, and then a NNC is used to implement classification of the test set in the feature space. For all of the three methods, their error rates are functions of the number of feature vectors or the dimensionality of feature spaces M. Denoting ξ as the obtained minimal error rate, M_{opt} as ξ's corresponding dimensionality of feature spaces (or joint feature spaces for MSC-HCF and Kmean-HCF), recognition results of the three methods in the test sets are as shown in Table 2, where the MSC-HCF obtains the lowest error rates. Improvement of average error rates by the MSC-HCF is 6.596% to the Yang-D-LDA and 2.245% to the Kmean-HCF. Considering the large test set (totaling 950 images), 6.596% and 2.245% mean up to 62.7 and 21.3 samples respectively. Also, both of the Kmean-HCF and the MSC-HCF perform much better than the Yang-D-LDA, and this demonstrates the advantage of the proposed HCF strategy.

Table 2. Comparison of Minimal Error rates.

Experiments	Yang-D-LDA		Kmean-HCF		MSC-HCF	
	ξ (%)	M_{opt}	ξ (%)	M_{opt}	ξ (%)	M_{opt}
1st run	15.16	38	10	105	8.11	120
2nd run	14.11	24	11.16	119	7.58	102
3rd run	14.32	131	9.37	100	8.11	118
Average	14.53	64.3	10.18	108	7.93	113

6. CONCLUSION

In this paper, a general framework to boost performance of traditional FR methods in large databases was introduced. The proposed framework uses the principle of "divide and conquer", by which the original complex recognition problem is decomposed into a set of simpler ones, where those traditional methods can be successfully applied. To this end, we proposed a clustering method based on a novel "separability criterion" to partition the large training database into a set of MSCs, which can be considered as "a mixture of LDAs". Based on these MSCs, we then introduced a novel hierarchical classification framework, which is able to take advantages of these obtained clusters, to implement efficient face recognition. Experiments in a compound database indicate that the error rate performance of the traditional D-LDA method under the proposed framework is able to be greatly improved.

Again, the proposed framework is general. Besides the D-LDA, other face representation methods such as PCA or Kernel PCA can be integrated into the framework to improve their performance. We are confident that the improvement will be more impressive with size of the databases increasing.

7. REFERENCES

[1] G.D. Guo, H.J. Zhang, and S.Z. Li, ""Pairwise face recognition"," in Proceedings of The Eighth IEEE International Conference on Computer Vision, Vancouver, Canada, July 2001, vol. 2, pp. 282–287.

[2] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in Neural Networks for Signal Processing IX, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds., 1999, pp. 41–48.

[3] Alex Pentland, Baback Moghaddam, and Thad Starner, ""View-based and modular eigenspaces for face recognition"," Proc. Computer Vision and Pattern Recognition Conf., pp. 1–7, June 1994.

[4] K.-K. Sung and T. Poggio, ""Example- based learning for view-based human face detection"," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 39–51, 1998.

[5] B. J. Frey, A. Colmenarez, and T. S. Huang, ""Mixtures of local linear subspaces for face recognition"," in Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, June 1998.

[6] Hua Yu and Jie Yang, ""A direct lda algorithm for high-dimensional data with application to face recognition"," Pattern Recognition, vol. 34, pp. 2067–2070, 2001.

[7] Aleix M. Mart nez and Avinash C. Kak, ""PCA versus LDA"," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228–233, 2001.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, ""Eigenfaces vs. Fisherfaces: recognition using class specific linear projection"," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711–720, 1997.