Entropy-based Iterative Face Classification

Marios Kyperountas[1], Anastasios Tefas[1], and Ioannis Pitas[1,2]

[1] Department of Informatics, Aristotle University of Thessaloniki, Greece
[2] Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece
{mkyper, tefas, pitas}@aiia.csd.auth.gr

Abstract. This paper presents a novel methodology whose task is to deal with the face classification problem. This algorithm uses discriminant analysis to project the face classes and a clustering algorithm to partition the projected face data, thus forming a set of discriminant clusters. Then, an iterative process creates subsets, whose cardinality is defined by an entropy-based measure, that contain the most useful clusters. The best match to the test face is found when one final face class is retained. The standard UMIST and XM2VTS databases have been utilized to evaluate the performance of the proposed algorithm. Results show that it provides a good solution to the face classification problem.

Keywords: face classification, entropy, discriminant analysis.

1 Introduction

In the past several years, great attention has been given to the active research field of face classification. For the Face Recognition (FR) problem, the true match to a test face, out of a number of N different training faces stored in a database, is sought. The performance of many state-of-the-art FR methods deteriorates rapidly when large, in terms of the number of faces, databases are considered [1, 2]. Specifically, the facial feature representation obtained by methods that use linear criteria, which normally require images to follow a convex distribution, is not capable of generalizing all the introduced variations due e.g. to large differences in viewpoint, illumination and facial expression, when large data sets are used. When nonlinear face representation methods are employed, problems such as over-fitting, computational complexity and difficulties in optimizing the involved parameters often appear [1]. Recently, various methods have attempted to solve the aforementioned problems. A widely used principle that has been used is the 'divide and conquer', which decomposes a database into smaller sets in order to piecewise learn the complex distribution by a mixture of local linear models. In [1], a separability criterion is employed to partition a training set from a large database into a set of smaller maximal separability clusters (MSCs) by utilizing a variant of linear discriminant analysis (LDA). Based on these MSCs, a hierarchical classification framework that consists of two levels of nearest neighbour classifiers is employed and the match is found. The work in [3] concentrates on the hierarchical partitioning of the feature spaces using hierarchical discriminant analysis (HDA). A space-tessellation tree is generated using the most expressive features (MEF), by employing Principal Component Analysis (PCA), and the most

discriminating features (MDF), by employing LDA, at each tree level. This is done to avoid the limitations linked to global features, by deriving a recursively better-fitted set of features for each of the recursively subdivided sets of training samples. In general, hierarchical trees have been extensively used for pattern recognition purposes. In [4], an owner-specific LDA-subspace is developed in order to create a personalized face verification system. The training set is partitioned into a number of clusters and only a single cluster, which contains face data that is most similar to the owner face, is retained. The system assigns the owner training images to this particular cluster and this new data set is used to determine an LDA subspace that is used to compute the verification thresholds and matching score when a test face claims the identify of the owner. Rather than using the LDA space created by processing the full training set, the authors show that verification performance is enhanced when an owner-specific subspace is utilized.

This paper presents a novel framework, onwards referred to as EbIC (Entropy-based Iterative Classification), which applies a person-specific iterative classification that is based on an entropy measure. The clustering and discriminant analysis parameters of EbIC are heavily affected by the characteristics of the test face. This methodology is not restricted to face classification, but is able to deal with any problem that fits into the same formalism. At this point, it is imperative that two terms that are frequently used in this paper are defined: 'class' refers to a set of face images from the same person, whereas 'cluster' refers to a set of classes. The i^{th} face class is denoted by Y_i whereas the i^{th} cluster by C_i. It should be mentioned that face images of one person may even be partitioned into multiple clusters; thus, each of these clusters will contain a class from that particular person.

Initially, the training and test face vectors are projected onto a LDA-space by employing Fisher's criterion [5, 6], thus producing the most discriminant features (MDF). Subsequently, k-means is used to partition the training data into a set of K discriminant clusters C_i and the distance of the test face from the cluster centroids is used to collect a subset of K' clusters that are closest to the test face. The cardinality of this subset is set through an entropy-based measure that is calculated by making use of the discrete probability histogram. The training data that reside in these K' clusters are merged and a new MDF-space of the merged face classes is found by applying LDA and k-means is once again used to partition the data into a set of clusters in a discriminant space. This process is repeated in as many iterations as necessary, until a single cluster is selected. Then, discriminant analysis is performed on this cluster, by using the data that reside in this cluster to produce the MDF-space, and the face class that is most similar to the test face is set as its identity match.

2 Adaptive Discriminant Clustering

The EbIC algorithm is an iterative process which, during each iteration, uses an adaptive MDF space that is closely related to the characteristics of the test face. More specifically, the set of clusters to be included in the training process that will define the future MDF space are selected based on how close they are to the test face in the

current MDF space. Let us assume that an image $\mathbf{X}$ of a test face is to be assigned to one of the Y distinct classes Υ_i that lie in the training set space $\mathcal{T}$. In addition, assume that each i^{th} class in $\mathcal{T}$ is represented by N_{Υ_i} images and the total number of training images is $N_{\mathcal{T}}$. Thus, the face images that comprise the training set $\mathcal{T}$ can be represented by $\mathbf{Y}_n,\ \ n=1,\ldots,N_{\mathcal{T}}$.

2.1 Linear Discriminant Analysis

In order to linearly transform the face vectors such that they become separable, they are projected onto an MDF space. Let S_W and S_B be within-class and between-class scatter matrices [7, 8] of the training set $\mathbf{Y}$. A well known and plausible criterion is to find a projection that maximizes the ratio of the between-class scatter vs. the within-class scatter (Fisher's criterion):

$$J(\mathbf{W}) = \frac{\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}}{\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}}. \tag{1}$$

Therefore, LDA is applied on $\mathbf{Y}$ and the discriminant matrix $\mathbf{W}$ of (1) is found. The training and test feature vectors are then projected to the MDF-space by

$$\mathbf{y}_n^{'} = \mathbf{W}^{\mathrm{T}} \mathbf{y}_n,\ \ n=1,...,N_{\mathcal{T}},\ \text{and} \tag{2}$$

$$\mathbf{x}^{'} = \mathbf{W}^{\mathrm{T}} \mathbf{x}. \tag{3}$$

where $\mathbf{y}_n$ and $\mathbf{x}$ are the training and test images in the form of vectors. Each training feature vector $\mathbf{y}_n^{'}$ is stored in a column of $\mathbf{Y}^{'}$.

2.2 Clustering Using k-means

The k-means algorithm is then employed in an effort to partition the training data into the Y distinct face classes. Given a set of N data vectors, realized by $\mathbf{y}_n,\ n=1,\ldots,N,$ in the *d*-dimensional space, *k*-means is used to determine a set of K vectors in $\Re^d$, called cluster centroids, so as to minimize the sum of vector-to-centroid distances, summed over all K clusters. The objective function of *k*-means that is used in this paper employs the squared Euclidean distance and is presented in [9]. After the K cluster centroids are found, e.g. $K=Y$, a single vector $\mathbf{y}_n^{'}$ can be assigned to the cluster with the minimum vector-to-cluster-centroid distance, among the Y distances that are calculated. The distance between each training feature vector and the Y centroids, $\boldsymbol{\mu}_i$, can be calculated by the Euclidean distance measure:

$$D_i^n\left(\mathbf{y}_n^{'}, \boldsymbol{\mu}_i\right) = \left\| \mathbf{y}_n^{'} - \boldsymbol{\mu}_i \right\|,\quad i=1,...,Y. \tag{4}$$

2.3 Entropy-based Generation of MDF-Spaces

Let us consider a set of K clusters, or partitions, in the data space $\mathcal{T}$. The surrounding *Voronoi region* of the *i*-th cluster is denoted as V_i. Theoretically, the a-priori probability for each cluster to be the best matching one to any sample vector $\mathbf{x}$ of the feature space is calculated as such, if the probability density function $p(\mathbf{x})$ is known:

$$P_i = P(\mathbf{x} \in V_i) = \int_{V_i} p(\mathbf{x}) d\mathbf{x}. \tag{5}$$

For discrete data, the discrete probability histogram can replace the continuous probability density function:

$$P_i = P(\mathbf{x} \in V_i) = \frac{\#\{j \mid \mathbf{x}_j \in V_i\}}{N}, \tag{6}$$

where $\#\{\cdot\}$ represents the cardinality of a set and N the size of the training data set whose members are $\mathbf{x}_j$, $j = 0,1,\ldots,N-1$. Let us consider a set of K partitions in the training data space $\mathcal{T}$ and their distribution $P = (P_1, P_2, \ldots, P_K)$. The entropy, a commonly used measure that indicates the randomness of the distribution of a variable, can be defined as [10]:

$$H = H(P) = -\sum_{i=1}^{K} P_i \log_2 P_i \tag{7}$$

An 'ideal' data partitioning separates the data such that overlap between partitions, e.g. the class overlap, is minimal, which is equivalent to minimizing the expected entropy of the partitions over all observed data.

In this paper, the entropy-based measure is calculated in a new data space $\mathcal{T}' \subset \mathcal{T}$, which consists of a subset that retains K' of the total K clusters that are generated by the *k*-means algorithm. Let us assume that the K' clusters contain Y' face classes. A needed assumption used to calculate the entropy is that a true match to the test face class $\mathcal{X}$ exists within the $\mathcal{T}'$ space. Let the probability for the i-th face class $\mathcal{Y}_i'$, that is now contained in $\mathcal{T}'$, to represent a true match for $\mathcal{X}$ be $P_i = p(\mathcal{Y}_i' \mid \mathcal{X})$. Since the prior probabilities $p(\mathcal{Y}_i' \mid \mathcal{X})$ are unknown, they can be defined using the discrete probability histogram, as in (6), as:

$$P_i = p(\mathcal{Y}_i' \mid \mathcal{X}) = \frac{N_{\mathcal{Y}_i'}}{N_{\mathcal{Y}'}}, \tag{8}$$

where $N_{\mathcal{Y}'}$ is the total number of face images contained in $\mathcal{T}'$, and $N_{\mathcal{Y}_i'}$ is the number of times that class i is represented in $\mathcal{T}'$, e.g. $N_{\mathcal{Y}_i'}$ different images of the person associated with class i are contained in $\mathcal{T}'$. The value of K' is limited by the threshold T_H applied on the entropy value, which, in order to guarantee a low computational cost is approximated by substituting (8) into (7), so that the following is satisfied:

$$-\sum_{i=1}^{K'} \frac{N_{\Upsilon_i'}}{N_{\Upsilon'}} \log_2\left(\frac{N_{\Upsilon_i'}}{N_{\Upsilon'}}\right) \leq T_H \,. \tag{9}$$

The approximated entropy values are used to guarantee that at each step of the EbIC algorithm an easier, in terms of the ability to achieve better separation among the classes, classification problem is defined. Threshold T_H is applied on the entropy value H to limit the number of different classes that Υ' will contain. Essentially, this is done by limiting the number of clusters K' that comprise Υ'.

In the new MDF-space, created using the face data from the K' clusters, LDA will attempt to discriminate the different classes found in each of the K' clusters. This enables the algorithm to formulate a clustering process that considers possible large variations in the set of images that represent each face class. For example, a portion of the set of images that corresponds to the i^{th} training person may present this person having facial hair, whereas others as not having facial hair. If these variations are larger than identity-related variations, then they are clustered into disjoint clusters. Thus, the match with the subset of the training images of class i whose appearance is most similar to the test face is considered, so the best match can be found.

3 Experimental Results

In this section, the classification ability of EbIC is investigated by observing FR experiments using data from the XM2VTS and UMIST databases. Essentially, as in most FR applications, the classification experiments that are carried out fall under the small sample size (SSS) problem where the dimensionality of the samples is larger than the number of available training samples per subject [11, 12]. The performance of EbIC is presented for various degrees of how severe the SSS problem is. This is done by providing recognition rates for experiments where each face class Υ_i is represented by the smallest to the largest possible number of training samples, N_{T}. Since EbIC employs discriminant analysis, the smallest possible value is 2. The largest possible value of training samples for each face class Υ_i is determined by the number of available images in this class, N_{Υ_i}, and by considering that at least one of these samples needs to be excluded in order to be able to evaluate the recognition performance for that particular class. The remaining images that do not comprise the training set are used to test the performance of EbIC, thus, they constitute the test set. The training and test sets are created by random selection on each set of the N_{Υ_i} images of each face class. To give statistical significance to our experiments, this random selection is repeated N_{R} times, thus, N_{R} recognition rates are averaged in order to obtain the final recognition rate R_{rec}.

The UMIST database consists of $K = 20$ different face classes, each of which is represented by at least $N_{\Upsilon_i} = 19$ images. Consequently, 17 recognition rates were

derived for training sets that contained $N_T = 2,\ldots,18$ images from each of the 20 face classes. Each corresponding rate was the average out of $N_R = 10$ repetitions. The XM2VTS database consists of $K = 200$ different face classes, each of which is represented by $N_{\gamma_i} = 8$ images. The number of clusters K' that are retained at each clustering level is selected by using (9). The face classes residing in the final cluster are projected to the MDF-space that is created by processing only this specific set of data. The face class that is closest to the test face in this MDF-space is selected as the true match in identity. Table 1 reports the mean recognition rates, R_{rec}, obtained for FR experiments carried out on both face databases, for $N_R = 10$ independent runs. The entropy-based measure, which is utilized to determine the number of clusters that should be retained, leads to more accurate results than the ones in [13], where a power function that converges to unity was used instead.

4 Conclusion

A novel face classification methodology that employs person-specific adaptive discriminant clustering is proposed and its performance is evaluated. By making use of an entropy-based measure, the EbIC algorithm adapts the coordinates of the MDF-space with respect to the characteristics of the test face and the training faces that are more similar to the test face. Thus, the FR problem is broken down to multiple easier classification tasks, in terms of achieving linear separability. The performance of this method was evaluated on standard face databases and results show that the proposed framework provides a good solution for face classification.

Acknowledgments

This work has been performed within the COST Action 2101 on Biometrics for Identity Documents and Smart Cards, and partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211471 (i3DPost).

Table 1. Mean recognition rates for various numbers of training samples per subject

UMIST				XM2VTS	
Training Samples N_T	R_{rec} (%)	Training Samples N_T	R_{rec} (%)	Training Samples N_T	R_{rec} (%)
2	58.9	11	96.6	2	31.8
3	81.1	12	96.9	3	92.1
4	89.8	13	97.0	4	95.9
5	91.2	14	97.2	5	96.7
6	91.8	15	97.7	6	97.2
7	94.1	16	97.9	7	**98.6**
8	94.5	17	98.3		
9	94.6	18	**99.1**		
10	95.4				

References

1. J. Lu and K.N. Plataniotis, "Boosting face recognition on a large-scale database", *in Proc. IEEE Int. Conf. on Image Processing (ICIP'02)*, Rochester, New York, USA, September 22-25, 2002.
2. G.D. Guo, H.J. Zhang, and S.Z. Li, "Pairwise face recognition", *in Proc. 8th IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 282-287, Vancouver, Canada, 2001.
3. D.L. Swets and J. Weng, "Hierarchical discriminant analysis for image retrieval", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 386-401, May 1999.
4. H.-C. Liu, C.-H. Su, Y.-H. Chiang; Y.-P. Hung, "Personalized face verification system using owner-specific cluster-dependent LDA-subspace", *in Proc. of the 17th Int. Conf. on Pattern Recognition (ICPR'2004)*, vol. 4, pp. 344-347, Aug. 23-26, 2004.
5. J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using LDA based algorithms", *IEEE Trans. on Neural Networks*, vol. 14, no. 1, pp. 195-200, Jan. 2003.
6. M. Kyperountas, A. Tefas, and I. Pitas, "Methods for improving discriminant analysis for face authentication", *in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, March 2005.
7. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, New York, NY: Academic Press, 1990.

8. M. Kyperountas, A. Tefas, and I. Pitas, “Face verification using locally linear discriminant models”, *in Proc. IEEE Int. Conf. on Image Processing*, vol. 4, pp. 469-472, San Antonio, 16-19 Sep., 2007.
9. F. Camastra and A. Verri, “A novel kernel method for clustering” , *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 801-805, May 2005.
10. M. Koskela, J. Laaksonen, and E. Oja, “Entropy-based measures for clustering and SOM topology preservation applied to content-based image indexing and retrieval”, *in Proc. 17th Int. Conf. on Pattern Recognition, vol.* 2, pp. 1005-1009, 2004.
11. J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, “Selecting kernel eigenfaces for face recognition with one training sample per subject”, *in Proc. IEEE Int. Conf. on Multimedia and Expo,* pp. 1637-1640, July, 2006.
12. M. Kyperountas, A. Tefas, and I. Pitas, “Weighted piecewise LDA for solving the small sample size problem in face verification”, *IEEE Trans. on Neural Networks*, vol. 18, no. 2, pp. 506-519, March 2007.
13. M. Kyperountas, A. Tefas, and I. Pitas, “Face recognition via adaptive discriminant clustering”, *in Proc. IEEE Int. Conf. on Image Processing*, pp. 2744-2747, San Diego, Oct. 2008.