ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 389-394

Facial Image Retrieval Using Sequential Classifiers

Srinivas Gutta and Harry Wechsler

Department of Computer Science
George Mason University
Fairfax, VA 22030
{sgutta, wechsler}@cs.gmu.edu

*This paper considers hybrid architectures for facial image retrieval using gender, ethnic origin and identity as retrieval cues. The hybrid approach consists of an ensemble of RBF networks and inductive decision trees (DT). Experimental Cross Validation (CV) results on a collection of 3006 frontal face images from the FERET database (corresponding to 1009 subjects) yield - (**a**) an average accuracy rate of 96%, 94% and 92% on the gender, ethnic and identity tasks respectively when their corresponding classifiers were used individually and (**b**) an average cumulative accuracy rate of 86% for the case, when the ethnic and identity classifiers were used jointly and 83% when all the three classifiers were used together.*

1. Introduction

As more and more image data are acquired, managing and manipulating them as images becomes an important issue to be resolved before we can take full advantage of their information content. Image database and visual information system technologies have become major efforts to address these issues [1]. By and large these systems take a text-based approach to indexing and retrieval. A good survey of such text-based query formation and matching techniques can be found in [2]. However there are several problems inherent in systems which are exclusively text-based. First, automatic generation of descriptive key words or extraction of semantic information to build classification hierarchies for broad varieties of images is beyond the capabilities of current machine vision techniques.

An alternative to relying on text is to work with descriptions based on properties, which are inherent in the images, themselves. The idea behind this is to, retrieve visual data by a query based on the visual content of an image itself. For example, given a facial database of criminals, a surveillance agent may want to retrieve specific images based on the characteristics of the image, such as, gender, ethnic origin, identity etc. Formulating a query for such a search would involve selecting one or more representative examples and then searching for images which resemble those examples.

No Facial Image Retrieval (FIR) systems exist in the literature that can accomplish the tasks of finding the gender, ethnicity and identification in one composite system. The only other FIR system appearing in the literature is that of MIT's Photobook [3] system employed for identification purposes only. They report an accuracy of 99 % on a database of 575 frontal FERET images. As opposed to Photobook, we describe a FIR system using a hybrid classifier approach on a database of 3006 images that can sequentially accomplish the tasks of finding the gender, ethnic origin and identity.

2. Classifier Architecture

The sequential classifier architecture is shown below in Fig. 1. The sequential classifier consists of four levels. At the first level we have a hybrid classifier consisting of ERBF and DT (see Section 3.3). At the second level we have two nodes namely, male and female. At each of these two nodes we have again two hybrid classifiers to classify an image based on four ethnic criteria namely that of, Caucasian, African, Asian and Oriental. Thus at the third level we would have eight nodes. At each of these nodes we have an RBF classifier. At the fourth level are the leaves, which are dependent on the number of unique individuals the RBF classifier has been trained on. The sequential classifier has been designed to perform either in an autonomous or in a semi-autonomous fashion. For example, the user can either give only the image example without any additional information or he can provide already known information about the image like gender, ethnic origin, etc. In the first case the image example would be sequentially classified by traversing the classifier tree while, in the other case the classification would directly jump to the stage where the FIR system needs to find additional description for that image.

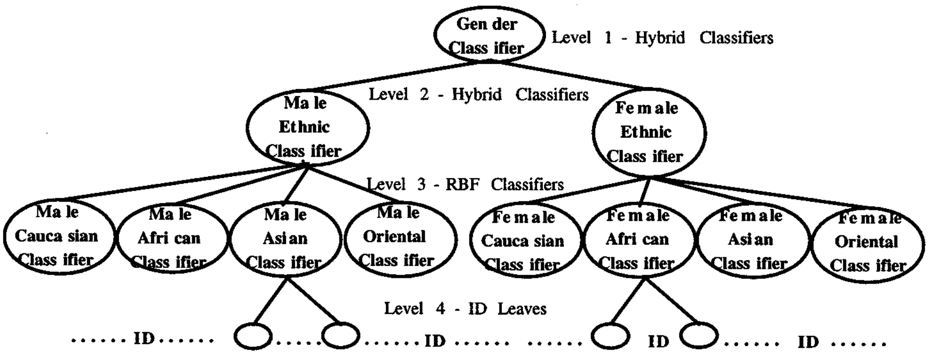

Figure 1. Sequential Classifier Architecture

3. Hybrid Classifiers

The hybrid classifier consists of an ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT). The reasons behind using RBFs are their ability to cluster similar images before classifying them. Using the RBF outputs, the decision trees (DT) implement the symbolic stage, because they provide for flexible and adaptive thresholds. The ensemble of radial basis functions (ERBF) implements the equivalent of query by consensus and they are trained on data reflecting the inherent variability of the input. Ensembles are defined in terms of their specific topology (connections and RBF nodes) and the data they are trained on. Both original data and possible distortions caused by geometrical changes and blur would induce robustness to those very distortions via generalization. As suitable decision boundaries ('thresholds') are hard to establish, this issue is addressed by interfacing a symbolic component to the ensemble of networks. The symbolic component is trained on the outputs produced by ERBF by choosing a random set of positive and negative examples corresponding to the respective classes ('genders or ethnic origins') to be learned, and yields decision boundaries defined as decision trees (DT).

3.1. Ensemble of Radial Basis Functions (ERBF)

An RBF classifier has an architecture very similar to that of a traditional three-layer back-propagation network [4]. Connections between the input and middle layers have unit weights and, as a result, do not have to be trained. Nodes in the middle layer, called BF nodes, produce a localized response to the input using Gaussian kernels.

The RBF input consists of n normalized face images fed to the network as 1D vectors. The hidden (unsupervised) layer, implements an enhanced k-means clustering procedure, where both the number of Gaussian cluster nodes and their variance are dynamically set. The number of clusters varies, in steps of 5, from 1/5 of the number of training images to n, the total number of training images. The width of the Gaussian for each cluster, is set to the maximum {the distance between the center of the cluster and the farthest away member - within class diameter, the distance between the center of the cluster and closest pattern from all other clusters} multiplied by an overlap factor o, here equal to 2. The width is further dynamically refined using different proportionality constants h. The output (supervised) layer maps face encodings ('expansions') along such a space to their corresponding class and finds the corresponding expansion ('weight') coefficients using pseudoinverse techniques. In our case the output layer consisted of two nodes corresponding to two classes 'male' and 'female'. Note that the number of clusters is frozen for that configuration (number of clusters and specific proportionality constant h) which yields 100 % accuracy when tested on the same training images.

For a connectionist architecture to be successful it has to cope with the variability available in the data acquisition process. One possible solution to the above problem is to implement the equivalent of query by consensus using ensembles of radial basis functions (ERBF). Ensembles are defined in terms of their specific topology (connections and RBF nodes) and the data they are trained on. Specifically, both original data and distortions caused by geometrical changes and blur are used to induce robustness to those very distortions via generalization. Two different versions of ERBF are proposed and described below.

3.1.1. ERBF1

The first model integrates three RBF components and it is shown in Figure 2. Each RBF component is further defined in terms of three RBF nodes, each of which specified in terms of the number of clusters and the overlap factors. The overlap factors ***o***, defined earlier, for the RBF nodes RBF(11, 21, 31), RBF(12, 22, 32), and RBF(13, 23, 33) are set to the standard 2, 2.5, and 3, respectively. The same RBF nodes were trained on original images, and on the same original images with either some Gaussian noise added or subject to some degree of geometrical ('rotation'), respectively. The intermediate nodes C_1, C_2, and C_3 act as buffers for the transfer of the normalized images to the various RBF components. Training is performed until 100% recognition accuracy is achieved for each RBF node. The nine output vectors generated by the RBF nodes are passed to a *judge* who would make a decision on whether the probe ('input') belongs to the male class or not. The specific decision used is similar to that of majority voting namely, {if *majority* (5) of the 9 outputs agree on a particular class then that probe belongs to that class}.

3.1.2. ERBF2

ERBF2 is derived from ERBF1 by increasing the number of images (3) used to train each class and by decreasing the number of RBF nodes from nine to three (Figure 3). Each RBF node is now trained on a mix of face images consisting of original ones and their distorted variations. The overlap factors, training remain the same as used for ERBF 1. During testing, nine output classes are generated, corresponding to the Cartesian product between the kind of input {original, variation with Gaussian noise, variation with rotation} and the kind of RBF node, and they are passed to a *judge*. The specific decision for gender classification remains the same as it was the case for ERBF1.

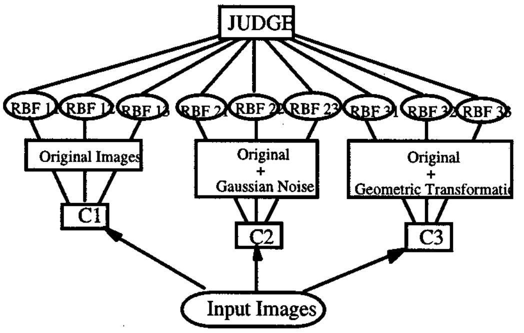

Fig. 2. ERBF1 Architecture

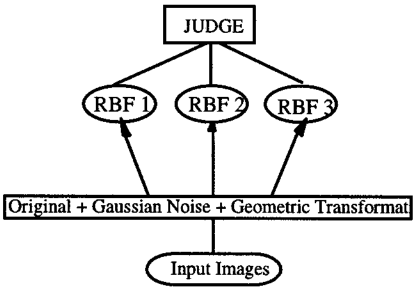

Fig. 3. ERBF2 Architecture

3.2. Decision Tree (DT)

The basic aim of any concept-learning symbolic system is to construct rules for classifying objects given a *training set* of objects whose class labels are known. The objects are described by a fixed collection of attributes, each with its own set of discrete values and each object belongs to one of two classes. The rules derived in our case will form a decision tree (DT).

The decision tree employed is Quinlan's C4.5 [5]. C4.5 uses an information-theoretical approach, the entropy, for building the decision tree. It constructs a decision tree using a top-down, divide-and-conquer approach: select an attribute, divide the training set into subsets characterized by the possible values of the attribute, and follow the same procedure recursively with each subset until no subset contains objects from both classes. These single-class subsets correspond then to leaves of the decision tree. The criterion that has been used for the selection of the attribute is called the *gain ratio criterion.*

3.3. ERBF (1,2) and DT (C4.5) Hybrids

Inductive learning, as applied to building a decision tree requires a special interface for numeric-to-symbolic data conversion. The ERBF output class vector (X_1, ... ,X_9) chosen for training for example is tagged as 'MALE' (positive example) or 'FEMALE' (negative example) for the gender task. The input from ERBF to C4.5 consists of a string of learning (positive and negative) events, each event given as a vector of discrete attribute values. Training involves choosing a random set of positive events and a random set of negative events. The C4.5 builds the classifier as a decision tree whose structure consists of *leaves*, indicating class identity, or *decision nodes* that specify some test to be carried out on a single attribute value, with one branch for each possible outcome of the test.

The decision tree is used to classify an example by starting at the root of the tree and moving through it until a leaf is encountered. At each non-leaf a decision is evaluated, the outcome is determined, and the process moves on.

4. Experiments

Face processing usually starts through the detection of a pattern as a face and boxing it, proceeds by normalizing the face image to account for geometrical and illumination changes using information about the box surrounding the face and/or eyes location, and finally it processes the face using appropriate image representation and classification algorithms. The results reported later on assume that the patterns corresponding to face images have been detected and normalized [6].

The database for our experiments comes from the standard FERET facial database [7] and comprises of 3,006 frontal images of resolution 256 x 384 encoded in 256 gray scale levels and corresponds to 1,009 unique subjects. It is to be noted that, each subject appears in the database as a pair ('fa' and 'fb'). An image of a subject taken at a different date is called a duplicate. The database also includes 494 subject duplicate pairs. The images were acquired within a span of 3 years. The total number of images of gender 'Male' is 1906 and of gender 'Female' is 1100 images. The total number of images of gender 'Male' corresponding to various ethnic origins (Caucasian, Asian, Oriental and African) is 1212, 256, 330 and 108 respectively. Similarly the total number of images of gender 'Female' corresponding to various ethnic origins (Caucasian, Asian, Oriental and African) is 722, 106, 144 and 128 respectively.

At the first level of the hierarchical classifier, namely that of the gender classifier we used a total of 30 random male (15 Caucasian, 5 Asian, 5 Oriental and 5 African) 'fa or fb' images and 30 female 'fa or fb' images of similar composition for training the connectionist component. An additional 60 random 'fb or 'fa' images corresponding to both male and female of similar composition were used to train the symbolic component.

At the second level of the hierarchical classifier we have two nodes namely that of, male and female ethnic classifiers. For the male ethnic classifier we again used a total of randomly selected 30 'fa or fb' male images of similar composition as explained above for training the connectionist component and another 30 randomly selected 'fb or fa' images of similar composition for training the symbolic component. The female ethnic classifier was trained similar to the male ethnic classifier by using two sets of randomly selected 30 'fa or fb' female images each for training the connectionist and symbolic components respectively.

At the third level of the hierarchical classifier we have an "ID" task. This was performed by using an RBF network analogous to ERBF2 network (see Section 3.1.2). Specifically experiments were now performed by training the RBF network with the number of unique subjects left. In the case when, the unique subjects image pairs are no longer available but, their duplicates are present then, their duplicates are substituted in place of the unique subject image pairs. Please note that the images used for training the hierarchical classifier for the first two levels was performed by randomly selecting the images corresponding to the unique subjects only. This process was repeated for a total of 20 times (CV cycles). At each CV cycle, the left over unique subject pairs and the available duplicate pairs not present in the unique subject pairs set was used for training at the third level of the hierarchical classifier.

Type of Task	Correct Classification (%)	Mis-Classification (%)
Gender	96	4
Ethnic	94	6
Identity	92	8
Case 1	90	10
Case 2	86	14
Case 3	83	17

Table 1. Average Results

Gender (Ethnic) Task	Correct Classification (%)	Mis-Classification (%)
RBF	70 (64)	30 (36)
ERBF1	79 (73)	21 (27)
ERBF2	82 (80)	18 (20)
ERBF1 with C4.5	90 (88)	10 (12)
ERBF2 with C4.5	96 (94)	4 (6)

Table 2. Classifier Comparison for Gender and Ethnic Tasks

Table 1 describes the average results over 20 cycles for all the classifiers at a particular level when used individually as well as the average cumulative results when the classifiers were used in a sequence. Specifically, (a) 'rowl' gives the average result for the case when the gender classifier was used individually, (b) 'row2' gives the result when the ethnic classifier was used individually, (c) 'row3' gives the result when the identity classifier was used individually, (d) 'row4' (casel) gives the average cumulative result when the gender and ethnic classifier were used jointly, (e) 'row5' (case2) gives the average cumulative result when the ethnic and identity classifier were used jointly and (f) 'row6' (case3) gives the average cumulative result when the gender, ethnic and identity classifier were used jointly. Table 2 above gives the average results over 20 cycles when comparison is made between the various classifiers for the gender and ethnic tasks respectively.

From the results reported in the above tables one can observe that when the connectionist ERBF model is coupled with an Inductive Decision Tree - C4.5 - the performance improves over the case when only the connectionist (ERBF) module is used. Specifically, we observe that the classification rate increased on the average by 14 %. Another observation one can make is that the ERBF2 model is better than the ERBF1 model. The plausible explanation is that training using more examples ('multiple displays') leads to better performance. We also note that the ERBF models reported above outperform single RBF networks. The reason for this last observation comes from ERBF models implementing the equivalent of a 'query by consensus' paradigm. Improved ERBF (vs RBF) performance can be also traced to the fact that the range for test images is (slightly) different from those encountered during training and that using more but slightly different nets ('referees') adds to the strength of the decision.

5. Conclusions

We have proposed in this paper a FIR architecture, which can sequentially classify images based on their gender, ethnic orientation and identity using the hybrid classifier approach and shown its feasibility using a collection of 3006 face images from the standard FERET data base (corresponding 1009 subjects). The hybrid architectures, consisting of an ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT), combines the merits of 'holistic' template matching with those of 'discrete' features based classifiers using both positive and negative learning examples.

Acknowledgements: This work was partly supported by the US Army Research Lab under contract DAAL01-93-K-0099.

6. References

[1] Furht B., Smoliar, S. W. and Zhang H. (1995), *Video and Image Processing in Multimedia Systems*, Kluwer Academic Publishers.

[2] Ang Y. H., Narasimhalu A. D. and Hawamdeh S. (1993) Image Information Retrieval Systems, in Handbook of Pattern Recognition and Computer Vision, 719-739, World Scientific Press.

[3] Pentland A., Picard R. W. and Sclaroff S. (1996) Photobook: Content-Based Manipulation of Image Databases, *Journal of Computer Vision* 18, No. 3, 233-254.

[4] Lippmann R. P. and Ng K. (1991), A Comparative Study of the Practical Characteristic of Neural Networks and Pattern Classifiers, Tech. Report 894, Lincoln Labs, MIT.

[5] Quinlan J.R. (1986), *C4.5 - Programs for Machine Learning*, Morgan Kaufmann.

[6] Gutta S., Huang J., Kakkad V., and Wechsler, H. (1998) Face Surveillance, *International Conférence on Computer Vision (ICCV)*, Bombay, India.

[7] Phillips P. J., Wechsler H., Huang J. and Rauss P. The FERET Database and Evaluation Procedure for Face Recognition Algorithms, *Image and Vision Computing*, 1998 (to appear).