International Conference on Neural Networks (ICNN'96)

Face Recognition Using Hybrid Classifier Systems

Srinivas Gutta and Harry Wechsler
Department of Computer Science,George Mason University, Fairfax, VA-22030
Telphone: 703-993-1533, Fax: 703-993-3729
{sgutta, wechsler}@cs.gmu.edu, http://chagall.gmu.edu/FORENSIC

ABSTRACT

This paper considers hybrid classification architectures and shows their feasibility on large data bases consisting of facial images. Our architecture, consists of an ensemble of connectionist networks - radial basis functions (RBF) - and decision trees (DT). This architecture enjoys robustness via (**i**) consensus provided by ensembles of RBF networks, and (**ii**) categorical classification using decision trees. The results reported in this paper on automatic face recognition using the FERET data base are encouraging when one considers that the size of our test bed is in excess of 350 subjects and the great variability of the data base. In addition we have also demonstrated the feasibility of our approach on queries aimed at the retrieval of frames ('images') using contextual cues.

1. Introduction

Face recognition is a difficult task mostly because of the inherent variability of the image formation process in terms of image quality and photometry, geometry, occlusion, change, and disguise. Two recent surveys on face recognition discuss these challenges in some detail[1][2]. All face processing systems available today can only perform on restricted data bases of images in terms of size, age, gender, and/or race, and they further assume well controlled environments. There are additional degrees of variability ranging from those assuming that the position/cropping of the face and its environment (distance and illumination) are totally controlled, to those involving little or no control over the background and viewpoint, and eventually to those allowing for major changes in facial appearance due to factors such as aging and disguise (hat and/or glasses).

This paper considers hybrid classification architectures and shows their feasibility on large data bases consisting of facial images. The hybrid connectionist - symbolic architectures are based on principles of hierarchical organization and they implement successive levels of abstraction in terms of information granularity. The architectures, consisting of an ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT), combine the merits of 'holistic' template matching with those of 'discrete' features using both positive and negative learning. This architecture enjoys robustness via (**i**) consensus provided by ensembles of RBF networks, and (**ii**) categorical classification using decision trees.

2. Background

As hierarchical modes of processing, also referred to as hybrid, have been shown to be fundamental for efficient and intelligent behavior, it should come as no surprise that the same should apply to learning and classification tasks as well. The concept of reductionism is a common practice in the development of intelligent systems - to design solutions to complex problems through a stepwise decomposition of the task into successive modules. Typically, in hybrid systems, reflexive tasks are assigned to the connectionist subsystem and deliberative tasks to the symbolic subsystem.

An early example of using ensembles of neural networks is due to Hamshire and Waibel[3]. The Meta - Pi classifier is a connectionist pattern classifier that consists of a number of source-dependent sub networks that are integrated by a combinational Time Delay Neural Network (TDNN) superstructure. The TDNN combines the outputs of the modules, trained independently, in order to provide a global classification. An important consideration for developing classification schemes, hybrid or otherwise, is the way the system is trained. Valid assumptions about performance in terms of generalization ability can be made only when the data used to train the system is representative of that likely to be encountered later on during field operation and that appropriate strategies, such as crossvalidation (CV), would partition the data into training, tuning, and testing sets, respectively. It is well known that a combination of many different predictors can improve predictions and thus enhance generalization ability. Query by committee[4] or reaching consensus using 'democracy'[5] are examples of such an approach and they are usually implemented using ensembles of networks If one were to define ambiguity as the variation of the output ensemble members over unlabeled ('test') data, the disagreement between networks can be quantified and corrective training take place[6]. An active learning scheme, corresponding to corrective training, would retrain on those test examples for whom the ensemble strongly disagrees.

3. Face Recognition

Identifying people seems straightforward -- people do it all the time in business and in social encounters. Automated recognition, however, requires computer systems to look through many stored sets of characteristics ('the gallery') and pick the one that matches best those features of the unknown individual ('the probe'). In most practical scenarios there are two possible recognition tasks to be considered[7]:

- MATCH: An image of an unknown individual is collected and the identity is found searching a large set of images. The size of the gallery, as well as the number of probes is typically 50-100, but in many applications it would be possibly in excess of one thousand.

- SURVEILLANCE: Rather than identifying a person, the system is now involved with verification and checks if a given probe belongs to a relatively small gallery, sometimes labeled as a set of intruders. The probe can range from individual faces to known faces displaying specific characteristics. The surveillance system is usually flooded with a large set of face images (e.g. video frame retrieval and/or airport security) and most of the faces, if not all of them, correspond to false positives.

An overall hybrid architecture, appropriate for face recognition tasks, is shown in Fig. 1. Face recognition 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 identifies the face using appropriate image representation and classification algorithms. The tools needed to detect face patterns and normalize them are discussed elsewhere[8][9], while this paper describes only the tools developed to realize and implement those stages of face recognition involved with classification tasks. The matching and surveillance tasks to be addressed by hybrid classifiers are (**i**) individual (ID) identification subject to displaying specific facial landmarks such as wearing glasses (see Sect. 6.2.2), and (**ii**) verification (see Sect. 6.1), respectively.

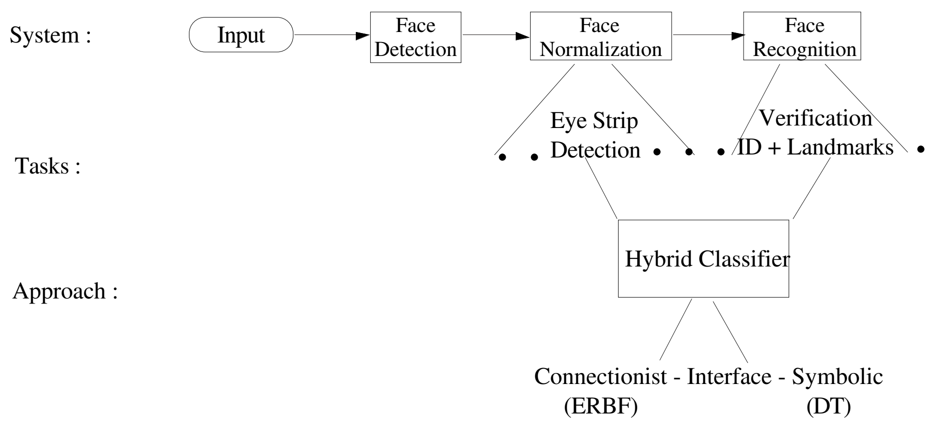

Figure 1. Automated Face Recognition (AFR)

4. Hybrid Classifiers

The hybrid classifiers consist of ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT). The reasons behind using RBF are its ability for clustering similar images before classifying them and the potential for developing in the future hierarchical classifiers where faces will be iteratively discriminated in terms of gender, race, and age, before final recognition takes place. Decision trees (DT) implement the symbolic stage using the RBF outputs because they provide for flexible and adaptive thresholds, and they can 'explain' the way classification and retrieval are eventually achieved. We now describe how to implement Ensembles of RBF (ERBF) and the hybrids consisting of ERBF and DT.

4.1 Ensemble of Radial Basis Function (ERBF) Networks

An RBF classifier has an architecture very similar to that of a traditional three-layer back-propagation network. 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 **G**aussian kernels. The basis functions (BF) used are Gaussians , where the activation level y_i of the hidden unit i is given by:

$$y_i = \Phi_i\left(\|X - \mu_i\|\right) = \exp\left[-\sum_{k=1}^{D} \frac{\left(x_k - \mu_{ik}\right)^2}{2h\sigma_{ik}^{\;2}}\right]$$

where h is a proportionality constant for the variance, x_k is the kth component of the input vector $X=[x_1, x_2, ..., x_D]$, and μ_{ik} and σ_{ik}^2 are the kth components of the mean and variance vectors, respectively, of basis function node i.. Each hidden unit can be viewed as a localized receptive field (RF). The hidden layer is trained using **k**-means clustering.

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). Ensemble 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 .

4.2.1 ERBF1

The first model integrates three RBF components and it is shown in Fig. 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 RBF components have been trained on the original image, the original image with some Gaussian noise added, and the original image with some degree of geometrical ('rotation') transformation performed on the image, respectively. Each one of the nine RBF nodes is trained until 100 % recognition accuracy is achieved by keeping the number of clusters fixed and varying the proportionality constant h. The corresponding nine output vectors thus generated are passed to a *judge* which makes a decision as to whether to accept or reject the input facial image. The actual decision as to accept or to reject the probe is given as: *if the norm of the average of all the nine outputs is greater than threshold* θ *then accept; else reject* .

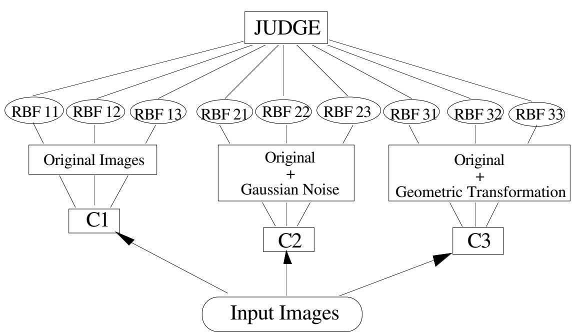

Fig. 2. ERBF1 Architecture

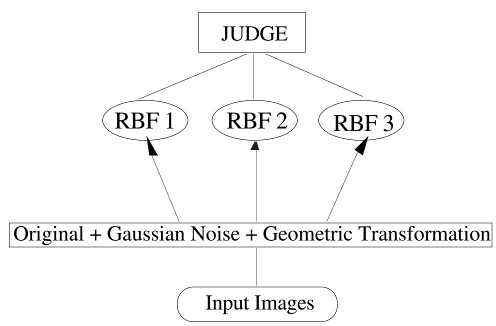

Fig. 3. ERBF2 Architecture

4.2.2 ERBF2

ERBF2 is derived from ERBF1 by increasing the number of images used to train each class and correspondingly decreasing the number of RBF nodes from nine to three (Fig. 3). This modification is achieved by passing the original image, the original image with Gaussian noise added and the original image with some geometrical (small rotation) transformation as training examples to the same node. The RBF nodes are further specified in terms of the number of clusters and overlap factors. As it was the case with ERBF1, training is performed until 100 % recognition accuracy is achieved for each RBF node**.** Nine output vectors, corresponding to the Cartesian product between the type of input and the type of RBF node, are generated as it was the case for ERBF1, and they are passed to a *judge* which makes a decision as to whether to accept or reject the input facial image. The actual decision as to accept or to reject is based on the same decision as for ERBF1: *if the norm of the average of all the nine outputs is greater than threshold* θ *then accept ; else reject* .

4.3 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[10]. 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* .

4.4 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 vectors (X1, ... ,X9) chosen for training are tagged as 'CORRECT' (positive example) or 'INCORRECT' (negative example) and the features are quantized to values ranging from 1 to 10. The input to the C4.5 DT 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 30 positive events and a random set of 100 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.

5. FERET

For the most part, the performance of face recognition systems reported in the literature has been measured on small databases, with each research site carrying out its experiments on their own database thus making meaningful comparisons and drawing conclusions impossible [11]. For the purpose of our application we use the FERET facial data base consisting of 1, 109 sets comprising 8, 525 images. Most of the sets consist of the following poses: two frontal shots ('fa' and 'fb'), 1/4 half (right and left) profiles ('qr' and 'ql'), 3/4 half (right and left) profiles ('hr' and 'hl'), and right and left (90 deg.) profiles ('pr' and 'pl'). Since large amounts of images were acquired during different photo sessions, the lighting conditions and the size of the facial images can vary**.** The diversity of the FERET data base is across gender, race, and age. The facial image sets were acquired without any restrictions imposed on expression and with two frontal images shot at different times during the photo session.

6. Experiments

The data base (DB) consists of 748 images taken over a span of two years. The database includes 374 ‘fa’ and 374 ‘fb’ of which 11 pairs of images are duplicates. All images are of size 256 x 384 using 256 gray scale levels. These images when passed to an Face Detection and Normalization system[8][9] produces 722 images - 350 'fa' and 350 'fb' images and 11 pairs of duplicate images. These cropped images are available at a resolution of 64 x 72.

6.1 Verification

First we report on experiments where, possibly for security reasons, the AFR system screens a large number of probes against some predefined gallery it has already been trained on. The experiments were carried out on images drawn randomly and the hybrid classifier consists of ERBF and C4.5, described in Sect. 4. The training strategy used is similar to that of $\boldsymbol{k}$ - fold cross validation (CV)[12]. In $\boldsymbol{k}$ - fold cross validation, the cases are randomly divided into k mutually exclusive test partitions of approximately equal size. The cases not found in each test partition are independently used for training, and the resulting classifier is tested on the corresponding test partition. The average error rates over all k partitions is the CV error rate. In our experiments, the first CV cycle on its first iteration takes as training set the first 50 'fa' frontal images, while testing on their corresponding 50 'fb' images and the remaining 311 'fa' and 'fb' images - on the second iteration the training set consists of the corresponding 50 'fb' frontal images, while testing on their corresponding 50 'fa' images and the remaining 311 'fb' and 'fa' images. In order for the 1344 ERBF output vectors for each cycle to be classified, quantization is performed and a random set of 30 output vectors are tagged as positive examples and another random set of 100 output vectors are tagged as negative examples. The remaining 1214 output vectors are tagged as test vectors. Once C4.5 completes its training, the remaining 1214 vectors are classified according to the decision tree learned by the system. Table 1 gives the results over the entire seven cycles for the case when ERBF1 is coupled with C4.5, while Table 2 shows the results when ERBF2 is coupled with C4.5. As a comparison we experimented also using only the connectionist ERBF model. Table 3 and 4 show the best results when ERBF1 and ERBF2 are used and θ is set to 0.65.

Cross Validation Cycle	Accepted (Correct) %	False Negative %	Rejected (Correct) %	False Positive %
A	89	11	100.00	0.00
B	87	13	100.00	0.00
C	91	09	99.16	0.84
D	90	10	99.58	0.42
E	89	11	98.58	1.42
F	92	08	97.50	2.50
G	85	15	99.50	0.50
Avg.	89	11	99.18	0.82

Table 1. ERBF1 and C4.5 Results

Cross Validation Cycle	Accept (Correct) %	False Negative %	Rejected (Correct) %	False Positive %
A	93	07	100.00	0.00
B	96	04	100.00	0.00
C	95	05	99.50	0.50
D	93	07	99.16	0.84
E	97	03	100.00	0.00
F	89	11	98.83	1.17
G	94	06	99.33	0.67
Avg.	93.85	06.15	99.54	0.46

Table 2. ERBF2 and C4.5 Results

Cross Validation Cycle	Accepted (Correct) %	False Negative %	Rejected (Correct) %	False Positive %
A	83	17	100.00	0.00
B	85	15	100.00	0.00
C	81	19	98.00	2.00
D	88	12	98.50	1.50
E	78	22	99.40	0.60
F	82	18	99.60	0.40
G	80	20	100.00	0.00
Avg.	82.43	17.60	99.30	0.70

Table 3. ERBF1 Results

Cross Validation Cycle	Accepted (Correct) %	False Negative %	Rejected (Correct) %	False Positive %
A	84	16	100	0.00
B	87	13	100	0.00
C	85	15	94.37	5.63
D	90	10	96.25	3.75
E	84	16	99.70	0.30
F	89	11	99.20	0.80
G	82	18	98.76	1.24
Avg.	85.85	14.15	98.32	1.68

Table 4. ERBF2 Results

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 verification performance increased on the average by 6 % while the false negative rate decreased by a similar amount of around 6%. Another observation one can make is that the ERBF2 model is better than the ERBF1 model. The plausible explanation is that having more examples to train on leads to better performance. We also note[9] 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. But however the primary disadvantage being that, it is very important to choose the right threshold. This is augmented by using the C4.5 system which adaptively learns the threshold rather than setting it *apriori.*

6.2 Query based Retrieval

Three types of retrieval queries, their realization, and the corresponding performance are described in the following subsections. The queries attempt to retrieve individual 'ID' subject to displaying contextual landmarks and they correspond to matching tasks.

6.2.1 Query1: 'Find all subjects wearing glasses'

The search engine searches the entire normalized data base for the presence of subjects wearing glasses using C4.5, which is trained on both positive examples, corresponding to wearing glasses, and negative examples for those without glasses. The data set used to train, disjoint to the data base we will later test on, includes 20 (correct / positive: '**+**' exs.) eye-strip with glasses of dimension 32 x 48, and 100 (incorrect / negative: '**-**' exs.) eye-strip without glasses. The actual eye-strips are usually of dimension 16 x 48 - the reason for using positive examples of larger size is that eye strip detection is enhanced when contextual cues given by the cheeks are taken into consideration. The set of eight features are extracted from the eye-strips using the Laplacian operator.

Since the original data base of 722 images consisted only of 3 subjects x 2 frontal poses wearing glasses, additional images of the same subjects wearing glasses were created by adding Gaussian noise and/or some geometrical change. In order to assess the robustness of the search engine 14 such additional images were created such that the total of images where glasses are present is now 20 out of 736. Cross validation results detect 19 out of 20 images where glasses were present corresponding to a 95 % accuracy, for a false positive rate of 0%.

6.2.2 Query2 'Find individual ID probe w/o glasses'

This task actually consists of two subtasks, **a:** '*Given an individual ID probe wearing glasses find all the instances of the probe with out glasses* ' and **b:** '*Given an individual ID probe without glasses, find all the instances of the probe with glasses* ', respectively. This query is implemented in two stages. First, a match stage using RBF network, and then, as explained above in Sect. 6.2.1, the presence of glasses or their absence is determined using C4.5.

Two cycles of cross validation, each one involving a random set of 50 images with three images ('probes') containing glasses or without glasses. Additional images were generated in a fashion similar to that described for Query1 in order to assess how robust the search engine is. The tables below give the results for Query2 - **a** -, and - **b** -, respectively. The results displayed show an average accuracy of about 90%, while the false negative rate is about 10%.

Cross Validation Cycle	Probe	Accepted (Correct) %	False Negative %	Rejectd (Correct) %	False Positive %
Cycle1	Probe1	90	10	96	4
	Probe2	85	15	98	2
	Probe3	95	05	96	4
Cycle2	Probe1	80	20	94	6
	Probe2	95	05	98	2
	Probe3	90	10	94	6

Table 5. Query2 - **a** - Results

Cross Validation Cycle	Probe	Accepted (Correct) %	False Negative %	Rejected (Correct) %	False Positive %
Cycle1	Probe1	85	15	94	6
	Probe2	85	15	98	2
	Probe3	95	05	98	2
Cycle2	Probe1	95	05	96	4
	Probe2	85	15	98	2
	Probe3	90	10	96	4

Table 6. Query2 - **b** - Results

7. Conclusions

Significant and robust results are achieved only when serious attempts are made to develop a large and representative test bed for a given application, face recognition for our purpose here, and when extensive experiments are conducted using proper evaluation procedures. We have proposed an architecture which addresses those concerns by considering hybrid classifiers and we have shown its feasibility using FERET - a large data base of facial images. The hybrid classifiers, consisting of an ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT), combine the merits of 'holistic' template matching with those of 'discrete' features using both positive and negative learning. The results reported in this paper on automatic face recognition are encouraging when one considers that the size of our test bed is in excess of 350 subjects and the great variability of the data base. In addition we have also demonstrated the feasibility of our approach on queries aimed at the retrieval of frames ('images') using contextual cues.

8. References

[1] Samal A., and Iyengar P., "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey", *Pattern Recognition*, Vol. 25, pp. 65-77, 1992.

[2] Chellappa R., Wilson C.L., and Sirohey S., "Human and Machine Recognition of Faces: A Survey, *Proc. of IEEE*, 83, 705-740, 1995

[3] Hampshire, J. B., and Waibel, A., The Meta-Pi Network: Building Distributed Knowledge Representations for Robust Multisource Pattern Recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **14**, 1992.

[4] Seung, H. S. *et. al.* Query by Committee, In *Proceedings of the fifth Workshop on Computational Learning Theory,* **5**, 287-294, San Mateo, Ca, Morgan Kaufmann, 1992.

[5] Battiti, R., and Colla, A. M., Democracy in Neural Nets: Voting Schemes for Classification, *Neural Networks, 7, 691-707, 1994.*

[6] Krogh, A., and Vedelsby, J., Neural Network Ensembles, Cross Validation and Active Learning, *NIPS*, **7**, Morgan Kaufmann, 1995.

[7] DePersia A.T, and P.J. Phillips The FERET Program: Overview and Accomplishments, 1995.

[8] Takacs B., and H. Wechsler , Face Location Using a Dynamic Model of Retinal Feature Extraction, *Proceedings of International Workshop on Automatic Face- and Gesture Recognition (IWAFGR)*, Zurich, Switzerland, June 1995.

[9] Gutta, S., J. Huang, D. Singh, I. Shah, B. Takacs, and H. Wechsler, Benchmark Studies on face Recognition, *Proceedings of International Workshop on Automatic Face - and Gesture Recognition (IWAFGR)*, Zurich, Switzerland, 1995.

[10] Quinlan, J.R., The Effect of Noise on Concept Learning*, Machine Learning: an Artificial Intelligence Approach*, R.S. Michalski, J.G. Carbonell and T.M. Mitchell, eds, Morgan Kaufmann , **2** 149-166, 1986

[11] Robertson, G., and I. Craw, Testing Face Recognition Systems, *Image and Vision Computing*, **19**, 609-614, 1994.

[12] Weiss, S. M. and C. A. Kulikowski, *Computer Systems That Learn*, Morgan Kaufmann, 1991.

Acknowledgements

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