Face Recognition Using Ensembles of Networks

S. Gutta+, J. Huang+, B. Takacs*, and H. Wechsler+

+Department of Computer Science
George Mason University
Fairfax, VA 22030

*Computational Sciences andInformatics
George Mason University
Fairfax, VA 22030

Abstract

*We describe a novel approach for fully automated face recognition and show its feasibility on a large data base of facial images (FERET). Our approach, based on a hybrid architecture consisting of an ensemble of connectionist networks - radial basis functions (RBF) and inductive decision trees (DT), combines the merits of 'abstractive' features with those of 'holistic' template matching'. The benefits of our architecture include (**i**) robust detection of facial landmarks using decision trees, and (**ii**) robust face recognition using consensus methods over ensembles of RBF networks. Experiments carried out using **k** - fold cross validation on a large data base consisting of 748 images corresponding to 374 subjects, among them 11 duplicates, yield on the average 87 % correct match, and 99 % correct surveillance ('verification').*

1. Introduction

Face recognition is a difficult task 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]. 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).

For most forensic scenarios there are two possible recognition tasks to be considered:

- **MATCH:** An image of an unknown individual is collected and the identity has to be found by searching a large set of images.
- **SURVEILLANCE:** The task is that of verification - whether a given probe does belong to a relatively small gallery, possibly consisting of just one image.

There are two major approaches for automated identification of human faces. The first approach, the *abstractive* one, extracts (and measures) discrete local features 'indexes' for retrieving and identifying faces, while standard statistical pattern recognition techniques are then employed for matching faces using these measurements. Examples of the feature extraction step include the discrete Karhunen-Loeve (KL) and the measurement of anthromorphic characteristics [3]. The other approach, the *holistic* one, conceptually related to template matching, attempts to identify faces using global representations. Characteristic of this approach are connectionist methods such as backpropagation using holons [4], principal component analysis (PCA), and singular value decomposition (SVD) using eigenfaces [5].

The approach described in this paper is hybrid in its nature and it is based on a psychologically plausible distinction between two types of cognitive operations: automatic, reflexive or low level (e.g., perception) *vs* controlled, deliberative or high level (e.g., reasoning). In the context of face recognition, hybrid architectures, consisting of connectionist networks and symbolic methods, would thus combine the merits of 'discrete' methods using numerical and symbolic values, with those of 'holistic' template matching, respectively.

We describe in this paper automated solutions for both the MATCH and SURVEILLANCE tasks, and present experimental results for large facial (image database) test beds as those made available under the FERET research program [6].

2. Ensembles of Networks and Consensus Methods

An early example of using ensembles of neural networks is due to Hamshire and Waibel [7]. 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. Battiti and Colla [8] have proposed means to combine the outputs of different neural network classifiers to improve the rejection-accuracy (ROC) rates and to make the combined performance better than that obtained from the individual components. The suggested concept of democracy is analogous with the human way of reaching a pondered decision - query by consensus.

Connectionist architectures are successful when they can cope with the variability of the image 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) and to train them 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. Specifically, both original data and distortions caused by geometrical changes and blur are used to train so robustness to different distortions is achieved via generalization. The specific realization of such ERBF is discussed in Sect. 4.

3. Methodology

The overall architecture is shown in Fig. 1. Face recognition usually starts through the detection of a pattern as a face and boxing it, proceeds then by normalizing the face image to account for geometry and illumination changes using information about the box surrounding the face and/or the eye location, and finally it identifies the face using appropriate classification algorithms. The tools developed to realize the AFR architecture and implement the face recognition process mentioned above are discussed in detail in Sect. 4.

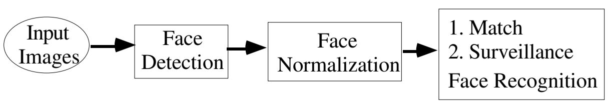

Figure 1. Architecture of the Automated Face Recognition (AFR) System

4. Automated Face Recognition (AFR)

We describe next each processing stage leading to face recognition and we address specific issues related to their computer implementation. The AFR addresses both the MATCH and SURVEILLANCE tasks as defined earlier. We experimented with a database of 748 images 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 levels.

4.1. Face Detection

To quickly locate the bounding box of the face (see Fig. 2) before more expensive facial feature (eyepair detection) algorithms are executed, we use a simple algorithm that operates on the edge image. First the edge image at resolution 64x96 pixels is created using the MATLAB implementation of the Marr edge detector. Horizontal and vertical projection profiles are computed to find the maximum projection ($P_{h,max}$ and $P_{v,max}$) values that would mark possible regions of interest. After thresholding/filtering the horizontal projection profile, by using only those lines where the projected values are greater than $0.8 \times P_{h,max}$, the upper boundary of the face box is found and it establishes a reference point for further processing. Once the upper boundary is detected, the algorithm searches for the maximal vertical projections within a certain distance from the center line to find the right and left sides of the bounding box. The lower boundary is finally computed using a fixed ratio of the upper boundary and the two edges on each side so the ratio height(h_o) / width(w_o) = 1.5. The accuracy of this stage, determined using visual inspection, is 94.5%. An incorrect box means that not all the facial features are present. The reason for failing to detect the right box stems primarily from low contrast. Note that all the images, including those for whom an incorrect box is found, are passed to the next stage, that of face normalization.

4.2 Face Normalization

Face normalization is carried out using the location of the eye strip. As a consequence two substages are implemented here: one would locate the eye strip and the other will use the midpoint of the eye strip for normalization. As it would be discussed later on learning involves both positive and negative instances ('examples').

4.2.1 Eye Strip Detection

Once the face box has been detected we proceed to locate the eye strip so an anchor for future normalization can be derived. The cropped ('boxed') face images are now made available at an increased resolution of 128 x 192, which is twice the resolution of the images used for cropping the face. The reason for affording increased resolution is that we have to seek finer facial details within a restricted area. The strategy for eye strip detection learns to label possible facial sites as eye strips using decision trees (DT) defined in terms of appropriate features. The eye strip detection process, consists of: data preparation, feature extraction, derivation of optimal decision tree (C4.5), and proper training and testing.

The data set used to train the eye strip detection comes from 40 images and would consist of 40 (correct / positive: '+' exs.) eye-strips of dimension 32x48 and 200 (incorrect / negative: '-' exs.) non eye-strip images. The actual eye-strips are usually of dimension 16x48 - the reason for using positive examples of larger size is eye strip detection would be enhanced when contextual cues given by the cheeks are taken into consideration. The eye strip is partitioned into a 2x3 arrays of windows where the eyes correspond to the left and right windows in the top row and immediately below we have the corresponding cheek windows.

The set of eleven features based on the Laplacian, ranking the Laplacian values, and local symmetry are used. Four features are computed and correspond to the mean of the Laplacian in each of the subwindows mentioned above except the middle ones. The leftmost and rightmost windows define now triangular areas whose base is the outer boundary of the strip and whose vertex is the corresponding mid-point across the boundary. The average values of the Lapacian in these triangles are ranked and four additional features in terms of equivalence classes are defined. As symmetry is important in describing facial landmarks and the face

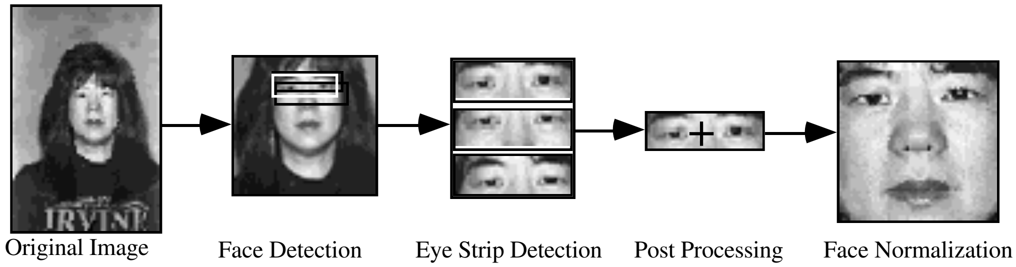

Figure 2. Example of Face Detection and Normalization

itself, the last three features measure symmetry characteristics. The features are defined across subtriangles derived from those defined above when split again into two right-angle triangles and measure the difference between the top three left and right subtriangles.

Inductive learning, as applied to the detection of eye-strips over the face region requires a special interface of numeric-to-symbolic data conversion. This has been achieved by tagging each of the 40 positive examples of eye-strips as 'CORRECT' and each of the 200 negative examples of non-eye-strips as 'INCORRECT'. The inductive learning system that has been used for this purpose is Quinlan's C4.5 [9]. The input to the C4.5 consists of a string of learning events, each event given as a vector of attribute values. C4.5 is an inductive machine learning system, which takes a set of positive examples and a set of negative examples and builds a classifier as a decision tree A 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 nonleaf a decision is evaluated, the outcome is determined, and the process moves on accordingly.

Once training has been completed and the learning system has found a suitable decision tree, specific testing and/or AFR system operation can proceed. This stage is performed by moving a window of size 32x48 with an overlap factor of size 6x8 pixels over the entire face box image. From each scanned window a set of eleven feature vectors are extracted . For each image, the windows that have been correctly classified are cut out of the face box image. As more than one eye strip is sometimes detected for each face box image, postprocessing such as winner-take-all (WTA) is needed, whereby the set of candidate eye strips is reduced to one strip. In cases where more than one eye strip still competes to win simple averaging of the coordinates yields the best location. An Example of eye strip detection for one face image is shown in Fig. 2. Note that there is a possibility for the decision tree classifier to fail in detecting eye strips; if this is the case, face recognition stops and the image is not processed any further. A total of 18 images were missed during the eye strip detection stage. The net result is that 722 images - 361 pairs - out of 748 images will be passed to the recognition stage, which corresponds to a success rate of 96.5%.

4.2.2 Face Normalization

Face normalization is performed to ensure that all the face images are of the same size for the classification stage. The eye strip detected in the previous step uniquely defines the anchor point for this normalization stage as the central point of the strip. The coordinates of the anchor point are mapped at resolution 128x192. Average measurements derived during face cropping and the midpoint of the eye strip just detected are used to 'cut' normalized faces for the recognition stage as shown in Fig. 2. The resulting face images are made available to the classification stage at a resolution of 64x72.

4.3. Face Recognition

The normalized face images are now ready for recognition. As the amount of information contained in each face image is quite large it seems appropriate that one would first try to cluster the data before its final classification. The Radial Basis Function (RBF) classifier [10] appears to be a natural choice for this task. 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 first discriminated in terms of gender, race, and age, before final recognition would take place. Such an approach is akin to that used by people when engaged in face recognition. We start this section by briefly reviewing some of the fundamentals behind RBF and then discuss how the MATCH and SURVEILLANCE tasks are actually performed.

4.3.1 RBF

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. That is, each hidden unit can be viewed as a localized receptive field. The hidden layer is trained using k-means clustering.

The most common basis function chosen is a Gaussian function. The outputs of the hidden unit lie between 0 and 1; the closer the input to the center of the Gaussian, the larger the response of the node. The activation level Z_j of an output unit is given by:

$$Z_j = \sum_i w_{ij} y_i + w_{0j}$$

where Z_j is the output of the j th output node, y_i is the activation of the ith BF node, w_{ij} is the weight connecting the ith BF node to the jth output node, and w_{0j} is the bias or the threshold of the jth output node. An unknown vector X is classified as belonging to the class associated with the output node j with the largest output Z_j.

4.3.2 MATCH

The RBF classifier described above has been applied to the MATCH task. Experiments were performed on 361 pairs of images, where each pair has both the frontals ‘fa’ and ‘fb’. Training is performed until either the 'fa' or 'fb' sets are perfectly classified, while testing takes place on the counterpart sets, 'fb' or 'fa', respectively. The first cycle takes a training set consisting of 350 'fa' frontal images, while testing on their corresponding 372 'fb' images - the additional 12 'fb' images correspond to duplicates. The second cycle would change training and testing roles between the 'fa' and 'fb' images. Table 1 below reports the results.

Cycle	Correct MATCH	Incorrect MATCH
1	88.89%	11.11%
2	84.23%	15.77%
Ave.	86.56%	13.44%

Table 1. Statistics for the MATCH Task.

Experiments were carried out when the REJECT / UNDECIDED option becomes available and the results are displayed in Table 2. The REJECT / UNDECIDED option corresponds to that case when the AFR system is undecided about what recognition should be made. Again two cycles are performed and the MATCH is accepted only if its strength is above the threshold θ. Otherwise no decision is made and the image is not classified ('matched').

average	Correct MATCH	Incorrect MATCH	REJECT Undecided
θ=.60	91.88%	8.12%	5.94%
θ=.65	93.58%	6.42%	7.16%
θ=.70	96.09%	3.91%	11.46%
θ=.75	98.38%	1.62%	13.47%

Table 2. Statistics for the MATCH Task with the REJECT (Undecided) Option.

4.3.3 SURVEILLANCE

We also experimented with the same fully automated face recognition system for the SURVEILLANCE task with an REJECT / UNDECIDED option. The experiments were carried out on images using two slightly different models. The major processing stages for the above system again include (i) face/head cropping, (ii) eye strip pair detection and normalization, and (iii) classification. The first two stages are similar in their implementation to those used for the MATCH task and they were described earlier.

The classification stage employs ERBFs - **E**nsemble of **R**adial **B**asis **F**unctions using the Radial Basis Function (RBF) neural network classifier as building blocks and the strategy used is similar to that of ***k*** - fold cross validation (CV) [11]. 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 training and testing roles between the 'fa' and 'fb' images would change. One can define in a similar fashion another seven CV cycles and the performance reported below is the average over seven cycles.

Training is done, as mentioned above, using ERBFs, and the REJECT / UNDECIDED option is implemented. ERBFs are trained over the original images and their slightly degraded (by Gaussian noise, blur, and/or geometrical transformations - small translation and rotation) corresponding frontal view images. The first model, ERBF1, consists of nine RBF networks, each slightly different from the other in terms of the number of clusters and overlap factors. The first three RBF networks are trained over the original images, the next three networks are trained over 'blurred' images, while the last three networks are trained over geometrically distorted images. The final decision as to whether to accept an image or to reject it is based on the following rule - *If the norm of the average of all (nine) outputs is greater than θ accept ('recognize') ; for (θ-$\Delta\theta$, θ) undecided ; else do not accept .* Average cross validation results corresponding to $\theta = 0.67$ are shown below, in Table 3, for different intervals $\Delta\theta$. Specifically, if the average of the ERBF output is within the interval $(\theta - \Delta\theta, \theta)$ no commitment is made as to the identity of the input and the image is labeled as REJECT / UNDECIDED.

av (CV)	Correct SURVEILLANCE	Incorrect SURVEILLANCE	REJECT (Undecided)
$\Delta\theta$ =.02	97.24%	2.76%	2.75%
$\Delta\theta$ =.06	97.96%	2.04%	5.14%
$\Delta\theta$ =.10	98.89%	1.11%	7.61%
$\Delta\theta$ =.14	99.89%	0.11%	10.22%

Table 3. Statistics for the SURVEILLANCE Task

For the second model, ERBF2, the number of RBF networks was reduced from nine to three by passing the original images and slightly degraded (by Gaussian noise, blur, and/or geometrical transformations - small translation and rotation) corresponding frontal view images to the same net. Each one of the three nets is different from the other in terms of clusters and overlap factors. During testing, nine output vectors 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. The final decision as to whether to accept an image or to reject it is based on the same rule as used for the first model. Average cross validation results are shown in Table 4, for different intervals $\Delta\theta$.

The rationale for the ERBF models comes from the need for robust performance. Specifically, one should expect 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. The second model, ERBF2, performs slightly better than ERBF1 possibly because of increased flexibility when more than one image per class is made available to each net. Future use of hybrid nets, as used for the eye strip detection, would make the need for empirical thresholds as those used by both ERBFs unnecessary, get rid of the REJECT / UNDECIDED option, and make the whole approach much more robust.

av (CV)	Correct SURVEILLANCE	Incorrect SURVEILLANCE	REJECT (Undecided)
$\Delta\theta$ =.02	98.68%	1.32%	0.95%
$\Delta\theta$ =.06	99.01%	0.99%	2.00%
$\Delta\theta$ =.10	99.43%	0.57%	3.51%
$\Delta\theta$ =.14	99.64%	0.36%	4.76%

Table 4. Statistics for the SURVEILLANCE Task

5. Conclusions

We described a novel approach for fully automated face recognition and showed its feasibility using a large data base of facial images (FERET). Our approach, based on a hybrid architecture consisting of an ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT), combines the merits of 'discrete and abstractive' features with those of 'holistic' template matching'. Training for face detection takes place over both positive and negative examples. The benefits of our architecture include (**i**) robust detection of facial landmarks using decision trees, and (**ii**) robust face recognition using consensus methods over ensembles of RBF networks. Experiments carried out using **k** - fold cross validation on a large data base consisting of 748 images corresponding to 374 subjects, among them 11 duplicates, yield on the average 87 % correct match, while 99 % correct surveillance ('verification') is achieved for a 2 % reject rate.

6. References

[1] Samal, A. and Iyengar, P. (1992), Automatic Recognition and Analysis of Human faces and Facial Expressions: A Survey, *Pattern Recognition*, Vol. 25, No. 5, 65-77.

[2] Chellappa, R., Wilson, C. L. and Sirohey, S. (1995), Human and Machine Recognition of Faces: A Survey, in *Proceedings of IEEE*, Vol. 83, No. 5, 705-740.

[3] Craw I.D, Tock B. and A. Bennet, "Finding Face Features", *European Conference on Computer Vision (ECCV)* , Genova, Italy, June 1992.

[4] Cottrell G.W. and Metcalfe J., "EMPATH: face, gender and emotion recognition using holons", in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 3, 564-571, R. P. Lippmann, R.P., Moody J. and Touretzky, D.S. (Eds.), Morgan Kaufmann, 1991.

[5] Turk M., and Pentland A., "Eigenfaces for Recognition", *J. of Cognitive Neuroscience*, Vol.3, No.1, pp. 71-86, 1991.

[6] 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, June 1995.

[7] 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, No. 7, 751-769,1992.

[8] Battiti, R. and Colla, A. M., "Democracy in Neural Nets: Voting Schemes for Classification", *Neural Networks* **7**, No. 4, 691-707,1994.

[9] 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 , pp. 149-166, 1986.

[10] Lippmann, R. P., and Ng, K.," A Comparitive Study of the Practical Characteristic of Neural Networks and Pattern Classifiers ", Technical Report 894, Lincoln Labs., MIT, 1991.

[11] 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.