Gender and Ethnic Classification Of Face Images

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

P. Jonathon Phillips
US Army Research Lab. Attn: AMSRL-SE-SE
2800 Powder Mill Rd.
Adelphi, MD 20783
jonathon@ragu.arl.mil

Abstract

*This paper considers hybrid classification architectures for gender and ethnic classification of human faces and shows their feasibility using a collection of 3006 face images corresponding to 1009 subjects from the FERET database. The hybrid approach consists of an ensemble of RBF networks and inductive decision trees (DT). Experimental Cross Validation (CV) results yield an average accuracy rate of - **(a)** 96% on the gender classification task and **(b)** 94% on the ethnic classification task. The benefits of our hybrid architecture include **(i)** robustness via query by consensus provided by the ensembles of RBF networks, and **(ii)** flexible and adaptive thresholds as opposed to ad hoc and hard thresholds provided by using only DT.*

1. Introduction

In order to interact socially, we must be able to process faces in a variety of ways. There is a vast amount of literature on social and cognitive psychology attesting to the impressive capabilities of humans at identifying familiar faces, as well as extracting information from both familiar and unfamiliar faces, including gender, race, and emotional state of the person [1]. Face processing 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 discuss these challenges in some detail [2,3].

Few attempts have been made to perform gender classification and the ones made used very small data sets, while we are not aware of any publication on ethnic classification. An early example of gender identification system (SEXNET) is due to Golomb et. al. [4]. In SEXNET, a back propagation network was trained to discriminate gender of human faces, yielding an accuracy of 91.9% on a data set of 90 exemplars corresponding to 45 male and 45 female subjects. The training set was composed of 80 exemplars and the remaining 10 exemplars were used for testing. Brunelli and Poggio [5] report an accuracy of 87.5% on a data set of 168 images corresponding to 21 male and 21 female subjects using 18 geometrical features. The system consisted of a hyper basis function network trained on the data sets of all minus one subject and tested on the excluded ones. A similar approach was used by Burton et. al. [6]. Using a discriminant function analysis they report an accuracy of 85.5% on a data set of 91 male and 88 female images respectively. Tamura et. al. [7] use a backpropagation network on a data set of 60 low resolution face images corresponding to 30 male and 30 female subjects, yielding an accuracy of 90%. The training set was composed of 30 faces (15 male and 15 female) and the testing set the remaining 30 faces. Recently Wiskott [8] reported an accuracy of 92% on a data set of 111 faces corresponding to 72 male and 39 female faces using Dynamic Link Matching (DLM) architecture.

As the size of the data sets used in the experiments reported above is quite restricted, no conclusions can be drawn about the ability of such methods to generalize and to scale up for large image databases, possibly consisting of several hundreds of thousands of images. This paper considers a hybrid classification architecture for gender and ethnic classification of human faces and it shows its feasibility using a collection of 3006 face images from the FERET database corresponding to 1906 images of gender male and 1100 images of gender female. The same database consists of 1932 images of Caucasian origin, 362 images of Asian origin, 474 images of Oriental origin and 238 images of African origin, respectively.

2. Hybrid Systems

Underlying the hybrid approach is the concept of reductionism, where complex problems are solved through stepwise decomposition. The hybrid approach 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

0-8186-8344-9/98 $10.00 © 1998 IEEE

(e.g., decision making or reasoning). Typically, in hybrid systems, reflexive tasks are assigned to the connectionist subsystem and deliberative tasks to the symbolic subsystem.

Intelligent hybrid (heterogeneous) systems [9] involve specific (hierarchical) levels of knowledge defined in terms of concept granularity and corresponding interfaces. Specifically, the hierarchy would include connectionist, fuzzy, and symbolic levels, with each level possibly consisting of ensemble architecture by itself. Note that such (homogeneous) ensembles render themselves as methods of choice for implementing data fusion techniques. As one moves upward in the hierarchical structure, we witness a corresponding degree of data compression allowing more powerful ('reasoning') methods to be employed on reduced amounts of data. We briefly review some examples characteristic of the conceptual ('granularity') levels referred to above and the means to connect the components of such hybrid architectures. An early example of homogeneous ensembles is due to Lincoln and Skrzypek [10]. Other examples of homogeneous ensembles include, democracy in neural nets [11], and committees of networks [12]. Ensembles of symbolic modules are usually referred to as multistrategy learning (AI) methods. As an example, Danyluk [13] presents Gemini, a system that integrates analytical and empirical learning. As an example of heterogeneous ensembles, Greenspan [14] has proposed an architecture for the integration of subsymbolic (connectionist) and symbolic ('rule-based') levels, using unsupervised (SOFM) and supervised learning (rule-based information theoretic approach), respectively.

The hybrid approach for gender and ethnic classification, described in this paper consists of connectionist and symbolic modules. The hybrid approach combines the merits of 'holistic' template matching' with those of 'discrete' methods using numerical and symbolic values, respectively. The connectionist stage is further defined in terms of ensembles of Radial Basis Function (RBF) networks, while the symbolic stage consists of Decision Trees (DT). The connectionist RBF networks are trained on different data sets corresponding to variations of the original data leading to increased ambiguity by employing different topologies for the networks themselves.

3. Hybrid Classifier Architecture

Face processing first detects a pattern as a face and then will box the face. It proceeds to normalize 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 results reported later on assume that the patterns corresponding to face images have been detected and normalized. The specific task considered herein is then that of gender and ethnic classification, i.e., discrimination of human faces as belonging to either a female or a male category and Caucasian, Asian, Oriental or to African categories respectively.

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, and they can interpret ('explain') the way classification and retrieval are eventually achieved. The hybrid architecture for the gender (ethnic) classification task is shown below in Fig 1.

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 similar to that of a traditional three-layer back-propagation network [15]. 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 basis functions (BF) used are Gaussians, where the activation level y_i of the hidden unit i is given by:

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

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.

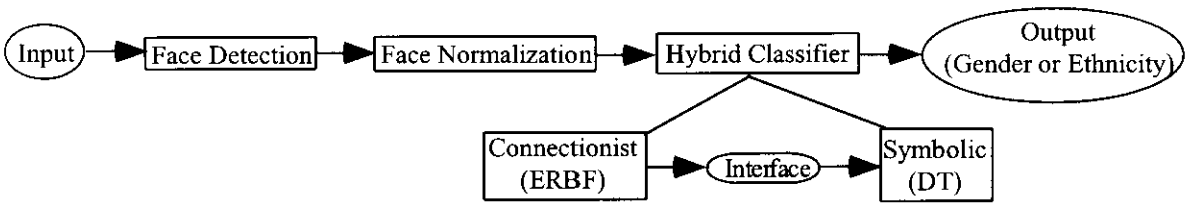

Figure 1. Hybrid Architecture for Gender and Ethnic Classification

The RBF input consists of n normalized face images pixels 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 hidden layer yields the equivalent of a functional facial base, where each cluster node encodes some common characteristics across the face space. 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' for the gender task and four nodes for the ethnic task. 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 [16].

3.1.1. 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 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 that particular class or not. The specific decision used is similar to that of majority voting namely, {if *majority* (5) of the 9 class outputs agree on a particular class then that probe belongs to that class}.

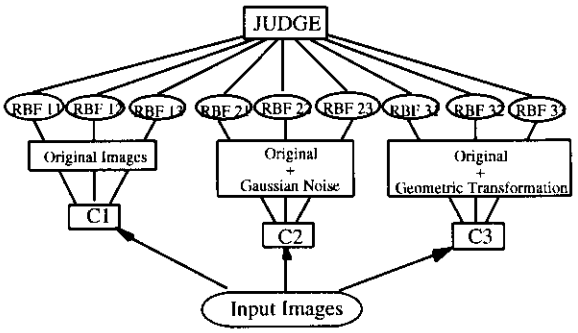

Figure 2. ERBF1 Architecture

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 (Fig. 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.

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 [17]. 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.*

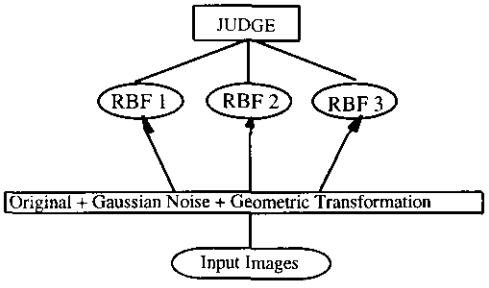

Figure 3. ERBF2 Architecture

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 is tagged as 'MALE' or 'FEMALE' for the gender task while, for the ethnic task one would tag the output class vectors according to their corresponding classes. 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. FERET Database

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 [18]. For the purpose of our application we use the FERET facial database consisting of 1, 934 sets comprising 14, 075 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 database is across gender, race, and age.

5. Experiments

The database for our experiments comes from the standard FERET facial database 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 an original pair ('fa' and 'fb'). An image of a subject taken at a different date is called a duplicate. Specifically, an image of a subject appearing in the database taken at a different date is labeled as 'duplicate1', while the modified (scaled up version of original face images) image of a subject is labeled as 'duplicate2'. These images are then resized to standard resolution of 256 x 384. The database now includes 494 ('duplicate1' + 'duplicate2') subject duplicate pairs. The images were acquired within a span of 3 years. The faces are manually located and normalized to a standard resolution of 64x72. The total number of images of gender 'Male' is 1906 and of gender 'Female' is 1100 images. The same database consists of 1932 images of Caucasian origin, 362 images of Asian origin, 474 images of Oriental origin and 238 images of African origin, respectively. The assignment of labels or ground truths ('Male' and 'Female' for the gender task and 'Caucasian', 'Asian', 'Oriental' and 'African') for the images was achieved through consensus among various people in the laboratory.

In Section 5.1 we report on the experiments conducted for the gender classification task, while in Section 5.2 we report the results for the ethnic classification task. A sample set of face images is shown in Fig. 4.

5.1. Gender Classification

Initially the set of 3006 images was divided into two sets of 2946 images and 60 images. The set containing 60 images (30 male and 30 female) was kept aside for training the DT. First we report on experiments when only a single RBF network is used, followed by when the two models of ensembles, and finally when the hybrid classifier described in Section 3 is used. The training and testing strategy used is a modified form of k - fold cross validation (CV) [19]. As the number of images corresponding to gender - male and gender - female are not equal, we divided the images in the following way. The images corresponding to gender - male are randomly divided into 62 mutually exclusive partitions of approximately equal size (1876/30=62.53) and 35 partitions (1070/30=35.6) for the female images. Next we randomly pick up one partition from the two (male and female) sets and used them for training the connectionist RBF network and the remaining partitions left over were used for testing. This process was repeated 20 times. The average error rate reported is the error rate over these 20 cycles. Table 1 gives the average CV results over 20 cycles.

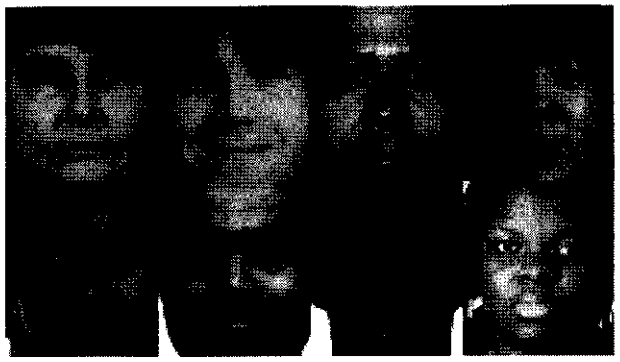

Figure 4. Sample Set of Face Images

The training and testing strategy used for ERBF1 and ERBF2 is similar to the one used for the RBF network. Table 1 also gives the average CV results over 20 cycles for the case when ERBF1 and ERBF2, respectively, are used.

For the case of the hybrid classifier consisting of ERBF and DT, again 20 cycles of CV were performed. In the case of hybrid classifier, training consists of two stages, namely that of training the ERBF's as well as training the DT. This has been achieved by first training the ERBF's as explained above and then using the class outputs (9) from the 60 images kept separate against the already trained ERBF's as an attribute vector for training the DT. Testing was then performed on the partitions left out during training the ERBF component. As an example, the total number of images for one cycle is 2886. The average CV results over all the 20 cycles are shown below in Table1.

Table 1. Results for Gender Classification Task

Gender Task	Correct Classification %	Mis-Classification %
RBF	70	30
ERBF1	79	21
ERBF2	82	18
ERBF1 with C4.5	90	10
ERBF2 with C4.5	96	4

5.2. Ethnic Classification

As in the gender task, the set of 3006 images were again divided into two sets of 2946 images and 60 images. The set containing 60 images (30 Caucasian, 10 Asian, 10 Oriental and 10 African) was kept aside for training the DT. Now we are left with a total of 1902, 352, 464 and 228 images corresponding to Caucasian, Asian, Oriental and African origins, respectively. As in the case of gender classification task, we report on experiments when only a single RBF network is used, followed by when the two models of ensembles, and finally when the hybrid classifier described in Section 3 is used. The training and testing strategy used is similar to the one used above. As the number of images corresponding to various ethnic categories are again not equal, we divided the images in the following way. The images corresponding to ethnic category caucasian are randomly divided into 63 mutually exclusive partitions of approximately equal size (1902/30=63.4), category asian into 11 partitions (352/30=11.73), category oriental into 15 partitions (464/30=15.46) and category african into 7 partitions (228/30=7.6). Next we randomly pick up one partition from each one of the ethnic sets and used them for training the connectionist RBF network and the remaining partitions left over were used for testing. This process was repeated 20 times. The average error rate reported is the error rate over these 20 cycles. This process remains the same for both ERBF1 and ERBF2 also. Table 2 gives the average CV results over 20 cycles.

The training and testing for the hybrid classifier remains the same as it was the case for the gender classification task. The output vectors here are tagged as belonging to one of the four ethnic origins. The average CV results over all the 20 cycles are shown in Table 2.

From the results reported in tables 1 and 2, 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% and 12% for the gender and ethnic tasks

respectively. 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') (see Section 3.1.2) 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.

Table 2. Results for Ethnic Classification Task

Gender Task	Correct Classification %	Mis-Classification %
RBF	62	38
ERBF1	74	26
ERBF2	82	18
ERBF1 with C4.5	86	14
ERBF2 with C4.5	94	6

6. Conclusions

We have proposed in this paper hybrid classifier architectures for gender and ethnic classification of human faces and showed their feasibility using a collection of 3006 face images corresponding to 1009 subjects from the FERET database. Cross Validation (CV) results yield an average accuracy rate of - (**a**) 96% on the gender classification task and (**b**) 94% on the ethnic classification task. The hybrid 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 based classifiers using both positive and negative learning examples.

The classifier architecture presented in this paper could serve in the building of hierarchical classifiers where faces would be sequentially discriminated in terms of gender, ethnicity and age before final recognition would take place similar to functional link architectures [8].

7. References

[1] D. Valentin, H. Abdi, A. Toole, and G.W. Cottrell, Connectionist Models of Face Processing: A Survey, *Pattern Recognition* 27(9): 1209-1230, 1994.

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

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

[4] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, SEXNET: A Neural Network Identifies Sex from Human Faces, in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 3, 572-577, Lippmann, R. P., Moody, J. E. and Touretzky, D.S., (Eds.), Morgan Kaufmann, 1990.

[5] R. Brunelli, and T. Poggio, Caricatural Effects in Automated Face Perception, *Biol. Cybern.* 69: 235-241, 1993.

[6] A. M. Burton, V. Bruce, and N. Dench, What's the Difference between Men and Women? Evidence from Facial Measurement, *Perception* 22: 153-176, 1993.

[7] S. Tamura, H. Kawai, and H. Mitsumoto, Male/Female Identification from 8 x 6 Very Low Resolution Face Images by Neural Network, *Pattern Recognition* 29(2): 331-335, 1996.

[8] L. Wiskott, Phantom Faces for Face Analysis, Pattern Recognition 30(6):837-846, 1997.

[9] L. R. Medsker, *Hybrid Intelligent Systems*, Kluwer Academic Publishers, 1995.

[10] W. P. Lincoln, and J. Skrzypek, Synergy of Clustering Multiple Back Propagation Networks, in *Advances in neural Information Processing Systems (NIPS)*, D. S. Touretzky (Ed.), Vol. 1, 650-657, Morgan Kaufmann, 1990.

[11] R. Battiti, and A. M. Colla, Democracy in Neural Nets: Voting Schemes for Classification, *Neural Networks* 7(4): 691-707, 1994.

[12] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[13] A. P. Danyluk, GEMINI: An Integration of Analytical and Empirical Learning, in *Machine Learning: A Multistrategy Approach* 4, R. S. Michalski, and G. Tecuci (Eds.), 189-215, Morgan Kaufmann, 1994.

[14] H. Greenspan, R. Goodman, and R. Chellappa, Texture Analysis via Unsupervised and Supervised Learning, in *Proc. of the International Conference on Neural Networks (ICNN)*, Vol. 1, 639-644, 1991.

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

[16] S. Gutta, and H. Wechsler, Face Recognition using Hybrid Classifiers. *Pattern Recognition*, Vol. 30, No. 4, pp. 539-553, 1997.

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

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

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

Acknowledgements

This work was partly supported by the DoD Counterdrug Technology Development Program, with the US Army Research Laboratory as Technical Agent, under contract DAAL01-97-K-0018.