Face and Hand Gesture Recognition Using Hybrid Classifiers

Srinivas Gutta, Jeffrey Huang, Ibrahim F. Imam, and Harry Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030
{sgutta, rhuang, wechsler}@cs.gmu.edu, iimam@aic.gmu.edu
http://chagall.gmu.edu/FORENSIC/

Abstract

This paper advances the methodology of hybrid classification architectures for face and hand gesture recognition tasks and shows their feasibility through experimental studies using the FERET data base and gesture images. The hybrid architecture, 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 'abstractive' matching using discrete features and subject to both positive and negative learning. The hybrid architecture, quite general as it applies to both face and hand gesture recognition, derives its robustness from (**i**) consensus using ensembles of RBF networks, and (**ii**) flexible matching using categorical classification via decision trees. The experimental results, proving the feasibility of our approach, yield (**i**) 93 % accuracy, using cross validation, for contents-based image retrieval (CBIR) subject to correct ID matching tasks, such as 'find Joe Smith with/without glasses', on a data base of 200 images, and (**ii**) 96 % accuracy, using cross validation, for forensic verification on a data base consisting of 904 images corresponding to 350 subjects (of whom 102 are duplicates). Cross validation results on the hand gesture recognition task yield a false negative rate of 3.6 % and a false positive rate of 1.8 % , using a data base of 750 images corresponding to 25 hand gestures.

Key words: content-based image retrieval (CBIR), decision trees (DT), ensembles of radial basis function (ERBF) networks, face recognition, FERET, hand gesture recognition, hybrid classifiers

1 Introduction

Faces are accessible 'windows' into the mechanisms that govern our emotional and social lives. The face is a unique feature of human beings. Even the faces of "identical twins" differ in some respects. Humans can detect and identify faces in a scene with little or no effort. This skill is quite robust, despite large changes in the visual stimulus due to viewing conditions, expression, aging, and distractions such as glasses or changes in hair style. There are several related subproblems: (**i**) detection of a pattern as a face, (**ii**) recognition, (**iii**) analysis of facial expressions, and (**iv**) classification based on physical features (Samal and Iyengar, 1992; Chellapa et. al., 1995). A system that performs these operations will find countless applications, e.g. criminal identification and retrieval of missing children, workstation and building security, credit card verification, and video-document retrieval. Automated recognition 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 - (i) MATCH: An image of an unknown individual is collected ('probe') and the identity is found searching a large set of images ('gallery'), and (ii) VERIFICATION: 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. Humans relate to each other also using their hands for communication. Computer recognition of hand gestures may provide a more natural human-computer interface. Another interesting application for hand gesture recognition is for developing 'smart rooms' [Pentland, 1996].

This paper advances the methodology of hybrid classification architectures for both face and hand gesture recognition tasks and shows their feasibility through experimental studies using the FERET data base and gesture images. 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 'abstractive' matching using discrete features and subject to both positive and negative learning. The hybrid architectures, quite general as they apply to both face and hand gesture recognition, derive their robustness from (**i**) consensus using ensembles of RBF networks, and (**ii**) flexible matching using categorical classification via decision trees.

2. Background

Underlying the hybrid approach is the concept of reductionism, where complex problems are solved through stepwise decomposition. Intelligent hybrid systems involve specific (hierarchical) levels of knowledge defined in terms of concept granularity and corresponding interfaces. Specifically, the hierarchy would include connectionist and symbolic levels, with each level possibly consisting of an ensemble architecture by itself, and with proper interfaces between levels. 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. An analogy to this strategy is focusing of attention as employed in visual perception. The advantages provided by each level consist of:

- Connectionism can handle the whole range of sensory inputs and their variability ('noise'). Its distributed nature provides for fault tolerance to missing and incomplete data. The output of such modules can be combined across ensemble of such networks. Last but not least, the output of such modules yields the sought after symbolic units needed for later stages of processing.
- Symbolic methods are compact and can fuse data from different sensory modalities and cognitive modes. As a consequence one can interpret the sensory input and explain it using meaningful coding units.

An early example of homogeneous ensembles is the Meta-Pi architecture suggested by Hamshire and Waibel (Hamshire and Waibel, 1992) for speech interpretation. Homogeneous ensembles of symbolic modules are usually referred to as multistrategy learning (AI) methods. As an example of heterogeneous ensembles, Greenspan (1991) has proposed an architecture for the integration of neural networks and rule-based methods using unsupervised and supervised learning for pattern recognition tasks.

3. Face and Hand Gesture Recognition

An overall hybrid architecture, appropriate for face and hand gesture 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. In the case of hand gesture recognition, same processes apply but normalization now ensures that all the hand gestures are off equal size. The tools needed to detect face (hand gesture) patterns and normalize them are discussed elsewhere [Gutta et. al, 1995], while this paper describes only the tools developed to realize and implement those stages of face (hand gesture) recognition involved in classification tasks. The matching and surveillance tasks to be addressed by hybrid classifiers for face recognition are (**i**) CBIR subject to correct ID match tasks, such as 'find Joe Smith with/without glasses', on a data base of 200 images, and (**ii**) surveillance (verification) on a data base consisting of 904 images corresponding to 350 subjects (of whom 102 are duplicates), and for hand gesture recognition they involve the recognition of specific gestures on a database of 750 images correponding to 25 hand gestures.

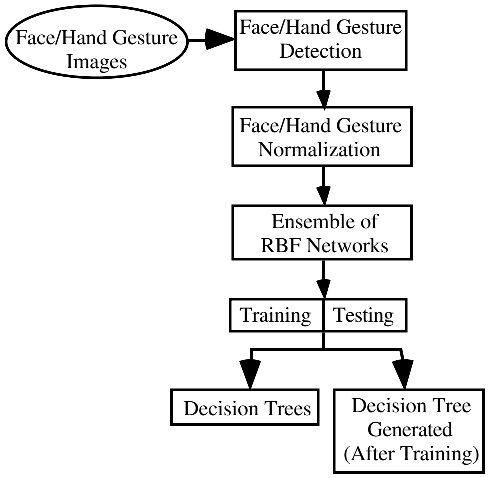

Figure 1. Automated Face / Hand Gesture Recognition (AFHGR) Architecture

4. Hybrid Classifier Architectures

The hybrid classifiers consist of an ensemble of connectionist networks - radial basis functions (RBF) - and inductive decision trees (DT). The reason behind using RBF is its ability for clustering similar images before classifying them. Decision trees (DT) implement the symbolic stage using the RBF outputs. 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 Figure 3. 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 gallery or not. The specific decision for face recognition is {**if** the norm of the average of all the nine outputs is greater than threshold θ **then** accept **else** reject}, and for hand gesture recognition is {**if** Max(R) is greater than threshold θ **then** accept **else** reject}, where

$$R_i = \sum_{i=1}^{9} C_{ij} \Big/ 9$$

i is the index for the number of networks, j for the number of classes and C_r is the output of the ith network for the jth class, where the threshold θ was set empirically.

4.2.2 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 4). Each RBF node is now trained on a mix of face images consisting of original ones and their distorted variations**.** The overlap factors, training remian the the same as used for ERBF 1. 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, and they are passed to a *judge*. The specific decision for face recognition remains the same as it was the case for ERBF1 while for hand gesture recognition it is {**if** Max(R) is greater than threshold θ **then** accept **else** reject}, where

$$R_i = \sum_{i=1}^{9} C_{ij} \Big/ 9,$$

i is the index for the number of output vectors, j for the number of classes and C_r is the output of the ith network for the jth class respectively and as in the case of ERBF1 the threshold θ was set empirically.

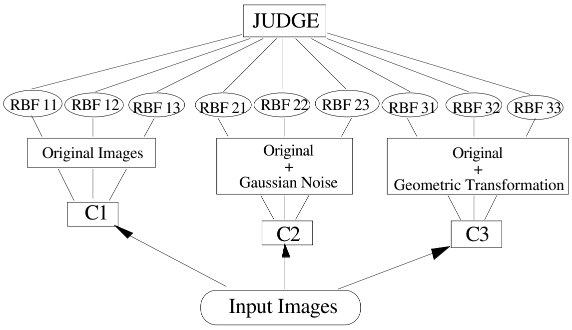

Fig. 3. ERBF1 Architecture

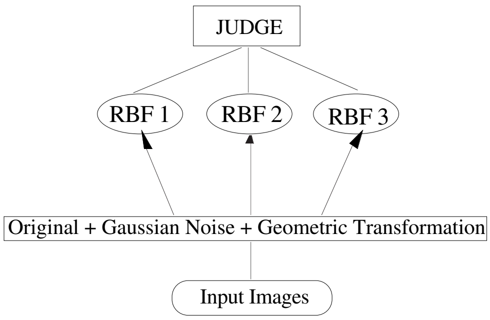

Fig. 4. ERBF2 Architecture

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 for face recognition is Quinlan's C4.5 [1986]. 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* .

The decision tree employed for hand gesture recognition is AQDT [Imam and Michalski, 1993]. AQDT learns a decision structure/tree from decision rules or examples by iteratively selecting an attribute to be a node in the structure, generates as many branches as the number of values of the selected attribute, associate all rules or examples with the appropriate branch, then if all rules or examples at any branch belong to one decision class, the system creates a leaf node for that decision class, otherwise, it repeat the same process.

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

Inductive learning, as applied to building a decision tree requires a special interface for numeric-to-symbolic data conversion. The ERBF output vector (X_1, ... ,X_9) chosen for training are tagged as 'CORRECT' (positive example) or 'INCORRECT' (negative example) and are quantized to values ranging from 1 to 10. The input to the C4.5 (AQDT) 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 positive events and a random set of negative events. The C4.5 (AQDT) 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. Image Acquisition

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 [Robertson and Craw, 1994]. To overcome such shortcomings, we have been developing over the last several years the FERET facial database so a standard tested for face recognition applications can become available [Gutta et. al., 1995] The FERET facial data base consists now of 1,109 sets (of whom 190 duplicate sets) comprising 8,525 images. 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 the two frontal images shot at different times during the photo session.

For the hand gesture recognition task images were acquired using an QuickTake-100 camera. The images were taken at a fixed resolution of 320x240 pixels encoded as 255 bits of gray scale levels. A total of 750 images corresponding to 25 hand gestures were taken from 15 subjects. From each subject 2 sets of 25 images were acquired. Each set was taken after a time lapse of 30 minutes to provide some variability in the orientation of the gesture. In addition a fixed distance of two and a half feet was maintained from the camera and the subject for each of the images acquired.

6. Experiments

As discussed earlier the experimental data reported herein is restricted to recognition only. We report first on the experiments carried out on the face recognition tasks, and then describe the results for hand gesture recognition . A sample set of normalized face and hand gesture images are shown in Fig. 2.

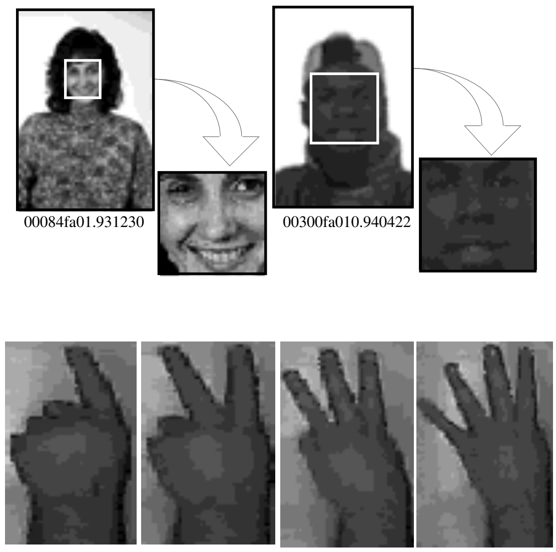

Figure 2. Examples of Normalized Face and Hand Gesture Images

6.1 Face Recognition

We started with an initial database of 952 frontal images drawn randomly from FERET. The database includes 476 ‘fa’ and 476 ‘fb’ of which 104 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 produce 904 images - 350 'fa' and 350 'fb' images and 102 pairs of duplicate images which corresponds to 95 % accuracy. These images are then made available at a resolution of 64 x 72. Note that the performance reported on face recognition tasks assumes that faces have been already located and normalized.

(i) CBIR Subject to Correct ID ('match') - 'Find Individual ID Probe with / without Glasses'

This task actually consists of two (a, b) subtasks, '*Given an individual ID probe wearing glasses find instance of the probe without glasses* ' and '*Given an individual ID probe without glasses, find instance of the probe with glasses* ', respectively. This query is implemented in two stages. First, a match stage seeks the identity of the ID probe using the original RBF network, and second, the presence of glasses or their absence is determined using C4.5. Note that the RBF network has been trained on variants of the original images in analogy to the ERBF2 network.

Two cycles of cross validation are performed, each one involving a gallery of 100 images, five images ('probes') wearing glasses or without glasses, and each probe is rendered 40 times to assess robustness. These 40 images were obtained by using duplicates as well by generating additional images by adding Gaussian noise and/or some geometrical change. The results give an average accuracy of about 93.3 %, while the false negative rate is about 6.7 % for subtask - a - <u>and</u> an average accuracy of about 93.5 %, while the false negative rate is about 6.5 % for subtask - b -, respectively.

(ii) Surveilling a Gallery of Images for the Presence of Specific Probes

First we report on experiments where, possibly for security reasons, the automatic face recognition (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 from batches 1, 2, 3, and 4, and the hybrid classifier consists of ERBF and C4.5, as described in Section 5. The training and testing strategy used is similar to that of k - fold cross validation (CV). As an example, the first CV cycle on its first iteration would implement the connectionist (ERBF) component using the first 50 'fa' frontal images, while learning the decision tree by randomly sampling positive and negative examples from the corresponding 50 'fb' images and the remaining 402 'fa' and 402 'fb' images. Note that the examples used to build the decision tree are generated using the already trained connectionist component. On the second iteration, the connectionist (ERBF) component is trained on the corresponding 50 'fb' frontal images, while learning the decision tree by randomly sampling positive and negative examples from the corresponding 50 'fb' images and the remaining 402 'fa' and 402 'fb' images. Note that images corresponding to subjects drawn from the gallery can be drawn from both the corresponding 'fa' or 'fb' images and also from the set of (102) duplicate (fa and fb) images available. The two iterations suggested above yield a sample space of 1708 ERBF output vectors (of size 9) from whom one would randomly sample the positive and negative examples required to train C4.5. Specifically, one now randomly selects 30 output (ERBF) quantized vectors tagged as positive examples and 100 output (ERBF) quantized vectors tagged as negative examples. The remaining 1578 output vectors are then tagged as test vectors. Table 1, below gives the average CV results for the case when the ERBF models were coupled with C4.5 (Hybrid 1 - ERBF1 and C4.5; Hybrid 2 - ERBF2 and C4.5). To assess the relevance of using hybrid classifiers we performed similar experiments where the classifier consists of only ERBF and RBF has been used by itself. Table 1 also gives the results for the case when ERBF1 and ERBF2, and RBF are used and the threshold θ is set to 0.65.

6.2 Hand Gesture Recognition

We started with an initial database of 750 hand gesture images corresponding to 25 distinct type of hand gestures taken from 25 subjects. All images are of size 320 x 240 using 256 gray scale levels. These images when passed to an Gesture Detection and Normalization system [Gutta et. al, 1995] produce 739 images which corresponds to 98.5% accuracy. These images are then made available at a resolution of 64x116. Note that the performance reported on hand gesture recognition tasks assumes that hand gestures have been already located and normalized.

Initially the preliminary experiments were carried out using C4.5 instead of AQDT with 15 different hand gestures. But as the number of gestures were increased from 15 to 25, the performance of C4.5 degraded and as a result AQDT was used instead. The results obtained were consistent when using either C4.5 or AQDT. Specifically, we obtain a false negative rate of 1.7 % and a false positive rate of 1 % when ERBF 2 and AQDT was used, while a false negative rate of 11.3 % and a false positive rate of 10 % was obtained when ERBF 1 and AQDT. The advantages of using AQDT over other decision tree learning programs comes from its adaptive capabilities including forcing user preferences on the learning process. The entire set of 750 normalized images have been split into two sets of 650 images corresponding to 13 subjects and 100 images corresponding to 2 subjects respectively. These 100 images are used later for generating (training) the decision tree.

The 25 different hand gestures are divided into 5 groups of 5 gestures each. A <u>*combination*</u> is defined as the union of 4 of the 5 groups to form 20 distinct classes (hand gestures). The left over 5^{th} group containing 5 gestures is used for testing. The reason behind this kind of split is to use the images corresponding to the 5th group for testing against the images present in the training set. Thus the total number of *combinations* would be 5. Each *combination* is randomly divided into 26 sets of equal size, where each set consists of 20 images corresponding to 20 distinct gestures. Training is performed by alternating among each one of the 26 sets. On its first iteration, the first set is used for training and the remaining 25 sets corresponding to 500 images and the 130 images corresponding to the five classes not present in that *combination* is used for testing. One could define in an similar fashion the remaining 25 iterations. Thus the number of internal cycles (iterations) for each *combination* would be 26. We present average CV results for the case when only an RBF network was used by itself, followed by using only ERBF(1,2). As in the case of face recognition the threshold (θ) for RBF, and ERBF(1, 2) was set at 0.65 respectively. The 100 images corresponding to 2 subjects kept separate are used for generating the decision tree. Training is performed by again splitting the 25 different hand gestures similar to that explained above. Table 2, below also gives the average CV results for the case when the ERBF models were coupled with AQDT (Hybrid 1 - ERBF1 and AQDT; Hybrid 2 - ERBF2 and AQDT).

CV Cycle	Accepted (Correct)%	False Negative %	Rejected (Correct) %	False Positive %
RBF	75.6	24.4	72.06	27.94
ERBF1	82.14	17.86	99.30	0.70
ERBF2	86.34	13.66	98.27	1.73
Hybrid 1	90.42	9.58	99.33	0.67
Hybrid 2	95.74	4.26	99.61	0.39

Table 1. Average CV Results for Face Recognition

CV Cycle	Accepted (Correct)%	False Negative %	Rejected (Correct) %	False Positive %
RBF	54.5	45.6	61.4	38.6
ERBF1	64.4	35.6	69.8	30.2
ERBF2	73.4	26.6	79.4	20.6
Hybrid 1	87.6	12.4	90.2	9.8
Hybrid 2	96.4	3.6	98.2	1.8

Table 2. Average CV Results for Hand Gesture Recognition

7. Conclusions

This paper described the methodology of hybrid classification architectures for face and hand gesture recognition tasks and shows their feasibility through experimental studies using the FERET data base and hand gesture images. The hybrid architecture, 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 'abstractive' matching using discrete features, subject to both positive and negative learning. The results obtained prove the feasibility and merits of hybrid classification architectures.

The Ensembles of RBF models (ERBF) outperform single RBF networks. The reason for this behavior comes from ERBF models implementing the equivalent of a 'query by consensus' paradigm. The experimental results support the idea that hybrid learning improves classification performance as the connectionist ERBF model coupled with an Inductive Decision Tree - C4.5 - yields better surveillance rates while decreasing the false negative rate. Another observation one can makes is that training with both original and distorted data, as it was the case with ERBF2, leads to improved performance (vs ERBF1).

8. References

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.

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.

Pentland, A. P. (1996), Smart Rooms, *Scientific American*, Vol. 274, No. 4, 68-76.

Hampshire, J. B. and Waibel, A. (1992), The Meta-Pi Network: Building Distributed Knowledge Representations for Robust Multisource Pattern Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 7, 751-769 .

Greenspan, H., Goodman, R. and Chellappa, R. (1991), Texture Analysis via Unsupervised and Supervised Learning, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Vol. 1, 639-644.

Gutta, S. and Wechsler, H. (1995), Face Recognition Using Hybrid Classifiers, *Pattern Recognition*, (to appear).

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

Imam, I. F. and Michalski, R. S. (1993), Learning Decision Trees from Decision Rules: A Method and Initial Results from a Comparative Study, *Journal of Intelligent Information Systems (JIIS)*, Vol. 2, No. 3, 279-304.

Robertson, G. and Craw, I. (1994), Testing Face Recognition Systems, *Image and Vision Computing*, Vol. 19, 609-614.

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