A HYBRID LEARNING APPROACH FOR BETTER RECOGNITION OF VISUAL OBJECTS

Ibrahim F. Imam*

SRA International
4300 Fair Lakes Court
Fairfax, VA 22033
iimam@verdi.iisd.sra.com

Srinivas Gutta

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

Abstract

Real world images often contain similar objects but with different rotations, noise, or other visual alterations. Vision systems should be able to recognize objects regardless of these visual alterations. This paper presents a novel approach for learning optimized structures of classifiers for recognizing visual objects regardless of certain types of visual alterations. The approach consists of two phases. The first phase is concerned with learning classifications of a set of standard and altered objects. The second phase is concerned with discovering an optimized structure of classifiers for recognizing objects from unseen images. This paper presents an application of this approach to a domain of 15 classes of hand gestures. The experimental results show significant improvement in the recognition rate rather than using a single classifier or multiple classifiers with thresholds.

1 Introduction

Recently, there has been a great interest in developing multimedia applications in wide-ranging fields. Communicative applications including audio and video systems require visual information about different objects to be able to recognize these objects, and communicate among each other (Maggioni, 1995; Freeman & Weissman, 1995; Kjeldsen & Kender, 1995). Such applications should be able to recognize objects within any environmental conditions. The reasons for such limitations include learning object classifiers using standard, noise free, and normalized objects; and using a non-adaptive strategy for recognizing new objects.

This paper introduces a new approach for learning optimized structures of classifiers for recognizing visual objects. In this paper, the term "*object*" denotes an image of one of the visual classes (e.g., hand poses). A *structure of classifiers* is a tree where each non-leaf node contains a classifier (e.g., a neural network), branches correspond to different outcomes or quantized intervals of the outcomes of each classifier, and leaves represent different classes of visual objects. The set of classifiers, used for building the decision structures, represents different visual alterations to the standard image of each object.

The main goal of this approach is to recognize a set of visual objects regardless of visual alterations of these objects in the corresponding images. This is done in two phases. The first phase is concerned with generating a set of classifiers (e.g., neural networks), one for each combination of visual alteration of the standard object and parameter settings of the classifiers. Only two alterations were considered, in this paper, by applying a set of geometrical transformations and noise to the original image of each object. The outputs from the alteration processes are called *altered objects*. In the second phase, another set of training images are tested by all classifiers. For each image, the set of values obtained from all classifiers along with the correct recognition are used for discovering an optimized structure of classifiers for recognizing objects in testing (unseen) images. To perform this task, we used a system, called AQDT-2 (Imam & Michalski, 1993), for learning task-oriented decision structures from examples. An *optimal decision structure* is a structure that contains the minimum number of nodes (classifiers), the minimum number of leaves, and correctly classifies the maximum number of testing examples (images of similar and different visual objects).

The methodology was applied on a hand gesture database created by the authors. The hand gesture database contains 15 different gestures. For testing, 9 cycles of two cross-fold testing method were applied to test the recognition rate. *The results obtained in this paper show a significant improvement in the recognition rate when using the proposed approach over using single classifier or ensemble of classifiers.*

2 Related Work

Carpenter *et al.* (1992) proposed a Fuzzy system, called ARTMAP, which achieves a synthesis of the Adaptive Resonance Theory (ART) between neural network and fuzzy logic by exploiting a close formal similarity between the computations of ART category choice and fuzzy membership functions. Greenspan, Goodman, and Chellappa (1994) proposed an architecture for the integration of neural networks and rule-based methods using unsupervised and supervised learning. This approach was used for pattern recognition tasks. Also, Towell and Shavlik (1994) presented a methodology for transferring

* Also a Research Affiliate of the MLI Laboratory at GMU.

symbolic knowledge into a neural network and for extracting rules from a trained neural network. The proposed approach defers from the above ones in the fact that it uses an adaptive methodology to combine multiple neural networks with decision tree approach.

An early example of using ensembles of neural networks, called Meta-Pi, was presented by Hamshire and Waibel (1992). The Meta-Pi classifier is a connectionist pattern classifier that consists of a number of source-dependent sub-networks. These sub-networks are integrated by a Time Delay Neural Network (TDNN) superstructure. The TDNN combines the outputs of the modules, trained independently, in order to provide a global classification.

Lincoln and Skrzypek (1990) proposed a clustering multiple back propagation networks to improve the performance and fault tolerance. Following training, a 'cluster' is created by computing the average of the outputs generated by the individual networks. The output of the 'cluster' is used as the desired output during training by feeding it back to the individual networks. The basic idea behind using such a strategy is based on the assumption that while it is possible to 'fool' a single back propagation network all the time one cannot mislead all of them at the same time.

Battiti and Colla (1994) proposed an approach to combine the outputs of different neural network classifiers to improve the rejection-accuracy rates and to make the combined performance better than that obtained from the individual components. Decisions are made based on the majority rule (concept of democracy). The concept of democracy is analogous with the human way of reaching a pondered decision query by consensus.

Several approaches to the problem of recognizing hand gestures have been proposed recently. One can distinguish between methods which assume a physically valid model of the hand (e.g., Quam, 1990; Sturman & Zeltzer, 1994) and those which do not extract or impose these type of 3-D constraints (e.g., Huang & Pavlovic, 1995; Lee & Kunii, 1993; Downton & Drouet, 1991).

3 The Approach

This paper presents a novel approach for learning optimized structures of classifiers for improving image recognition. The proposed approach consists of two phases. In the first phase, all images of visual objects are converted into appropriate format (e.g., digital) and segmented. Then for each image a set of altered objects are generated. An *altered object* is a transformation of the standard object such that both should be recognized as members of the same class (e.g., same gesture). For each alteration process, a classifier is used for learning classification for all object classes. The second phase is concerned with determining an optimized structure of classifiers for recognizing unseen or new images.

Figure 1 illustrates the proposed approach. All images are segmented and normalized. Then for each image, two identical copies were produced one by adding noise and the other by rotating it. This process is repeated three times using new classifiers with different parameter settings. All nine classifiers were trained using a portion of the training examples (set #1). Then, all classifiers are tested against the remaining portion of the training examples (set #2). The output values from testing the nine classifiers and the correct recognition (i.e., decision class) provided by the user are combined into a data vector, called a *record of recognition*. All records of recognition are combined together to form a set of examples used later to build the optimized structure of classifiers. Building the optimized structure of classifiers is done by the program AQDT-2 (Imam & Michalski, 1993). This program is selected over other decision tree programs because it can optimize the obtained structures using a variety of cost functions.

3.1 Phase I: Learning Classification of Visual Objects

The first step of capturing an object from an image is to locate the horizontal and vertical boundaries of the object. The method utilized here uses a simple algorithm that operates on the edge image using the Sobel's edge extraction method. All images are passed to the object normalization process. Each object (or an image) is normalized before starting the object recognition phase. This rest of this subsection illustrates the process of

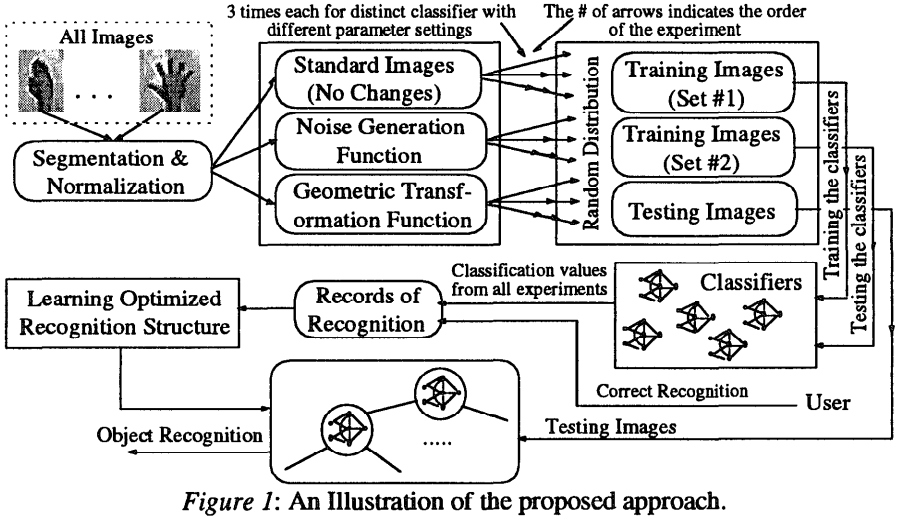

Figure 1: An Illustration of the proposed approach.

building, training and testing a single RBF classifier and an Ensemble of RBF classifiers (ERBF) and using both for object recognition. The ERBF classifiers are used later for acquiring the records of recognition.

Single Classifier: The construction of an RBF network is similar to the construction of any neural network. the number of input nodes in each RBF neural network was always set equal to the number of input images. The number of output nodes was set equal to the number of decision classes. The number of hidden nodes are optimally found, in each single RBF, by varying the number of clusters from fifth to equal number of the input nodes. At each stage of the variation, two additional parameters, the overlap factors and the proportionality constants were changed. These changes are repeated until all training examples (images) were classified correctly by the classifier. The same process is repeated for the cases when training on the original images with Gaussian noise and original images with geometrical transformation. The reason for using RBF network is because of its ability to cluster similar images before classifying them. A new object is assigned to the class with the highest value. To compare the performance of the RBF with the proposed approach, the training set #1 is used for training the classifier and the testing set is used for testing it.

Ensemble of Radial Basis Functions (ERBF): The proposed ERBF model integrates three *RBF components*, C1, C2 and C3, as shown in Figure 2. Each RBF component is further defined in terms of three RBF nodes, each of which specified in terms of the number of clusters and the overlap factors. The three RBF components have been trained on the original images, the original images after adding Gaussian noise, and the original images after geometrically rotating the objects with certain degree Ω, respectively. To recognize a new object, the image is tested by all the nine networks. Each network produces an output value for each class, called *classification value*. The sum of all 9 classification values for each class is called the *recognition rate*, R, for that class. An object X is a member of class C_i, if the recognition rate of this class is greater than the rate of all other classes, equation (1).

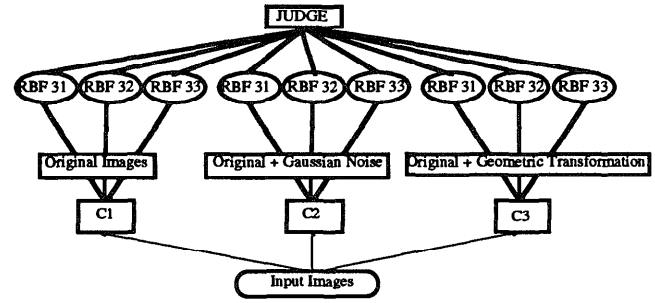

Figure 2: ERBF Architecture.

$$R(X) = \max_{C_j} \sum_{i=1}^{9} \left| O(N_i, C_j) \right| \quad (1)$$

where $O(N_i, C_j)$ is the value assigned by classifier N_i (i= 1, ..., 9) to class C_j (j= 1, ..., m; m is the number of decision classes) when testing the image X.

To generate the records of classification, The classifications provided by different classifiers and the correct recognition are grouped together to form one record. Figure 3 shows a description of the method. Note that the training set #2 is used in this phase as a testing set.

3.2 Phase II: Learning Optimized Structures of Classifiers

The second phase in the proposed approach is concerned with learning an optimized structure of classifiers for recognizing different objects from new images. *The proposed method is based on learning descriptions of situations to determine if the classification should be used for recognition or for selecting the best classifier for recognition*. Figure 4 illustrate the methodology used to determine an optimized structures for object recognition. To obtain such a structure, the AQDT-2 learning system (Imam & Michalski, 1993) is used to learn a task-oriented structure of classifiers that recognizes any new object using the minimum number of classifiers.

Input: Training (set #1.1) and testing (set #2.1) sets of segmented and normalized images.
Output: Classifications of a set of objects in unseen images.

Step 1: For all images (in set #1.1 and set #2.1), generate two identical sets of images (sets #1.2, #1.3, #2.2, and #2.3) .
Step 2: For all image in sets #1.2 and #2.2 add Gaussian noise. For all images in sets #1.3 and #2.3 add geometric transformation.
Note: All images in sets "#1x" are used for training the classifiers, and those in sets "#2x" are used for testing the classifiers.
Step 3: Create three identical copies of each set of images (e.g., #1.1.1, #1.1.2, #1.1.3 for set #1.1). Create a set of 9 RBF networks with different k-mean clustering one for each training set (start with #1).
For each classifier, steps 4 to 5 are repeated:
Step 4: Each classifier is trained using the corresponding set of images.
Step 5: When the training process of each classifier is finished, the corresponding set of testing images is used for testing the network (e.g., testing on set #2.a for training set #1.a. where a is any number or combination of numbers).
Step 6: The classification values obtained from testing each image using the 9 classifiers along with the correct recognition of the object were combined into a data vector.

Figure 3: Learning object classifiers.

Input: A set of training examples (records of recognition) and testing unseen images.
Output: An optimized decision structure that recognizes any new gesture (with or without alterations).

Step 1: Quantize all attribute values in the records of recognition.
Step 2: (Optional) Determine a set of disjoint rules describing the training examples.
Step 3: Specify the learning task for AQDT-2 (in this case, the decision structure should have minimum number of classifiers and minimum number of levels).
Step 4 to Step 6 are repeated until learning task is satisfied.
Step 4: Run AQDT-2 (initially with its default settings) to obtain a decision structure.
Step 5: Compare the information of the obtained decision structure with the optimal one. If the new structure is more optimal, store the values of the cost functions and the optimizing criteria. If the stopping criteria is satisfied, exit.
Step 6: Change the tolerance of the cost functions in AQDT-2 to give high preferences for some attributes. Go to step 4.

Figure 4: The method for learning optimized structure for shape recognition.

The advantages of using AQDT-2 over other decision tree learning programs is that it has adaptive capabilities including forcing partial attribute ranking, restructuring the decision structure according to a set of criteria and cost functions, etc. To select an attribute, AQDT-2 used the disjointness criterion which ranks attributes according to how their values discriminate the decision classes. For the experiments presented in this paper, the learning task was set to minimize the number of neural networks used in the obtained structure, and to maximize the recognition accuracy.

4 The Experiment

This section introduces an application of the proposed approach to a database of hand gestures. These images were taken by a KODAK Quick Take 100 Camera. The database contains 150 images of hand gestures. Images were taken from 5 different persons (2 sets of 15 images per each, Figure 5). The second set of images has been taken after a time gap of 30 minutes. The experiment was performed 3 times. For each time, a set of five different gestures (third, second, then first five gestures respectively) were considered as one decision class. This was done to increase the complexity of the problem.

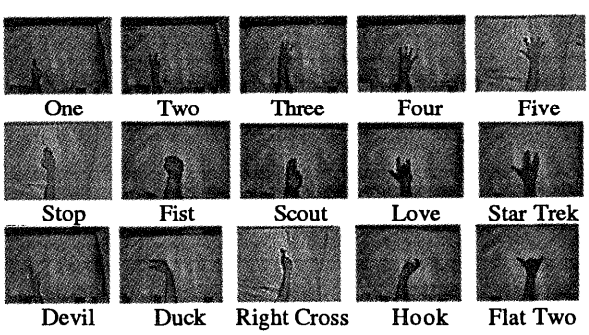

Figure 5: Images of hand gestures used in the experiment.

For each testing combination, a set of 9 testing cycles were performed. A testing cycle uses a two cross-fold method to evaluate each method. This is done by splitting all gestures, first, into two sets. The first set is labeled training set #1. The second set is divided into two subsets with ratio of 1:8. The smaller one is labeled training set #2, and the larger is labeled testing. The sizes of the training sets #1 in the 9 cycles are 7%, 13%, 20%, 27%, 33%, 40%, 47%, 53%, and 60%, respectively.
Original images were of size 240x320. All images were converted from RGB format to gray scale images of intensity varying from 0-255. All 150 hand gesture images were segmented and normalized, then all images were visually verified. Images were normalized to standard size of 64x116 pixels. Two visual alterations are used in this paper: 1) A gaussian noise with 0 mean and a variance of 10 was added to each image to generate new set of images; 2) A geometric transformation of 10^0 was applied to the original images to generate a different set of images.

The Recognition Rate of RBF and ERBF:
To compare the recognition rate of using single RBF classifier and ensemble of classifiers ERBF, the training set #1 is used for training the classifiers and the testing set was used for reporting their performances (Note: training set #2 was not used). Table 1. shows results obtained from testing a single RBF classifier. Each row in this table presents the results obtained from single combination. Table 2 shows results of testing the ERBF. The recognition process is done using equation (1).

Table 1: RBF Results.

Testing Comb.	Accepted (Correct)	False Negative	Rejected (correct)	False Positive
1	56.9%	43.1%	72.8%	27.2%
2	59.6%	40.4%	68.2%	31.8%
3	61.6%	38.4%	77.3%	22.7%

Table 2: ERBF Results.

Testing method	Accepted (Correct)	False Negative	Rejected (correct)	False Positive
1	71.6%	28.4%	81.1%	18.9%
2	69.8%	30.2%	82.2%	17.8%
3	72.4%	27.6%	77.8%	22.2%

Learning an Optimized Structure of Classifiers: To obtain the records of recognition from each experiment, the training set #1 was used for training the classifiers, while the training set #2 was used for testing. All outputs from each classifier were quantized according to a uniform distribution of the length of the interval (e.g., 10 intervals each of length 0.1). A discriminant descriptions of the decision classes of gestures were obtained by the AQ15 learning system (Michalski, et al, 1986). This intermediate process usually improves the overall performance. AQDT-2 used these rules to determine an optimized decision structure to classify any new gesture using the minimum number of networks. The program parameters were set to run 10 iterations with variable costs for all attributes, variable degree of generalizations, and minimizing the number of nodes and levels in the structure.

Each iteration uses random setting for the costs of the nine attributes. Each attribute represents one classifier. The program first learns a decision structure with its default settings. Then it changes the cost of one classifier at a time and re-learns a new structure. If the size of the structure is decreased or changed within a given tolerance, or the number of classifiers is decreased or changed within another tolerance, the system keeps the cost of that attribute and changes another attribute cost. The paper reports the results obtained from the default settings and the best 5 iterations.

Figure 6 shows a decision structure that is learned for one of the cycles of the testing combination #1. This decision structure contains 161 tests (e.g. the sum of number of tests from the root to any leaf node) divided over 56 paths (a path is a connection from the root to a leaf node). *Thus, the average (integer) number of classifiers needed to classify an unseen hand gesture is 3*. The total number of networks needed to classify any gesture belongs to the given group is 6 networks. About 72% of all possible gestures can be classified using only four networks (N4, N5, N7, and N8). For the different testing combinations described above, the AQDT-2 program was very successful in discovering optimized decision structures using subset of the neural networks used in Figure 6. In the second testing combination, only five networks were used to derive a decision (N3 was excluded). The average (integer) number of networks needed to derive a decision was also 3. In the third testing experiment, the same set of networks were used to obtain a decision.

Table 3 shows the error rate when testing the structure obtained by AQDT-2 with its default settings. It also shows the median and mean of error rates of the best five iterations (structures with minimum number of classifiers and have minimum number of average tests). Figure 7 shows a comparison of the error rate of using one RBF network (Baysian classifier), using a combination of nine different classifiers (ERBF) (majority voting), and using the proposed approach for image recognition. The results show a significant improvement in the recognition rate using the combination of the proposed approach.

Table 3: The error rate of using combination of ERBF and AQDT-2 for image recognition.

Testing Combination	Error Rate		
	Default Settings	Best 5 iterations	
		Median	Mean
1	4.3%	3.76%	3.6%
2	5.2%	3.12%	3.15%
3	3.25%	2.41%	2.4%
Average	4.25%	3.1%	3.05%

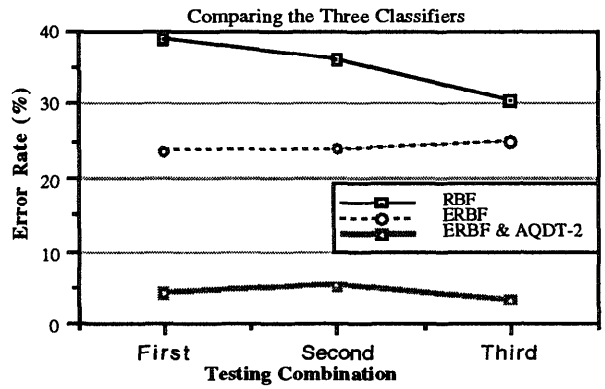

Figure 7: The error rate of using RBF, ERBF, and ERBF+AQDT-2 for gesture recognition.

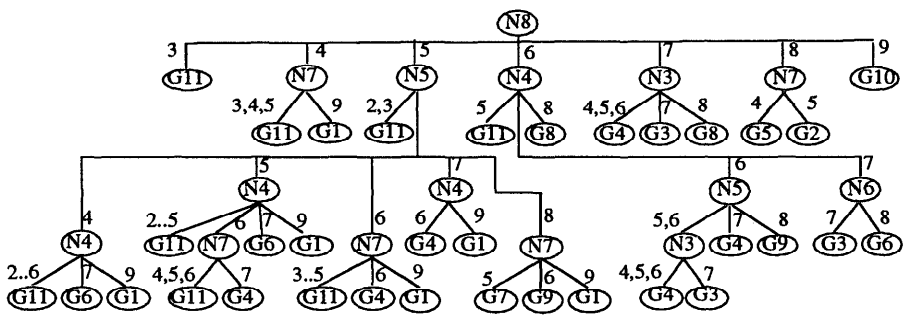

Figure 6: An optimized structure of classifiers for object recognition.

5 Conclusion

The paper introduces a new approach for improving the recognition rate of visual objects. The approach separates the process of learning classifications of a set of visual objects from the process of recognizing new objects. For learning object classifications, the proposed methodology learns classification of all objects from the original images and from different sets of altered images. In this research, only two alterations were

considered by either geometric rotating them or adding noise to the original image. For each original or altered set of images, three classifiers with different parameter settings were used to learn classification of different classes of objects. To recognize a new object, a structure of classifiers is obtained by the AQDT-2 learning system. This structure is used as a plan for recognizing objects in unseen images. The maximum classification value obtained by this classifier is used to either assign a class to the object in the image or select another classifier to test the image. The role of AQDT-2 is to determine the minimum set of classifiers needed for recognition and in which order the testing should take place.

The paper presented an application on a database of hand gestures. Three experimental combinations were performed on the data to analyze the performance of the methodology. In each combination, a subset of classes were grouped together as one class to increase the complexity of the problem. The results show a significant improvement in the recognition rate of new objects using the new approach against using a single RBF or ensemble of RBFs for recognition. The method allows more flexibility in recognizing visual objects. More analysis is needed to study the relationship between the number of alterations used and the complexity of obtained structures and the performance of object recognition.

Acknowledgments: The authors thank Zoran Duric, Mark Maloof, and Nirmal Warke for reviewing an earlier draft of this paper. This research was supported partially by the Forensic Lab. at GMU through the US Army Research Lab under Contract DAAL01-93-K-0099; and partially by the MLI Laboratory at GMU through the Advanced Research Projects Agency under grant No. N00014-91-J-1854 administered by the Office of Naval Research, in part by the Advanced Research Projects Agency under grants F49620-92-J-0549 and F49620-95-1-0462 administered by the Air Force Office of Scientific Research.

References

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

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., and Rosen, D.B., 1992. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps, *IEEE Trans. on Neural Networks*, Vol. 3, No. 5, pp. 698-713.

Downton, A.C., and Drouet, H., 1991. Image analysis for model-based sign language coding, *In Proceedings of the 6th International Conference on Image Analysis and Processing*, pp. 637-644.

Freeman, W.T., and Weissman, C.D., 1995. Television control by hand gestures, In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR)*, pp. 179-181, Zurich.

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*, Vol. I, pp. 639-644.

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

Huang, T.S., and Pavlovic, V.I., 1995. Hand Gesture Modelling, Analysis and Synthesis, In Proceedings of the *International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR)*, pp. 73-79, Zurich.

Imam, I.F. and Michalski, R.S., 1993. Learning Decision Trees from Decision Rules: A method and initial results from a comparative study. The International Journal of Intelligent Information Systems JIIS, Vol. 2, No. 3, pp. 279-304, Kluwer Academic Pub., MA.

Kjeldsen, R., and Kender, J., 1995. Visual Hand Gesture Recognition for Window System Control, *International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR)*, pp. 184-188, Zurich.

Lee, J., and Kunii, T.L., 1993. Constraint-based hand modeling and tracking, *Models and Techniques in Computer Animation*, pp. 110-127, Tokyo, Springer Verlag.

Lincoln, W.P., and Skrzypek, J., 1990. Synergy of Clustering Multiple Back Propagation Networks, *Advances in Neural Information Processing Systems (NIPS)*, Touretzky, D.S., (Ed.), Vol. 2, pp. 650-657, Morgan Kaufmann Publishers, San Francisco, CA.

Maggioni, C., 1995. GestureComputer—New Ways of Operating a Computer, *International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR)*, pp. 166-171, Zurich.

Michalski, R.S., Mozetic, I., Hong, J., and Lavrac, N., 1986. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, *Proceedings of AAAI-86*, pp. 1041-1045, Philadelphia, PA.

Quam, D.L., 1990. Gesture Recognition with a Dataglove, *Proceedings of the IEEE National Aerospace and Electronics Conference*, Vol. 2.

Sturman, D.J., and Zeltzer, D., 1994. A Survey of glove-based input, *IEEE Computer Graphics and Applications*, Vol. 14, pp. 30-39.

Towell, G.G., and Shavlik, J.W., 1994. Refining Symbolic Knowledge Using Neural Networks, *Machine Learning: A Multistrategy Approach*, Vol. IV, pp. 405-438, Michalski, R.S. and Tecuci, G. (Eds.), Morgan Kaufmann Publishers, San Francisco, CA.