# An Automatic Face Recognition System Using the Adaptive Clustering Network

Murali M. Menon and Eric R. Boudreau

■ We have developed an Automatic Face Recognition (AFR) System that uses the Adaptive Clustering Network (ACN)—a hybrid classifier that combines neural network learning with statistical decision making. The ACN automatically groups similar faces into the same cluster and creates new clusters for novel input faces. During training, the ACN updates the clusters continuously, and multiple clusters are created for the same subject to accommodate variations in the presentation of the subject's face (for example, changes in facial expression and/or head orientation). With incremental training, new subjects and further variations of existing subjects may be added on line without retraining the classifier on previous data. During the testing process, the ACN associates an input face with the cluster that most closely matches the face. The ACN minimizes misidentifications by reporting completely novel input faces as "unknown."

In addition to the ACN classifier, the overall AFR system includes a preprocessing stage that removes the background in an image and centers the face in the image frame, and a feature-extraction stage that compresses the data. The system requires relatively simple processing and has been implemented in software on a SUN workstation. The preliminary results have been encouraging. Using imagery of eight subjects taken with a video camera, the system achieved a correct-classification performance of 99% with no misidentifications.

FACE RECOGNITION is a task that humans perform quite easily under conditions in which there is little or no control over the environment and the orientation, distance, and state of the subject. In general, however, this seemingly simple task is a formidable machine-vision problem. The difficulty arises from the lack of a clear representation of what uniquely defines an individual face. The Automatic Face Recognition (AFR) system that we have developed uses the Adaptive Clustering Network (ACN), a classifier that automatically selects a set of features for distinguishing different faces.

The AFR system consists of three processing stages: a preprocessing module removes the background and centers the face, a feature-extraction module provides data reduction, and the ACN classifier identifies the subject. We used sample images containing subjects at three orientations with varying facial expressions to train the system. During the training process, the ACN automatically forms multiple clusters for each subject to account for variations in the input imagery, for example, differences in a subject's facial expression and/or head orientation. After all of the training imagery has been processed, the system associates each of the formed clusters with a particular subject. During testing, the system operates as a content-addressable memory in which an input is identified as the subject

associated with the cluster that most closely matches the input. The AFR system minimizes misidentifications by reporting completely novel inputs as "unknown." For very large databases the system can also be used to produce a list of the most likely candidates for a given input. If multiple frames of a given subject are available, then the system reports the name of the subject that occurred most frequently.

The architecture of the system offers several important benefits for AFR. The system can be trained incrementally, facilitating the addition of new subjects without the need for a complete retraining. Incremental training also allows the system to recognize a subject correctly over time despite variations in hair style, the presence/absence of glasses and facial hair, and the effects of aging. In addition, the system design takes advantage of the multiple frames that can be obtained from a real-time video observation of a subject. The system uses the multiple frames to accumulate evidence for a particular decision, thereby reducing the effect of single-frame errors.

Initially, we evaluated the system on a database of approximately 250 images containing eight different subjects at three head orientations. As shown in Figure 1, the eight subjects—four males and four females—were of various ages and races. The subjects were encouraged to converse and change their facial expressions during the videotaping, as shown in Figure 2, to increase the variability of the data. We trained the system on half of the frames and used the remaining half for testing. During training, the system formed a total of 25 clusters, with each subject represented by approximately three clusters. During testing, the system correctly identified 99% of the inputs on a single-frame basis. The remaining 1% of the inputs were labeled "unknown," resulting in 0% misidentification. On a multiple-frame basis, the system achieved a correct performance of 100%.

For a larger number of subjects and/or less constrained environments, we can maintain the high performance level of the AFR system by adding another

**FIGURE 1.** Samples of imagery of eight subjects.

independent measurement. For example, an audio-signal processing path that uses a subject's voice as input to a separate ACN could be combined with the video processing path to increase the probability of correctly identifying a subject. In fact, such a system could use the audio signal to make a decision even when a subject's face is obscured or outside the camera's field of view.

We have implemented the AFR system completely in software on a SUN SPARCstation 1 workstation. The training, which consisted of two passes of the training set through the system, required about 25 min (4 sec/image). The testing was quite rapid—less than 2 sec were required to identify an input. We designed the processing algorithms to exploit parallel/vector computer architectures, which can yield significant improvements in throughput.

Using both imaging and nonimaging data, we have evaluated the ACN classifier extensively on other object-recognition problems. The results of these evaluations have been the subject of several Lincoln Laboratory reports [1]. The ACN was initially utilized for automatic target recognition with laser radar range and passive-infrared imagery. The classifier has also been used to identify aircraft with pulsed-radar range returns (nonimaging). Another application was target detection from synthetic-aperture radar (SAR) imagery. Currently the ACN is the basis for an automatic ship classification system for inverse SAR (ISAR) imagery. The challenging problem of face recognition has provided an opportunity for further evaluation of the ACN on a complex pattern-recognition task.

(Note: The ISAR work was the subject of the article "An Automatic Ship Classification System for ISAR Imagery," which appeared in an earlier issue of this journal [2].)

Future enhancements to the ACN will lead to hierarchical architectures that incorporate multiple feature sets and perform multiresolution recognition. The long-term goal is to develop a classifier that is capable of addressing complex pattern-recognition problems such as unconstrained AFR in which the subjects are viewed from a wide range of aspects against complex and changing backgrounds under different lighting conditions.

## System Description

The Automatic Face Recognition (AFR) system is shown in Figure 3. The system contains a preprocessing stage that removes the background and centers a subject's face in an image frame, a feature-extraction stage that compresses the data, and a classification stage that identifies the face. In our work, we have used imagery acquired via a video digitizing board that converts an NTSC-format video signal from a video camera into a 720 × 480 pixel image

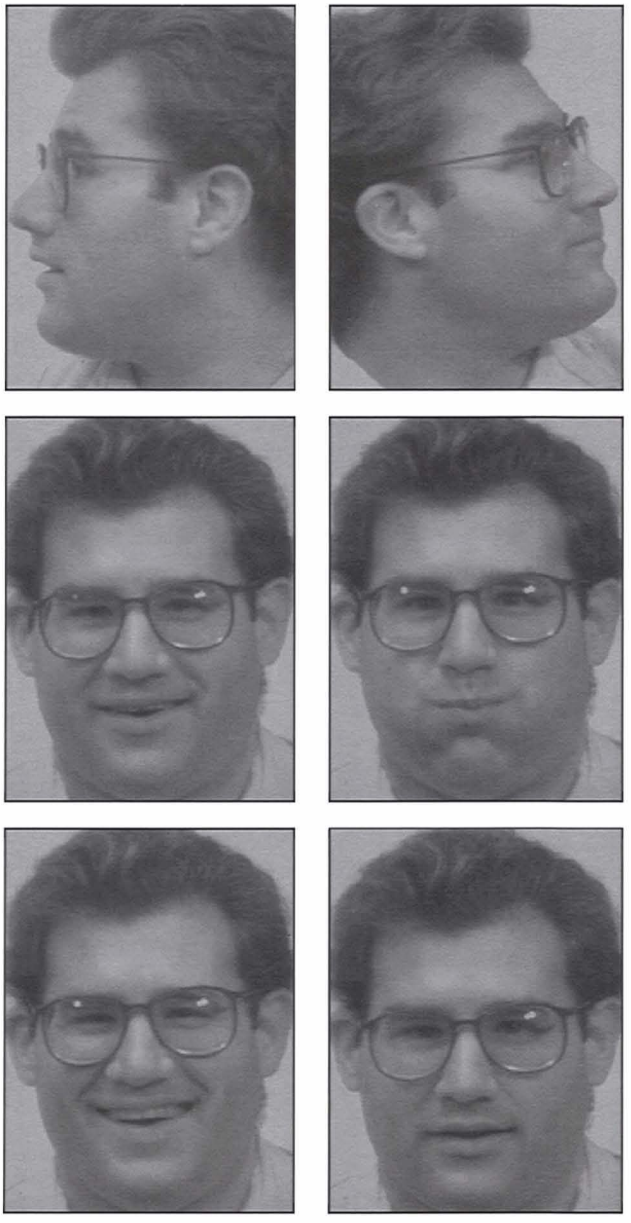

**FIGURE 2.** Samples of imagery of the same subject.

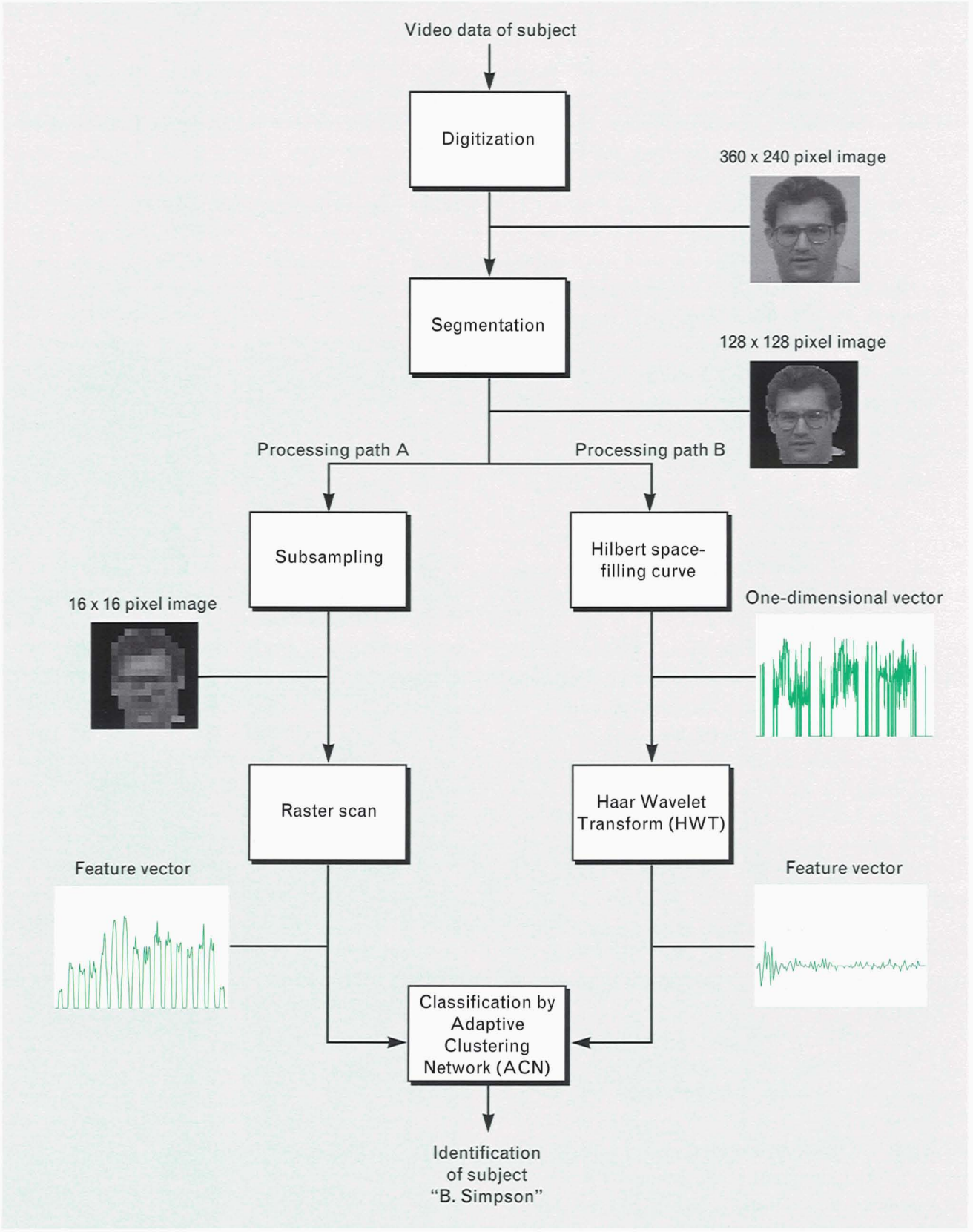


**FIGURE 3.** Automatic Face Recognition (AFR) system.

with a resolution of 7 bits/pixel.

*Preprocessing*

The AFR system subsamples the face imagery by a factor of 2 (resulting in a 360 × 240 pixel image) to reduce the processing load in subsequent stages. The boundaries of the face are then found with the Sobel edge detector [3]. The system applies a threshold to the edge-detected image to eliminate weak edges while retaining the outline of the head. The threshold used is 0.5 standard deviations above the mean intensity value of the edge-detected image. In the reduced image, pixels that lie outside of the head boundary are eliminated (i.e., set to zero). The system then finds the center of the face by calculating the pixel-intensity-weighted center of mass. A 128 × 128 region containing a frame with a centered face can then be extracted from the reduced image.

*Feature Extraction*

We have evaluated two approaches to feature extraction, both of which are shown in Figure 3. Feature extractors are used for data compression to restrict the classifier resource requirements. For instance, halving the length of the features reduces the requirements for ACN processing and storage by a factor of 2. Minimizing the processing and storage requirements becomes crucial when a large number of subjects must be processed.

The first approach to feature extraction (processing path A in Figure 3) uses the preprocessed image directly by subsampling the image to various pixel resolutions (64 × 64, 32 × 32, 16 × 16, 8 × 8, and 4 × 4) and then converting the subsampled image to a raster-scanned feature vector. This approach requires little processing overhead, but it produces feature vectors that are sensitive to the absolute image intensity (as opposed to the image intensity contrast).

The second approach (processing path B in Figure 3) uses a Hilbert space-filling curve [4], similar to the Peano curve [5], to sample the two-dimensional image into a one-dimensional vector. A one-dimensional Haar Wavelet Transform (HWT) [6] is then performed, and the lower-frequency Haar coefficients are used as features (see the box, entitled "Haar Wavelets for Multiresolution Signal Processing," on page 100). This approach avoids the computationally intensive two-dimensional Karhunen-Loeve Transform (KLT) [7] or Discrete Cosine Transform (DCT) [8] used in many AFR systems, and reduces the "blocking" artifacts associated with these two-dimensional methods. Another advantage of HWT is that the Haar coefficients are based on intensity contrast, rather than the absolute intensity. As a result, the feature vector is not sensitive to changes in overall lighting levels.

Figure 4 compares the two approaches to feature extraction: subsampling versus reconstruction from Haar wavelet coefficients.

*Classification*

After feature extraction, the salient information in the feature vector needs to be isolated. One approach looks for a fixed set of relationships that define a human face (e.g., the distance between a person's eyes) and classifies subjects based on these measurements. The difficulty with this approach is in the selection of a metric that will provide separability across a wide range of subjects. The ACN classifier, on the other hand, *automatically* determines the relevant information in the input by calculating the correlation between different feature vectors. The ACN is essentially a clustering algorithm that automatically groups similar feature vectors together to form clusters, and uses a *slow learning rule* (described below) for adapting the cluster centers as new data are encountered. By grouping multiple frames from a subject into a single cluster, the ACN avoids forming a lookup table of the training data.

The ACN shares many common attributes with other classifiers described in the literature. The ACN is unique, however, because it combines a slow learning rule with statistical methods to assign subject names to clusters for single- and multiple-frame classification. A classifier similar to the ACN is the Restricted Coulomb Energy (RCE) model [9], which also forms new clusters automatically, but the RCE does not adapt the cluster centers as new data are encountered. The *k*-means algorithm [10] updates the cluster centers by averaging all exemplars associated with a cluster, while the ACN uses slow learning to reinforce localized regions selectively. The standard *k*-NN algorithm [11] simply stores the entire training

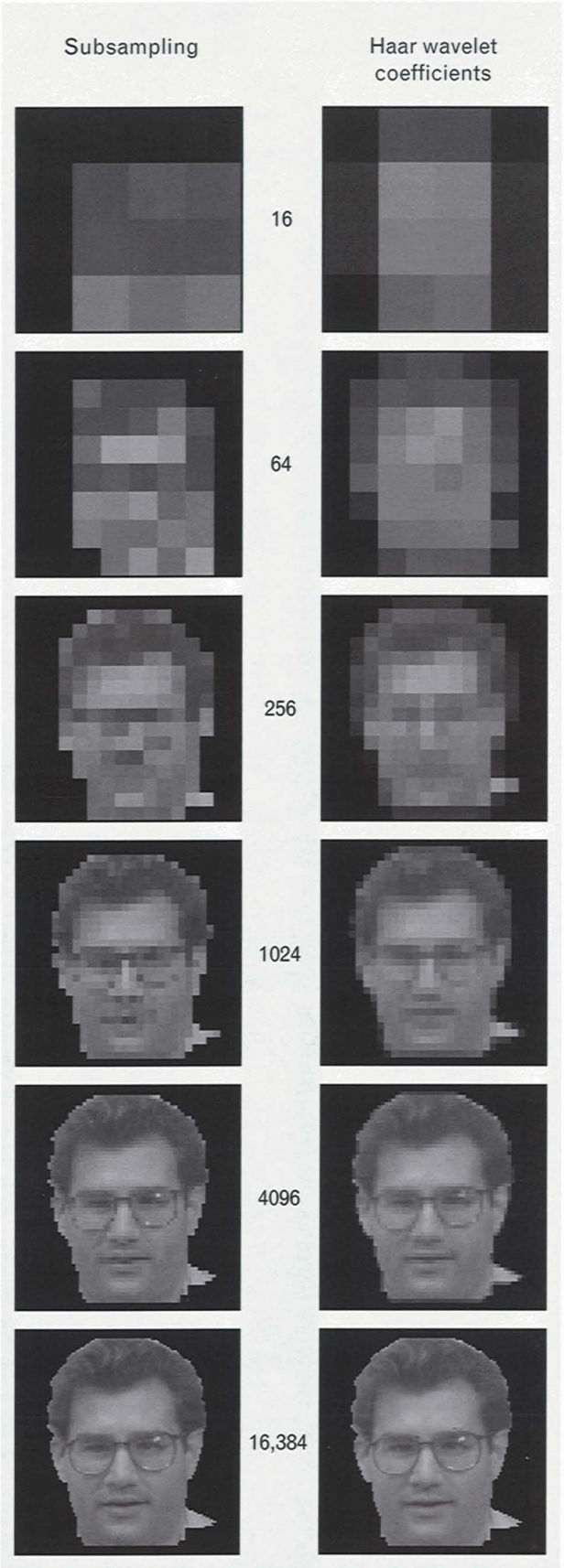


**FIGURE 4.** Comparison of two approaches to feature extraction: subsampling versus a reconstruction from Haar wavelet coefficients. The first column of images shows the subsampling approach (processing path A in Figure 3). From top to bottom, the images represent different pixel subsamplings: 4 x 4 (16 pixels), 8 x 8 (64 pixels), 16 x 16 (256 pixels), 32 x 32 (1024 pixels), 64 x 64 (4096 pixels), and 128 x 128 (16,384 pixels). Note the smearing of details at the lower pixel resolutions. In fact, at resolutions of 16 x 16 (256 pixels) and below, the structure of the face is unrecognizable. The second column of images shows a reconstruction from Haar wavelet coefficients (processing path B in Figure 3). From top to bottom, the images represent different numbers of coefficients used: 16, 64, 256, 1024, 4096, and 16,384. Note that, although details of the face have been eliminated at the lower resolutions, the underlying structure of the face has been maintained and is still recognizable down to 256 coefficients. Comparing the 256-pixel subsampled image with the 256-coefficient Haar image, we see that the face is more clearly discernible in the Haar image. This result was expected because the subsampling processing indiscriminately throws away information in the images. The selection of a low number of Haar coefficients, on the other hand, simply eliminates details of the face while maintaining the fundamental gross characteristics. Hence, for the lower resolutions, the additional information content of the Haar reconstruction justifies the extra processing that the wavelet feature extraction requires.

set. During testing, the most frequently occurring class (from the training set) among the $k$-nearest neighbors to the input is used to label the input. The Learning Vector Quantizer (LVQ) [12] clusters the training data but uses a computationally intensive *supervised* scheme for adjusting the cluster centers to reduce classification errors.

### *Training*

The ACN [13, 14], illustrated in Figure 5, groups inputs together into clusters based on the correlations between the inputs. For face recognition, the training correlation $c_{\text{train}}$ between an input feature vector $\mathbf{F}_{\text{input}}$ and an existing cluster $\mathbf{F}_k$ is defined as

$$c_{\text{train}}^{k} = \frac{\mathbf{F}_{\text{input}} \cdot \mathbf{F}_k}{\left\|\mathbf{F}_{\text{input}}\right\| \left\|\mathbf{F}_k\right\|}.$$

If the best match from the above equation is less than the training correlation threshold $\lambda_{\text{train}}$ (i.e., if

$\max_k c^k_{\text{train}} < \lambda_{\text{train}}$), then a new cluster is added. Otherwise, the best-match cluster is updated according to a simple "learning" rule given by

$$\mathbf{F}_k^{\text{new}} = \mathbf{F}_k^{\text{old}} + \gamma\left(\mathbf{F}_{\text{input}} - \mathbf{F}_k^{\text{old}}\right),$$

where $\gamma$ is the "learning" rate, which governs the rate of update of a cluster $\mathbf{F}_k$. For $\gamma = 1$, the input replaces the cluster; for $\gamma = 0$, no update occurs. Typically, $\gamma$ is set to a value between 0.10 to 0.25 to suppress small variations caused by noise and changes in a subject's facial expression, and to reinforce consistent features of the subject's face. After the training set has been processed, each cluster is assigned a class name that is associated with a particular subject.

### *Cluster Class Contrast*

The ACN is an unsupervised neural network in the sense that the feature vectors are grouped together into clusters regardless of what subject they represent. Therefore, it is possible that a given cluster may have been formed from several different input subjects, as illustrated in Figure 6. The class name (i.e., subject name) applied to a cluster is that class (i.e., subject) which has the highest frequency of occurrence in the cluster. Ideally, most of the entries in a cluster would be associated with the same subject. But if different subjects produce similar feature vectors, then no single class will form a majority in the cluster. During testing, confusable subjects (i.e., subjects with similar feature vectors) will match with such clusters, producing incorrect classifications. To reduce the number of misclassifications, the ACN reviews the clusters after training, and those clusters in which no single class dominates are labeled "unknown." During testing, the confusable subjects will be reported as unknown, thereby reducing the number of incorrect identifications.

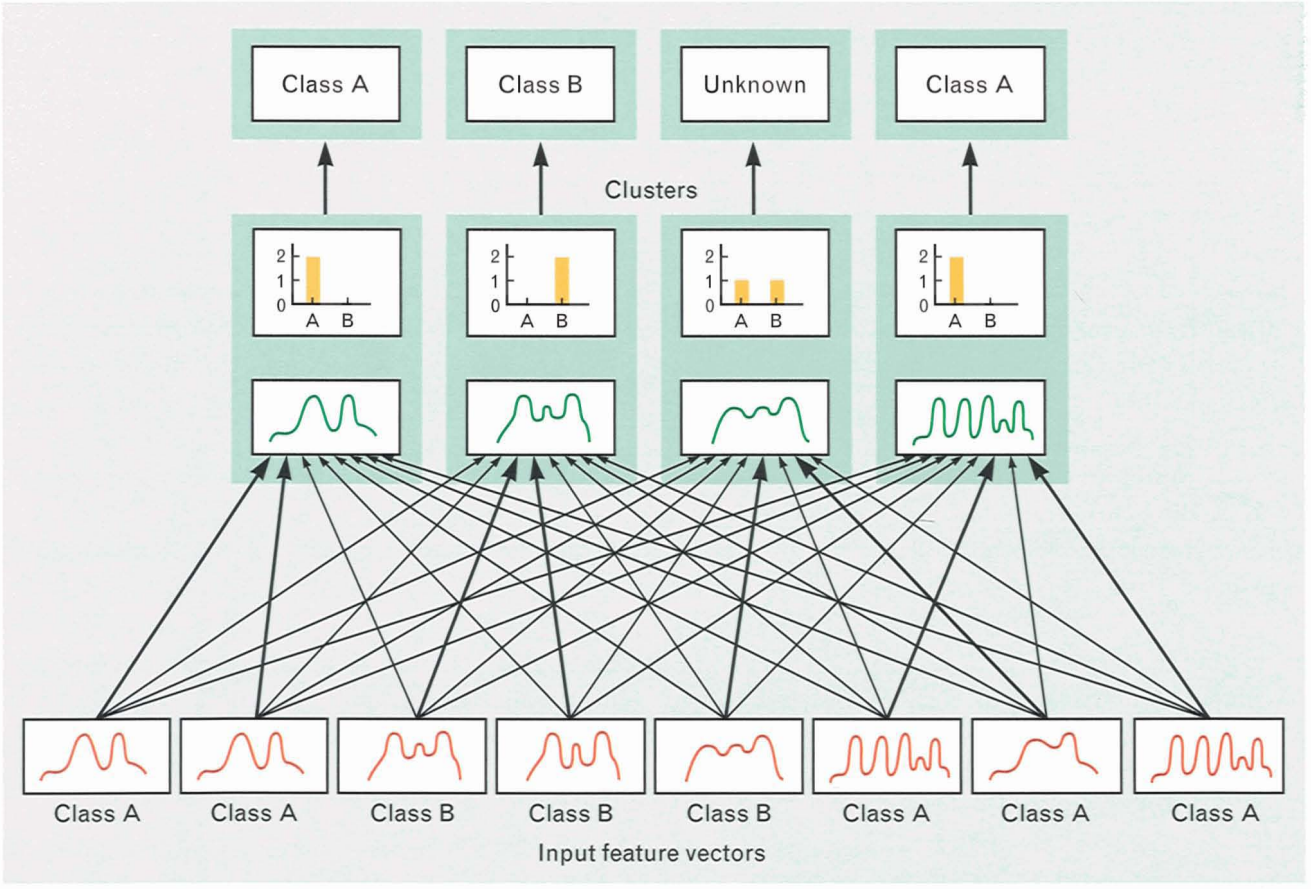


**FIGURE 5.** Adaptive Clustering Network (ACN). For details of the network, see the main text.

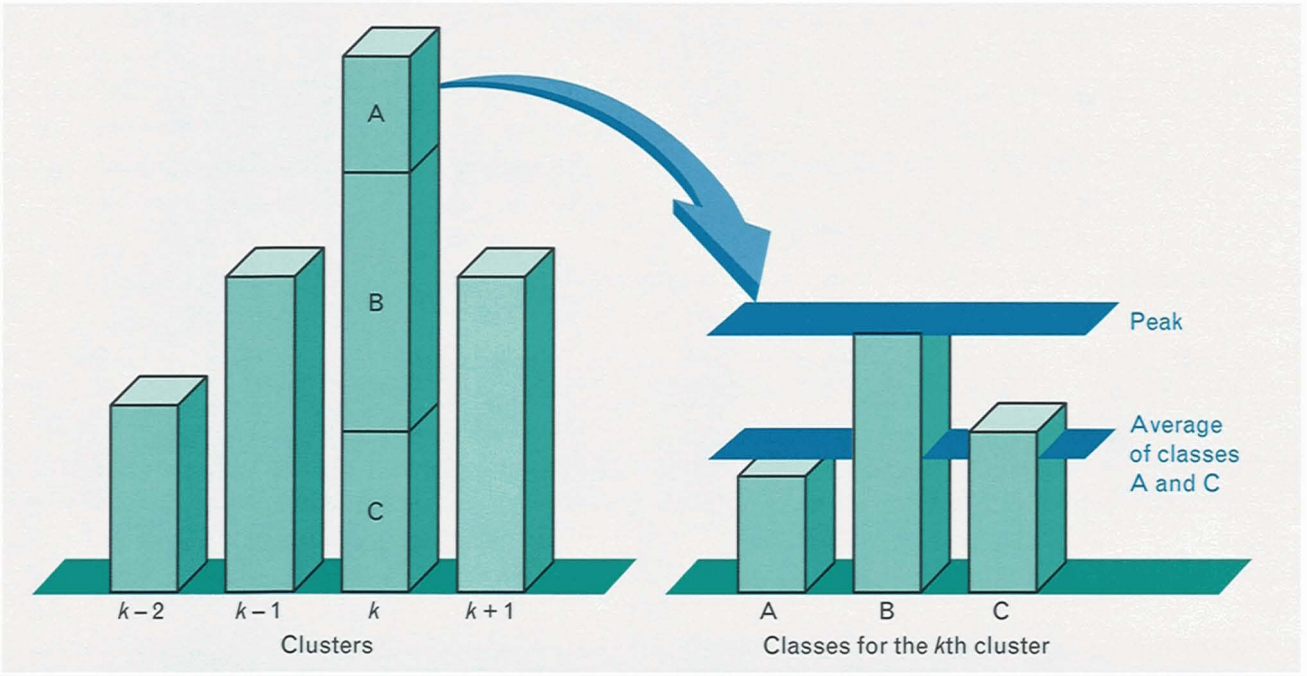


**FIGURE 6.** Cluster class contrast. The class with the highest frequency of occurrence in a cluster becomes the class name that is applied to that cluster. Thus, for the *k*th cluster, the class name B is assigned. To determine the degree to which the peak class dominates the other classes in a given cluster, we use a measure called the *cluster class contrast*, which is calculated for each cluster by taking the number of entries for the peak class and subtracting the average number of entries for the remaining classes. The result is then normalized by dividing by the total number of entries in the cluster.

A measure called the *cluster class contrast* is used to determine the degree to which the peak class dominates the other classes in a given cluster. The value is calculated for each cluster by taking the number of entries for the peak class and subtracting the average number of entries for the remaining classes. The result is normalized by dividing by the total number of entries in the cluster. Thus the class contrast for a cluster formed from a single class will be unity. The class contrast for a cluster in which no single class dominates will be close to zero, because the difference between the peak and the mean will be very small. After training, clusters whose class contrast measurements fail to meet a given threshold are labeled "unknown." A potentially adverse effect of this technique is the elimination of all of the clusters that represent a given subject by setting the threshold too high. This situation may occur for low $\lambda_{train}$ settings, because such settings increase the likelihood that multiple subjects will fall into the same cluster. The subjects whose clusters were eliminated will always be reported as unknown during testing.

### *Testing*

During testing, the ACN calculates the correlation between an input and each of the stored clusters. The ACN then sorts the clusters by their correlation to the input and forms a class histogram based on the top *k*-nearest neighbor (*k*-NN) clusters. The relative contribution of each cluster to the class histogram is determined by the cluster class contrast described earlier. A *k*-NN class contrast based on the peak in this histogram is calculated. If this class contrast exceeds a given threshold, then the input is identified as the class associated with the peak in the histogram. Otherwise, the input is reported as "unknown." The case of $k = 1$ labels the input with the class of the best-match cluster.

### *Multiple-Frame Processing*

To accumulate evidence for a particular decision and to reduce the effects of single-frame errors, we used multiple frames obtained from real-time video obser-

vations of the subjects. For each observation, a class histogram was constructed from the single-frame test results. The peak in the class histogram indicated the most probable subject identification, whereas the lack of a distinct peak caused the observation to be reported as "unknown."

The *observation class contrast* (analogous to the cluster class contrast) is a measure of the degree to which the peak class dominates the other classes for an observation. The value is calculated for each observation by taking the number of frames for the peak class and subtracting the average number of frames for the remaining classes. The result is normalized by dividing by the total number of frames in the observation. An observation is labeled "unknown" if its class contrast fails to exceed a given threshold.

## Results and Discussion

### *Training*

The training set contained 130 images of eight different faces with each face represented at a left facing (–90°), straight ahead (0°) and right facing (90°) orientation. We presented the entire training set to the ACN over two repetitions to allow the cluster structure to stabilize. A learning rate of 0.25 was used to suppress any minor variations in the imagery.

### *Correlation Threshold*

The training correlation threshold $\lambda_{train}$ determines the cluster structure. To study the effect of $\lambda_{train}$ on system performance, we trained the ACN by using various $\lambda_{train}$ values. For each different value, we presented the training set to the ACN in 10 separate trials in which the data were ordered in different sequences. The mean and standard deviation of the percent of the number of images that the system classified correctly, the percent that the system labeled "unknown," and the number of clusters formed were tabulated for each $\lambda_{train}$ setting. Figure 7 presents the percent correct, percent incorrect, and percent unknown as a function of $\lambda_{train}$.

If $\lambda_{train}$ is set too low, then all inputs will map to only one or a few clusters. If $\lambda_{train}$ is set too high, the inputs will be separated to the extent that each input will form its own cluster. Typically, $\lambda_{train}$ is chosen with Figure 8 such that a maximum performance level is obtained for a minimum number of clusters.

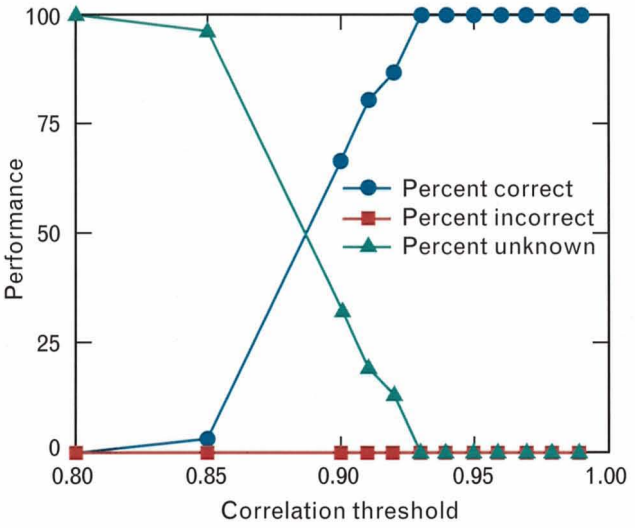


**FIGURE 7.** Training performance as a function of the training correlation threshold $\lambda_{train}$.

### *Training Results*

Using Figure 8, we set $\lambda_{train}$ equal to 0.93, which resulted in the formation of 25 clusters. The number of clusters formed corresponded approximately to the eight subjects, with three views per subject. Five of the 25 clusters contained only a single entry and were thus not considered reliable. The system labeled these single-entry clusters as "unknown," leaving 20 clus-

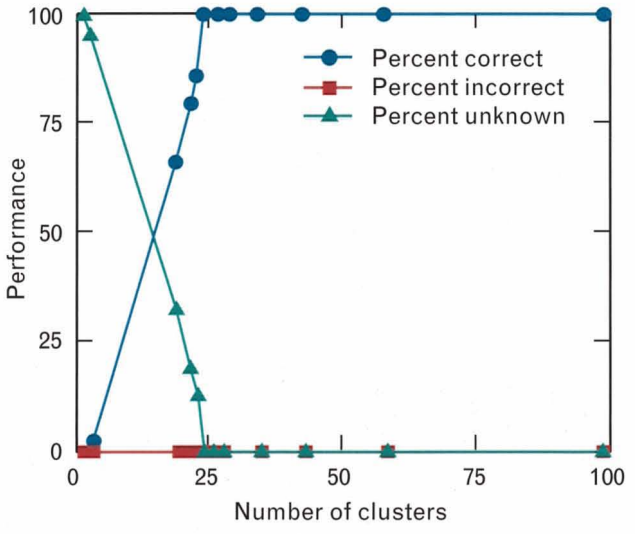


**FIGURE 8.** Training performance as a function of the number of clusters formed.

**Table 1. Class Confusion Matrix: Training**

| | *Class Reported by System* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Actual Class* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *Unknown* |
| *1* | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *2* | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *3* | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 |
| *4* | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 |
| *5* | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 0 |
| *6* | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 |
| *7* | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 | 0 |
| *8* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 |

ters with a valid class label.

The training performance of the system is summarized in a class confusion matrix in Table 1. The rows of the matrix represent the actual classes of the faces while the columns contain the classes that the system reported. Thus all of the off-diagonal entries would be zero for the ideal case in which the system classifies every face correctly, as is true in Table 1. Clearly the AFR system is able to separate the various faces at each of the different orientations.

To determine the effect of spatial resolution on the training performance of the system (processing path A of Figure 3), we subsampled the 128 × 128 input imagery to pixel resolutions of 64 × 64, 32 × 32, 16 × 16, 8 × 8 and 4 × 4. Figure 9 presents the results for a $\lambda_{train}$ setting of 0.93. Note that the performance is constant down to a resolution of 16 × 16 pixels and then degrades rapidly for 8 × 8 and 4 × 4 images. Still, even for 4 × 4 images the system labels inputs as "unknown" (53%) with few

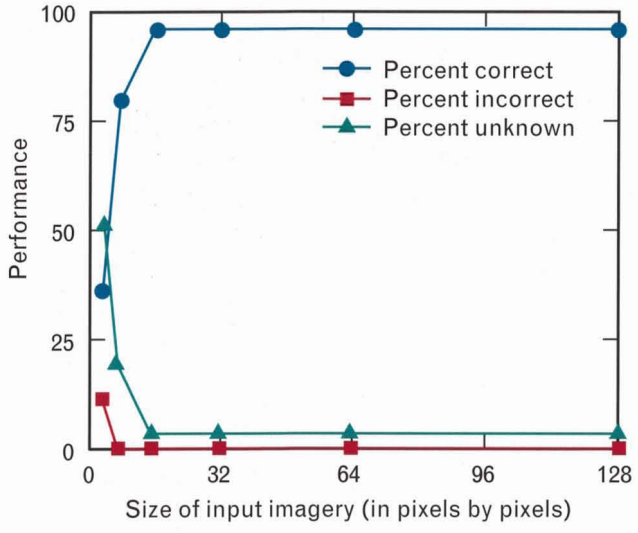


**FIGURE 9.** Training performance as a function of the input image size.

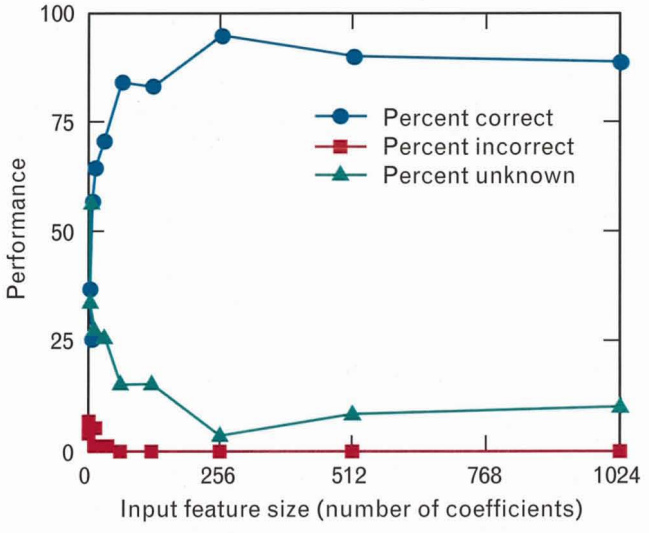


**FIGURE 10.** Training performance as a function of the number of Haar wavelet coefficients.

**Table 2. Class Confusion Matrix: Testing**

| | *Class Reported by System* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Actual Class* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *Unknown* |
| *1* | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *2* | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *3* | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 |
| *4* | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 |
| *5* | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 |
| *6* | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 | 0 |
| *7* | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 |
| *8* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 |

misidentifications (12%).

We then evaluated the system with extracted features (processing path B of Figure 3) rather than the raw imagery. Using a Hilbert space-filling curve followed by the calculation of the Haar wavelet coefficients for feature extraction, we sampled each image into a one-dimensional vector. Figure 10 presents the training results for various numbers of wavelet coefficients (see the box, entitled "Haar Wavelets for Multiresolution Signal Processing," on page 100). Note that the wavelet coefficients reduced the number of misidentifications at the lower resolutions. Using three wavelet coefficients (scale = 2), the system classified only 5% of the imagery incorrectly, compared with 12% for the raw subsampled 4 × 4 imagery.

### *Testing Results*

We tested the system with imagery that was exclusive of the training set. The testing set contained 123 im-

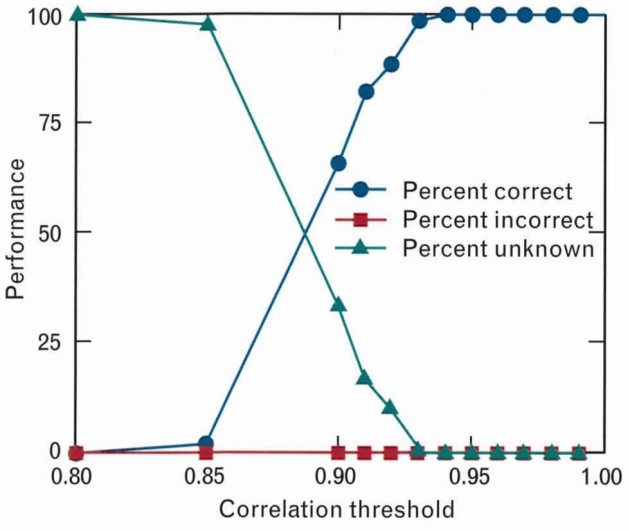


**FIGURE 11.** Testing performance as a function of the training correlation threshold $\lambda_{train}$.

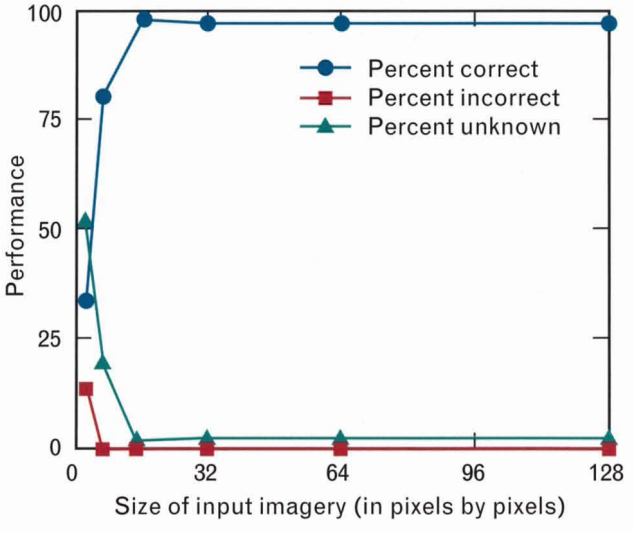


**FIGURE 12.** Testing performance as a function of the input image size.

# HAAR WAVELETS FOR MULTIRESOLUTION SIGNAL PROCESSING

THE BASIC IDEA behind multiresolution processing is that a signal should be modeled with a set of *basis functions* that operate over a range of resolutions to allow for changes in the signal that can occur at a multitude of different scales. For the basis functions, we can use *wavelets,* which are a general set of functions that incorporate both position and scale in their definition. A particular class of wavelets uses a prototype basis function from which all other functions can be derived by the appropriate shifting and scaling of the prototype. Hence a particular class of wavelets will have the same shape but differ in the location and resolution of their action on the signal. (Note: For a comprehensive review of wavelets, see Reference 1.)

We can approximate a signal by taking a linear combination of the wavelet basis functions $\phi$ multiplied by a set of wavelet coefficients $c$. For feature extraction, wavelets are a convenient way to represent the signal because the wavelet coefficients (which serve as the features) correspond to the correlation of the prototype function with the signal at various positions and scales. Hence a signal that varies quickly (high frequency) will result in large-magnitude coefficients at small scales and small-magnitude coefficients at large scales. For our face-recognition application, the feature set is the magnitude of the first $M$ coefficients. The value of $M$ can be varied to find the minimum number that would correctly separate the subjects.

In our present work, we have used Haar wavelets [2]. Figure A contains an example of a one-dimensional Haar wavelet with eight basis functions over four scales. Note that the same shape of the basis function is shifted at each

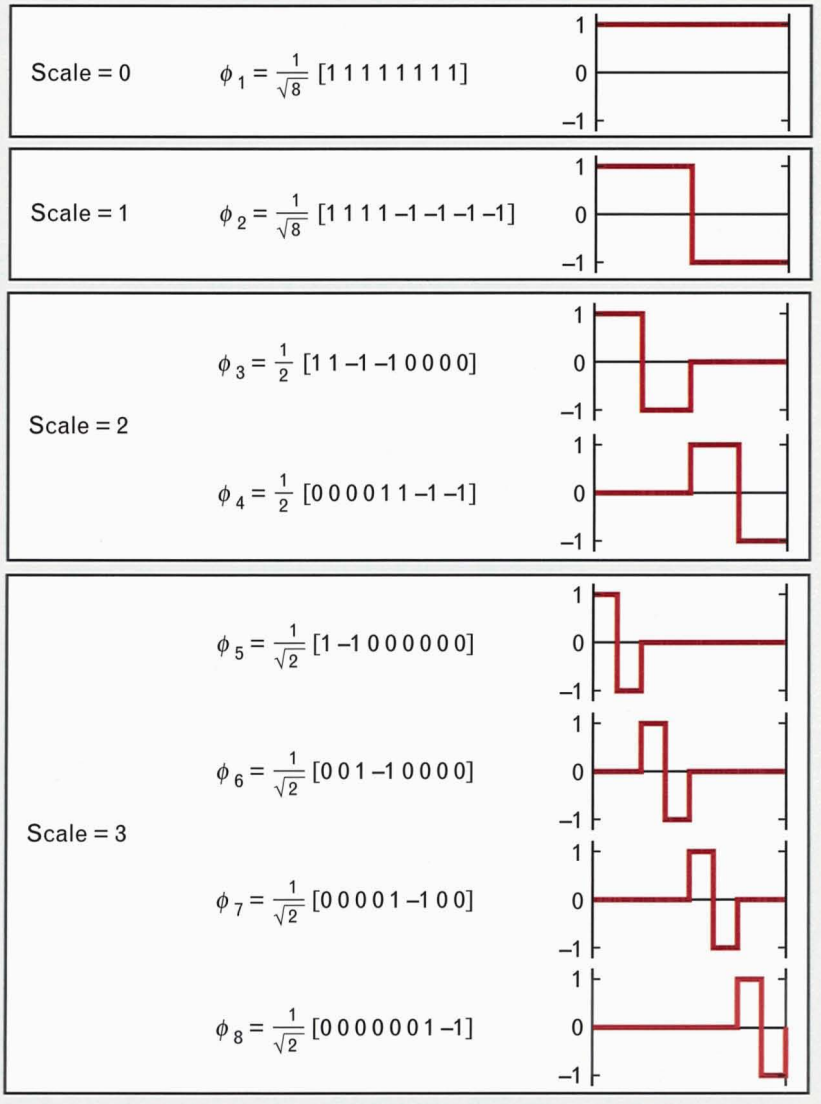


**FIGURE A.** Example of a one-dimensional Haar wavelet with eight basis functions over four scales.

scale and is attenuated with increasing scale. At the finest scale used (scale = 3), the basis functions are sensitive to transitions over just two elements in the input signal. The Haar wavelet has the property that the number of basis functions is always a power of 2.

As mentioned earlier, we can approximate an input signal $y$ by using a linear combination of the basis functions over all scales and shifts:

$$\hat{y}(n) = \sum_{j=1}^{M} c_j \phi_j(n).$$

Because the Haar wavelet forms an orthonormal basis, the coefficients $c$ can be derived with the following equation:

$$c_k = \sum_{n=1}^{N} y(n)\phi_k(n),$$

where $N$ = number of elements in the input vector.

In the current face-recognition application, we have used the shift-dependent coefficients because the faces are always centered in a region of fixed size. For a more general application, a feature set that is not sensitive to the position of the face in the field of view is needed. We can derive a set of features that is invariant to a shift in the signal by calculating the root-mean-square magnitude of the coefficients across all locations (shifts) at a given scale. The length of this feature vector will be the number of scales required to approximate the signal adequately. Further work with such a feature set is necessary to determine if the shift invariance influences the separability of the signals.

*References*

1. O. Rioul and M. Vetterli, "Wavelets and Signal Processing," *IEEE Signal Process. Mag.* **8**, 14 (Oct. 1991).
2. A. Haar, "Zur Theorie der Orthogonalen Funktionensysteme," *Math. Anal.* **69**, 331 (1910).

ages of eight different faces, with each face represented at the same three orientations (–90°, 0°, and 90°) used for training. The clusters formed during each of the 10 training trials were used to initialize the system prior to testing. Figure 11 shows that the best testing performance for the fewest number of clusters formed was obtained with the ACN trained at $\lambda_{train} = 0.93$ (25 clusters formed). At that $\lambda_{train}$ setting, the system performance was 99% correct, 1% unknown, and 0% misidentifications. Table 2 shows the class confusion matrix for testing. Using multiple-frame processing, the system achieved a correct performance of 100%.

We studied the effect of spatial resolution on the testing performance by using different subsampled image sets (processing path A of Figure 3). Figure 12 shows the results. The system performance is constant down to a $16 \times 16$ image size; at smaller sizes the performance degrades rapidly. The results, however, are biased toward "unknown" classifications rather than misidentifications. For a $4 \times 4$ image size, the system classified 14% of the images incorrectly, and 53% as "unknown."

We also tested the system by using Haar wavelet coefficients as the feature vector at various wavelet scales (processing path B of Figure 3). Figure 13 shows the results. Note that the wavelet feature extraction achieved a significant decrease in the number of misidentifications at the lowest resolution (three wavelet coefficients). At this resolution, there were only 5% misidentifications, compared to 14% for the raw $4 \times 4$ imagery.

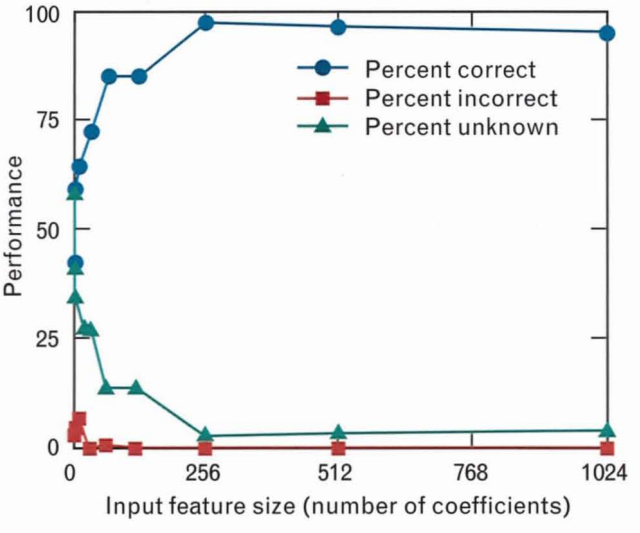


**FIGURE 13.** Testing performance as a function of the number of Haar wavelet coefficients.

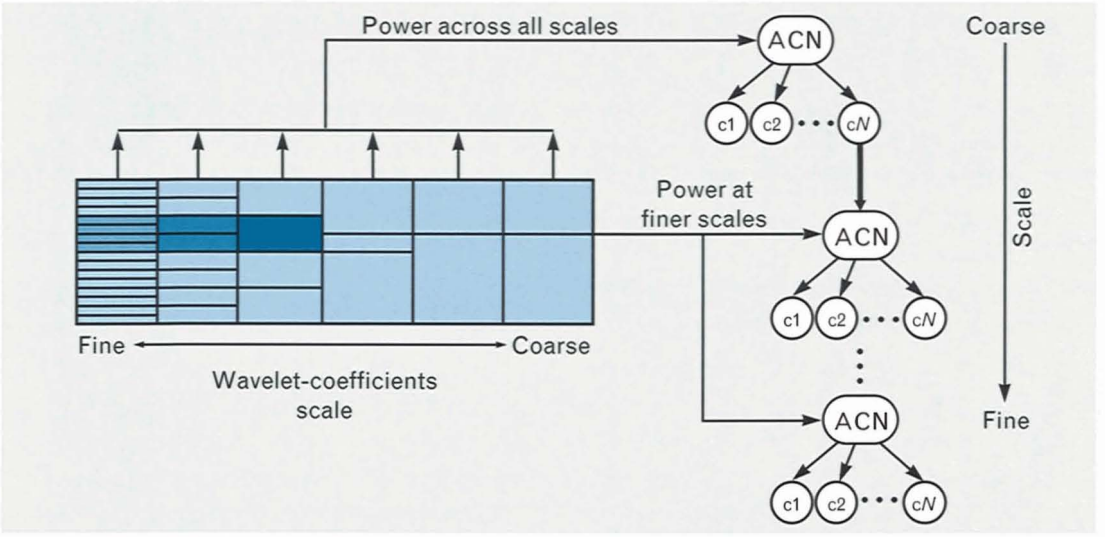


**FIGURE 14.** ACN architecture using a coarse-to-fine hierarchy with clusters c1 through c*N*.

## Summary

We have developed and evaluated a preliminary Automatic Face Recognition (AFR) system that uses the Adaptive Clustering Network (ACN) to separate different subjects automatically by grouping similar faces together into clusters. The ACN slow-learning rule accommodates small variations in the data, e.g., different facial expressions of the same subject, by selectively updating the cluster centers.

To assess the AFR system's performance, we used a small database of imagery in which the subjects were positioned at a fixed distance from the camera against a benign background. Without using feature-extraction techniques, the system identified 99% of the images correctly with no misidentifications. This performance was constant from the full resolution (128 × 128 pixels) imagery down to the subsampled (16 × 16) imagery. For the lowest resolution (4 × 4) imagery, the performance degraded to 33% correct with 14% misidentifications. Using Haar-wavelet feature extraction, the system was able to reduce the number of misidentifications at the lower resolutions while maintaining the performance level at the higher resolutions. With multiple-frame processing, the system performance increased to 100% correct.

## Future Work

Future work will include thorough evaluations of the AFR system on larger populations (the next study will involve 50 to 200 subjects) typically encountered in access-control applications. Datasets need to be collected periodically to track the system performance as the appearance of the subjects changes over time. To maintain a high level of performance for large populations, we will investigate the use of a hierarchy of ACN classifiers to perform multiscale processing. Both fine- and coarse-scale features from wavelet feature extraction could be utilized with this approach. We will also evaluate the fusion of both image and voice data by a system that consists of two separate ACN classifiers, one dedicated to video imagery and the other to features extracted from an audio signal.

### *Hierarchical Architecture*

We could extend the ACN to a radial basis function (RBF) [15] architecture to form an arbitrarily close approximation to any nonlinear mapping. In this approach, the ACN could be used to form the clusters in the first layer of the RBF; a least mean square (LMS) algorithm [16] is typically used to find the weights to the second mapping layer. A drawback of using an RBF architecture is that both the clustering and mapping must be updated to account for new training data.

A more flexible ACN-based architecture creates an additional ACN for every cluster that lacks a distinct peak class and has a reasonable number of entries

(typically 10 or more). The newly created ACN attempts to separate the input patterns by using a different feature set. This architecture forms a tree-like structure that can incorporate different features from the same dataset, e.g., a hierarchy of networks that process coarse to fine features from the dataset. The architecture can also fuse different datasets to improve the overall classification by using different ACNs for the different datasets. The process of adding new ACNs can occur during training so that, when a new dataset becomes available, it can be processed without the need to retrain the system on the previous dataset.

For the AFR application, the hierarchical structure would consist of a first layer of ACNs for distinguishing between general facial characteristics such as a person's overall head shape, the presence of facial hair, and the spatial relation of the eyes to the nose and mouth. The following layers would utilize increasingly finer features to distinguish between people with similar general features. The number of layers following each ACN would be determined directly by the data and would vary for each of the different general facial types. The ACN would automatically determine the necessary general feature types and the successive layers. This approach is well suited to multiresolution representations resulting from wavelet-transform feature extraction, as shown in Figure 14. The first-layer network would use the average RMS power of the wavelet coefficients across all scales, followed by networks that use the coefficients at different shift positions at finer scales. The feature set for networks beyond the first layer would be formed by setting a power threshold to locate (segment) wavelet coefficients of interest across the different shift positions. We will investigate this approach in future work.

### *Data Fusion*

In many cases the AFR task is complicated by, among other factors, the occlusion of a subject or significant changes in a subject's appearance. Thus a more reliable person-recognition system should incorporate another independent measurement, such as that of a subject's voice, to improve the probability of correctly identifying the subject. When an additional measurement of differing dimensionality and data rate is used, a methodology is needed to combine data from the disparate sources (e.g., a video image and an audio signal).

We can structure multiple ACN classifiers to make decisions based on multiple data sources. Figure 15 illustrates an architecture that fuses data from $N$ domains to arrive at a final classification decision. Note that an ACN has been dedicated to each domain, and that the output from each ACN is a class histogram, not a final decision. To arrive at a final decision, the system combines the class histograms from each ACN to form a cumulative class histogram. Using the cumulative histogram, the system calculates the class contrast measure, which indicates the dominance of the peak class (see the subsection "Cluster Class Contrast"), and compares the measure to a given thresh-

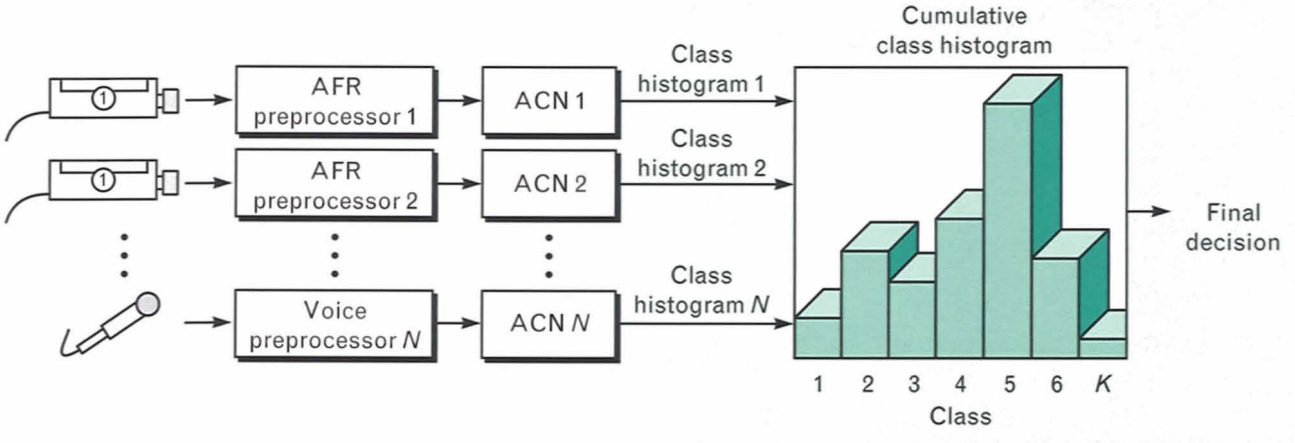


**FIGURE 15.** Data fusion with ACNs.

old. If the class contrast measure exceeds the threshold, then the decision is the peak class; otherwise, a decision is not made.

The architecture of Figure 15 offers several key advantages over typical data-fusion systems:

- The final class decision does not require the simultaneous presence of data in all domains because of the use of a cumulative histogram.
- The ACN class histograms from each domain can be formed from single- or multiple-frame presentations, and the histograms act as a buffering stage to allow different data rates in each domain.
- New domains can be incorporated easily without a restructuring or retraining of the existing system because the data fusion occurs within the cumulative histogram.
- New data can be incrementally introduced into each domain without affecting the remaining domain classifiers.

We expect the performance of a person-recognition system that uses both video imagery of a subject's face and an audio signal of the subject's voice to surpass the performance of the single-channel AFR system. The audio signal will complement the video imagery by providing a class histogram even when the subject's face is obscured or outside the camera's field of view.

## Acknowledgments

Alfred Gschwendtner, Group Leader of the Opto-Radar Systems Group, supported and encouraged this research, Lorraine Prior provided software for controlling the video digitizer board, and Danny Corbosiero assembled the data-collection devices used to obtain the preliminary database of face imagery. The authors are also grateful to the Lincoln Laboratory staff who participated in posing for this study.

This work was sponsored by the Department of the Air Force.

# REFERENCES

1. B. Moghaddam, K.J. Hintz, and C.V. Stewart, "Space-Filling Curves for Image Compression," *SPIE* **1471**, 414 (1991).
2. J.H. Shapiro, private communication.
3. T. Poggio and F. Girosi, "A Theory of Networks for Approximating and Learning," Artificial Intelligence Laboratory Memo 1140, MIT (1989).
4. D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**, eds. D.E. Rumelhart and J.L. McClelland (MIT Press, Cambridge, MA, 1986), pp. 318–362.
5. S. Phuvan, T.K. Oh, N. Caviris, Y. Li, and H.H. Szu, "Texture Analysis by Space-Filling Curves and One-Dimensional Haar Wavelets," *Opt. Eng.* **31**, 1899 (1992).
6. D.L. Reilly, L.N. Cooper, and C. Elbaum, "A Neural Model Category Learning," *Biol. Cybern.* **45**, 35 (1982).
7. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis* (John Wiley, New York, 1973).
8. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press, Orlando, FL, 1972).
9. T. Kohonen, "An Introduction to Neural Computing," *Neural Networks* **1**, 3 (1988).
10. Private communications.
11. *ACN User Guide* (MIT, Cambridge, MA, 1994).
12. M.M. Menon, E.R. Boudreau, and P.J. Kolodzy, "An Automatic Ship Classification System for ISAR Imagery," *Linc. Lab. J.* **6**, 289 (1993).
13. R.J. Schalkoff, *Digital Image Processing and Computer Vision* (John Wiley, New York, 1989).
14. P.F. Suarez, "Face Recognition with the Karhunen-Loeve Transform," thesis, Air Force Institute of Technology (Dec. 1991).
15. J.R. Goble, "Face Recognition Using the Discrete Cosine Transform," thesis, Air Force Institute of Technology (Dec. 1991).
16. M. Menon and E. Boudreau, "Pattern Recognition with Statistical Classification," application no. 08/122/705, U.S. Patent Office (filed 16 Sept. 1993).

**MURALI M. MENON**
is currently a research staff member in the Opto-Radar Systems Group. He received a B.S., an M.S., and a Ph.D. degree in chemical engineering from Case Western Reserve University. His research interests include applied pattern recognition, signal processing, and image processing, with special interests in wavelets and artificial neural networks.

In 1987 Murali joined Lincoln Laboratory, where he has worked on applications of neural networks for processing sensor data, including the design of automatic target recognition (ATR) systems. He is currently involved with the transfer of technology to the commercial sector, and has established a cooperative research and development agreement with a company in the Boston area to develop an automated quality control system based on neural networks and advanced signal processing methods. In 1990 he chaired a workshop on the applications of neural networks to real-world machine-vision problems at the IEEE Neural Information Processing Systems (NIPS) conference. He is in the process of writing a book chapter entitled "The Neocognitron: Prospects for Scaling to Practical Applications," for *Progress in Neural Networks* **II**, to be published in 1994. Murali is a member of the IEEE.

**ERIC R. BOUDREAU**
received a B.S. degree in electrical engineering from Worcester Polytechnic Institute, and worked at COMPDATA Services Corp. before joining Lincoln Laboratory four years ago. He is currently an assistant staff member in the Opto-Radar Systems Group, where his focus of research has been on the development and evaluation of automatic target recognition (ATR) systems. He has worked on the detection of stationary targets in laser-radar and passive-infrared imagery, and recently participated in the implementation of a portable global surveillance demonstration system. Eric is a member of Eta Kappa Nu.