Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture

Mark Rosenblum, Yaser Yacoob, Larry Davis
Computer Vision Laboratory
University of Maryland
College Park, MD 20742

Presented at the IEEE Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, Nov. 1994

Abstract

In this paper a radial basis function network architecture is developed that learns the correlation of facial feature motion patterns and human emotions. We describe a hierarchical approach which at the highest level identifies emotions, at the mid level determines motion of facial features, and at the low level recovers motion directions. Individual emotion networks were trained to recognize the 'smile' and 'surprise' emotions. Each emotion network was trained by viewing a set of sequences of one emotion for many subjects. The trained neural network was then tested for retention, extrapolation and rejection ability. Success rates were about 88% for retention, 73% for extrapolation, and 79% for rejection.

1 Introduction

Visual communication plays a central role in human communication and interaction. This paper explores methods by which a computer can recognize visually communicated facial actions- facial expressions. Developing such methods would contribute to human-computer interaction and other applications such as: multi-media facial queries, low-bandwidth transmission of facial data and face recognition from dynamic imagery.

Research in psychology has indicated that at least six emotions are universally associated with distinct facial expressions. Several other emotions, and many combinations of emotions, have been studied but remain unconfirmed as universally distinguishable. The six principle emotions are: happiness, sadness, surprise, fear, anger, and disgust (see Figure 1). In this paper we focus on these emotions.

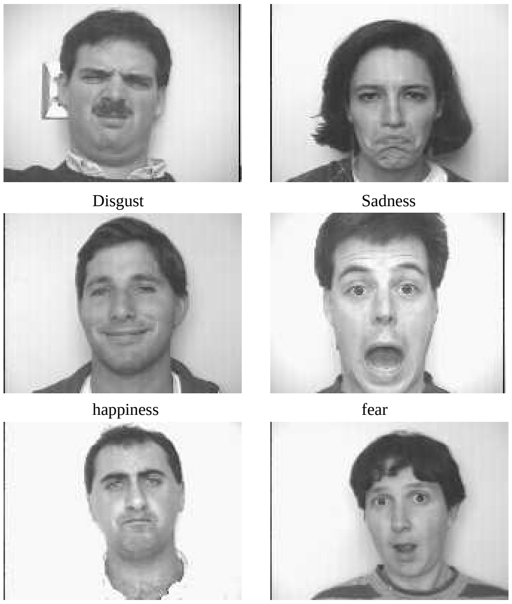

Disgust
Sadness
happiness
fear
Anger
Surprise

Figure 1: Six universal expressions

Most psychology research on facial expression has been conducted on "mug-shot" pictures that capture the subject's expression at its peak. These pictures allow one to detect the presence of static cues (such as wrinkles) as well as the position and shape of the facial features. Few studies have directly investigated the influence of the motion and deformation of facial features on the interpretation of facial expressions (a review of the relevant psychological aspects of recognizing facial expressions appears in [11]). Bassili [2] suggested that motion in the image of a face would allow emotions to be identified even with minimal infor-

The support of the Advanced Research Projects Agency (ARPA Order No. 6989) and the U.S. Army Topographic Engineering Center under Contract DACA76-92-C-0009 is gratefully acknowledged.

mation about the spatial arrangement of features. The subjects of his experiments viewed image sequences in which only white dots on the dark surface of the person displaying the emotion are visible. The reported results indicate that facial expressions were more accurately recognized from dynamic images than from a single static image. Whereas all expressions were recognized at above chance levels in dynamic images, only happiness and sadness were recognized at above chance level in static images. As illustrated in Figure 2, Bassili identified principle facial motions that provide powerful cues to the subjects for recognizing facial expressions. These results do not explicitly associate the motion patterns with specific face features or muscles since such information was unavailable to the experiment subjects.

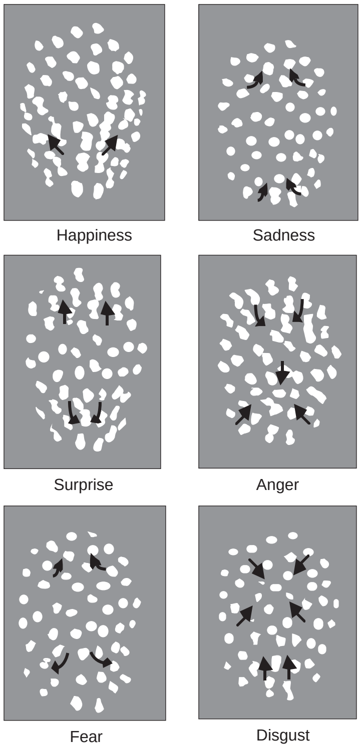


Figure 2: Motion cues for facial expression [2]

Building on these results we explore the potential of motion analysis in an autonomous system.

The problem of recognizing facial expressions has recently attracted attention in the computer vision community [4,8,10].

Yacoob and Davis proposed an approach for analyzing and representing the dynamics of facial expressions from image sequences [10]. This approach is divided into three stages: locating and tracking prominent facial features (i.e., mouth, nose, eyes, and brows), using optical flow at these features to construct a mid-level representation that describes spatio-temporal actions, and applying rules for classification of mid-level representation of actions into one of six universal facial expressions. On a sample of 46 image sequences of 32 subjects displaying a total of 105 emotions, the system achieved a recognition rate of 86% for 'smile,' 94% for 'surprise,' 92% for 'anger,' 86% for 'fear,' 80% for 'sadness,' and 92% for 'disgust.' Blinking detection success rate was 65%.

Connectionist architectures have been used in visual classification problems with great success [6,3]. The classification of visual imagery, however, has mainly focused on static imagery. Seibert and Waxman [7] recently developed a system that performed object recognition using the object's rigid motion. The neural network learned correlations between different aspect views of an object, and as the network observed a sequence of the object moving in space, it accumulated evidence of the object it was viewing.

The work reported here explores the use of a connectionist learning architecture for identifying the non-rigid motion pattern characteristics of facial expressions. The neural network views variable length sequences of images of a human subject instead of a single static image. The connectionist approach could replace the expert rules developed in [10], and may allow developing person-specific learning capabilities.

2 Overview of our approach

The following constitute the framework within which our approach for analysis and recognition of facial expressions is developed:

- The face is viewed from a near frontal view throughout the sequence.
- The overall rigid motion of the head is small between any two consecutive frames.
- The non-rigid motions that are the result of face deformations are spatially bounded, in practice, by an nxn window between any two consecutive frames.

The system is similar to [10] in the tracking and optical flow computation but differs in the analysis and interpretation of motion patterns. The system is composed of the following components:

- Optical flow computation: Optical flow is computed at the points with high gradient at each frame. Our algorithm for flow computation is based on a correlation approach proposed by Abdel-Mottaleb et al. [1]. It computes subpixel flow assuming that the motion between two consecutive images is bounded within an nxn window.

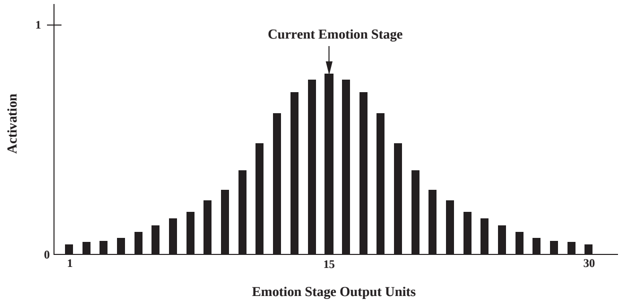
- Region tracking: We assume that, for each feature, we can initially compute a rectangular region that encloses it. Such an algorithm has been recently proposed for range data by Yacoob and Davis [9] and a similar algorithm could be developed for intensity images. Our algorithm tracks these regions through the remainder of the sequence. The tracking is based on the localization of points with high gradient and the optical flow fields computed at these points.
- A connectionist architecture for learning what facial motion information and relations are important to the determination of emotion. This system learns using a training set which consists of sequences of images from a diverse set of human subjects experiencing the same emotion.

3 The Inputs and Outputs of the NN

We perform three stages of preprocessing on the input sequence before providing input to the neural network. The first stage generates a sequence of images which represents the instantaneous optical flow of the image sequence. The second stage extracts, using tracking techniques, the important facial features from the optical flow sequence (i.e., the right and left eyebrows and the mouth). The third stage performs a log-polar transformation on the feature motion images of the sequence. This transformation compresses the outer extremities of the feature images for the purpose of reducing the effects of size variance. Size variance occurs because of subjects' varying distance from the camera, motion during the image sequences, and the natural variation in the sizes of subjects' features.

The output of the system could be structured so an output is associated with each emotion. This representation, however, does not provide enough spread for the neural network to learn effectively. An intermediate output representation is required to provide this spread. The intermediate output representation chosen represents the stage of an emotion to which an input image of a sequence belongs. The activation of an output unit in this representation corresponds to the network's confidence that the emotion of the current sequence is in the stage corresponding to the particular output unit.

Pomerleau [6] found that when there exists a proximal relation between output units the supervised learning unit activations should reflect this relation. In our application, an output unit represents a stage of an emotion and the stages are related by the obvious temporal proximal relation. If the current training vector is to reflect that the emotion is currently in stage N, then output unit N should be set with the greatest activation, while output units $N+1$ and $N-1$ should be set with a slightly lower activation and so on until the boundaries of the output vector are reached. Pomerleau used a Gaussian function to set the training activations. We also used a Gaussian, placing its peak on the current stage in the output training vector and setting the output units corresponding to the value of the Gaussian at that position in the vector (see Figure 3).



The Gaussian weighting function applied to desired-output vector corresponding to the fifteenth stage of an emotion sequence

Figure 3: The Gaussian weighted output vector

4 The Network Architecture

The complexity of recognizing facial expressions required dividing the emotion detection problem into three layers of decomposition. The first layer is by emotion (see Figure 4), and occurs at the network level- we train a separate network for each emotion (i.e., six separate networks in total). During training, a network in this layer is only exposed to one emotion for multiple subjects. The second layer is at the facial component level. This decomposition is internal to each of the emotion tuned networks. Each emotion network is broken into subnetworks, where each subnetwork specializes in a particular facial component. Since we are focusing on three facial components, each emotion network consists of three subnetworks. A component tuned subnetwork only uses the portion of the input vector that corresponds to its component specialization. The third layer is by direction sensitivity, and further decomposes the component subnetworks. In other words, these "subsubnetworks" are sensitive to one direction of motion for a specific preassigned facial component for a specific emotion. In order to capture all resultant motions, we use the four direction sensitivities of up, down, right, and left.

The fusion of information from each of the six emotion tuned networks is performed by a process external to these networks, and can be connectionist or hand coded in nature. We developed a hand coded scheme (discussed below) which combines the outputs of all six emotion networks. The fusion of information from the internal subnetworks is done internally in each of the emotion networks. The fusion is done implicitly through the coupling of these component subnetworks through the output units of the individual emotion network.

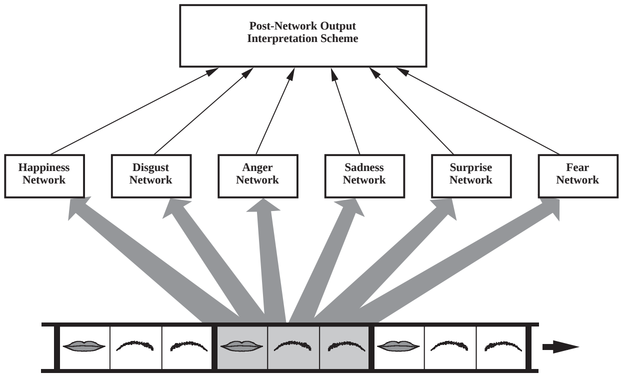


Figure 4: The hierarchy of networks based on emotion decomposition

5 The Basic Building Block

Because of their ability to directly represent prototypical situations of the application in the receptive field centers, we chose to use a modified version of the radial basis function network (RBFN) [5] as the architecture for the basic building blocks. In the following section we discuss the enhancements we made to the basic RBFN architecture to handle the temporal relations associated with this problem.

An RBF receptive field is a response region in N dimensional input space, with an N component center coordinate. The input space can be considered as an image space, since the input units are clamped directly to the values of the pixels of an image feature retina. Since each coordinate in image space corresponds to a unique image on the image retina, the receptive field centers also correspond to unique images on the image retina, and these function as the application templates. The maximum response of a receptive field occurs when an input image is situated at the same location as the center of the receptive field, and the response degrades in a Gaussian fashion as the Euclidean distance of the input image to the receptive field center increases.

5.1 The Spatio-Temporal Building Blocks

The RBFN architecture is not well suited to handle temporal relations. A significant part of the task of analyzing sequences of images, is being able to relate information from one frame to the next. Thus, enhancements that will allow past information to contribute to the current response are necessary. For example, in the 'surprise' emotion, the eyebrows move downward at the end of the emotion, and in 'anger' the eyebrows move downward at the beginning of the emotion. In order to determine whether the eyebrows are moving downward in the 'surprise' or 'anger' emotion, it is necessary to determine what happened to the eyebrows before they moved downward.

Past information is incorporated into the input vector by using feedback from the previous state of the input vector multiplied by a decay constant. Input units that use self feedback are called "context units"[3]. The activation function for each input unit in our architecture is:

$$C_i(t) = \begin{cases} 1 & \text{if } \alpha C_i(t-1) + I_i(t) > 1 \\ \alpha C_i(t-1) + I_i(t) & \text{if } \alpha C_i(t-1) + I_i(t) < 1 \end{cases}$$

where $C_i(t)$ is the activation of input unit i at time t, $C_i(t-1)$ is the activation of input unit i at time $t-1$, α is the decay constant, and $I_i(t)$ is the current input to unit i at time t. The decay constant is set so that remnants of previous motions linger for a portion of the sequence. If motion occurs for several iterations at the same pixel location in the input image, the input unit activation that corresponds to that pixel location becomes saturated and is set to the maximum activation level of one.

Each emotion subnetwork consists of receptive fields tuned to the particular facial feature, and the weights fully connecting those receptive fields to the output units. The set of receptive fields corresponding to a particular feature and for each motion direction are further tuned to become sensitive to only portions or subsequences of the input sequence. In other words the component receptive fields become sensitized to stages of the emotion sequence for the component and direction they are assigned to. A receptive field center image or template is set by integrating the motion images for a subsequence of the receptive field's assigned facial component and motion direction sensitivity. Any position in the summed image that has a value greater than zero is set to one. This is similar to how the input vector is calculated using a decay constant, except in this case the decay constant is set to one, and the subsequence of images has a start and end frame in the sequence. Figure 5 shows how a subsequence is used to set the center of a receptive field for a simple sequence of a ball moving across the retina. It is important to note that an input vector can never perfectly match a receptive field center template unless the decay constant for calculating the input vector is set to one.

In order to minimize the problem of overloading and under-utilizing receptive fields, we defined a parameter which represented the minimum number of pixels that must be turned on during the image summing stage to set a receptive field center with the summed image. If the number of "on" pixels during the summation crossed over this minimum threshold, no additional images were incorporated into the summed image and a receptive field center was set with the current accumulated image. The result was that portions of the sequence where significant motion occurred were spread over more center templates for higher temporal resolution, and portions of the sequence where little motion occurred were fit into fewer center templates for lower temporal resolution.

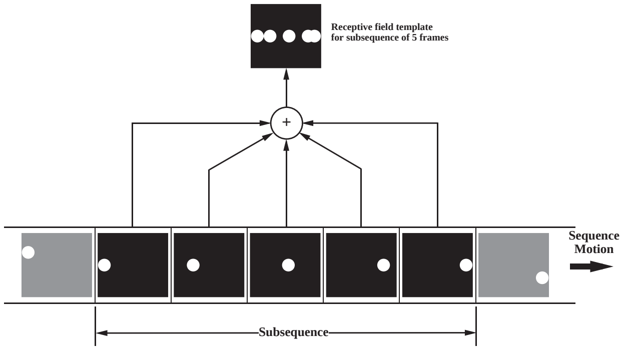


Figure 5: The approach to setting the receptive field centers from subsequences.

In addition to using past information to set the center positions of the receptive fields, past information is also used to determine the activations of the receptive fields during the training and usage modes. The determination of a receptive field activation is similar to the determination of the activation of an input unit, in that the activation from the previous time step is factored into the activation at the current time step using a decay constant. The activation of a receptive field in our architecture is determined by the following equation:

$$\rho_i(\vec{x}_{t+1}, t{+}1) = \begin{cases} 1 & \text{if } \gamma\rho_i(\vec{x}_t, t) + V_i(t) > 1 \\ \gamma\rho_i(\vec{x}_t, t) + V_i(t) & \text{otherwise} \end{cases}$$

where β is the decay constant and

$$V_i(t) = \frac{\exp\left(-\beta_i\ (\vec{x}_{t+1} - \vec{a}_i)^T(\vec{x}_{t+1} - \vec{a}_i)\right)}{\sum_j \exp\left(-\beta_j\ (\vec{x}_{t+1} - \vec{a}_j)^T(\vec{x}_{t+1} - \vec{a}_j)\right)}$$

Like the input vector determination, the receptive field response can become saturated, in which case the activation is set to one.

6 Experiments and Results

We use two forms of analysis of the emotion network outputs: absolute and relative analyses. For our preliminary experiments, we only trained two emotion networks; one for the 'smile' emotion, and another for the 'surprise' emotion. The test stage included image sequences of 'smile,' 'surprise,' and 'anger' emotions. Anger was used as a null reference since neither of the trained networks was tuned for 'anger.'

Before we discuss the methodology and results we define the terminology used. The term *familiar-face* indicates that the face used is that of a person that the system has seen in the *training* session. For such a face there can be two types of sequences, *familiar-* and *unfamiliar-sequences*. The former denotes those image sequences that were used in the training, and the latter indicates these sequences of the *familiar-face* that are new to the system.

6.1 Absolute Analysis

In order to evaluate the performance of the neural network architecture, we conducted experiments that measure the network's *retention, extrapolation,* and *rejection* ability. Retention refers to the ability of the network to perform successfully on familiar sequences. Extrapolation refers to the ability of the network to perform successfully on sequences of unfamiliar faces. Rejection refers to the ability of the network to reject a sequence that did not express the emotion that the network was tuned for.

To measure the performance of the system relative to the above criteria we divided the experiments into four categories. The first category encompassed familiar sequences, and it measured the networks retention ability. In the second category, unfamiliar faces were tested in order to measure the extrapolation ability. The third category included unfamiliar sequences of familiar faces and it measured a smaller increment of extrapolation than the second category. The fourth category included sequences of emotions that the tuned network did not specialize in (these can be for any type of emotion and face) and it measures the rejection rate of the network.

For each of the 'smile' and 'surprise' emotions, we trained two networks that only differed in receptive field width, and we tested each network using the four test categories. Each network was trained for 100,000 iterations, and the receptive field widths for SMILENET 1 and SURPNET 1 were larger than the receptive field widths for SMILENET 2 and SURPNET 2 (see Table 1). The 'smile' and 'surprise' networks were trained with 20 and 14 subjects, respectively. The output vector for each network represented 40 stages of an emotion. We used the criterion of at least seven stages being turned on to signify that the network recognized the emotion of a sequence, and we used an iteration confidence threshold of 0.155 to increment a stage counter for a frame of the sequence. Table 2 shows the results from the absolute analysis.

network	mouth rf widths	eyebrow rf widths
SMILENET 1	1	1
SURPNET 1	1	1
SMILENET 2	.694	.563
SURPNET 2	.694	.563

Table 1: The *relative* receptive field width settings

network	familiar seq	unfam. face	unfam. seq.	foreign expr.
SMILENET 1	16/20=80%	2/4=50%	7/7=100%	29/41=71%
SURPNET 1	13/14=93%	5/6=83%	3/3=100%	39/52=75%
SMILENET 2	16/20=80%	2/4=50%	4/7=57%	32/41=78%
SURPNET 2	13/14=93%	2/6=33%	3/3=100%	46/53=87%

Table 2: The results of the absolute analysis

In Table 3 we further break down category 4 to compare the rejection rate of 'anger,' 'smile' and 'surprise.' The results indicate that 'anger' got the best rejection rates.

network	anger	surprise	smile
SMILENET 1	16/18=89%	13/23=57%	–
SURPNET 1	17/18=94%	–	22/31=71%
SMILENET 2	16/18=89%	16/23=70%	–
SURPNET 2	18/18=100%	–	28/31=90%

Table 3: The further breakdown of category 4

The results indicate that the retention rates are higher than the extrapolation rates. In Table 3 the rejection rates for 'surprise' were better than those for 'smile' for the three emotions. For the 'smile' and 'surprise' networks with the same receptive field widths, the 'surprise' network had a much higher rejection rate of the 'smile' emotion than the 'smile' network had of the 'surprise' emotion. The larger detectable motion of the 'surprise' emotion improved performance for all four test categories, thus, improving retention, extrapolation and rejection of the 'surprise' networks over the corresponding width size 'smile' networks.

Also from Table 2 and Table 3 we can see that larger receptive field widths enhanced extrapolation abilities of the networks (categories 2 and 3), but at the same time reduced the retention and rejection rates (categories 1 and 4). Since one of the main goals of this research was to determine if a network could learn the commonalities of an emotion over a wide population from a small sample set, wider receptive field widths are better suited for our application. On one hand, if the receptive fields widths for a network are too large, thus over-generalizing, then all the receptive fields will respond with equally large activations, and the categorizing ability of the network is lost. On the other hand, if the receptive field widths are too small, the receptive fields will respond crisply to training patterns, but will have negligible responses to test patterns that only vary slightly from the training patterns, thus possessing no generalization ability. Therefore a retention/extrapolation trade-off exists between large and small receptive field widths.

6.2 Relative Analysis

Since it was our intention to teach a network extrapolation instead of retention, we focused our relative analysis on networks SMILENET 1 and SURPNET 1, which had better extrapolation performance because of their relatively larger receptive field widths. The relative analysis is dependent on the results of the absolute analysis. Similarly, in the relative analysis we defined four test categories to measure retention, extrapolation and rejection. The first category tests familiar sequences of 'smile' or 'surprise.' The second category tests sequences of unfamiliar faces. The third category tests unfamiliar sequences of familiar faces in at least one of the two training sets. The fourth category tests expression sequences foreign to both networks. Since we trained on the 'smile' and 'surprise' emotions, the only emotion sequences in the fourth category were those of 'anger.'

In the relative analysis, we compare the responses of the two networks; the thresholding is done in the earlier absolute stage of analysis. In the case of two networks, we have four possible combinations of outputs: Yes/Yes, No/Yes, Yes/No, and No/No (where a "Yes" signifies that a network recognizes a sequence as its specialization emotion, and a "No" signifies the network did not recognize the emotion). The Yes/No and No/Yes responses are straight forward, in that the relative emotion response is taken as the emotion of the network that responded with a "Yes". The No/No relative response also represents a clear answer that neither network recognizes the emotion of the sequence. The Yes/Yes response is ambiguous however, and is resolved by the relative analysis. To resolve the Yes/Yes ambiguity, the absolute output statistics of each network for the ambiguous sequence are compared. We used the number of stages turned on as the comparison statistic. The network that had the highest number of stages turned on was declared the winning network, and the resultant emotion was determined to be the specialization emotion of that network. The Yes/Yes ambiguous response was possible in test categories 1, 2, and 3; thus, the relative ambiguity resolution was expected to improve the performance for these three categories. Table 4 shows the results from the relative analysis after the ambiguity resolution for categories 1, 2 and 3.

familiar seq.	unfam. face	unfam. seq.	foreign expr.
30/34=88%	11/15=73%	11/12=92%	14/18=77%

Table 4: The results of the relative analysis

In order to compare the absolute and relative analyses, the absolute performances for the SMILENET 1 and the SURPNET 1 are combined into one performance measure based on a weighted average of the number of test cases for each network in each test set category, except category 4, since it does not apply. Table 5 shows the combined results from the absolute analysis compared with the results from the relative analysis for each category. The results show an expected slight performance improvement for categories 1 and 2, and an unexpected slight reduction in performance for category 3 between the absolute and relative analysis. The reduction in performance for category 3 was caused by the incorrect network having a higher score than the correct network.

analysis	familiar seq.	unfamiliar face	unfamiliar seq.
absolute	85%	70%	100%
relative	88%	73%	92%

Table 5: Comparison of the absolute results with the relative results

7 Conclusion

In this paper, we developed a human emotion detection system based on radial basis function network. By training the network, it was able to learn the correlations between facial feature motion patterns and specific emotions. In order to capture the temporal relations that are important to emotion detection several enhancements were made to the underlying network architecture. In order to make the problem more tractable, the emotion detection problem was decomposed at several levels: emotion, facial feature, and motion direction sensitivity levels. For our preliminary experiments, of the six universal human emotion expressions, we trained networks to recognize the 'smile' and 'surprise' emotions. Our experiments were designed to test a network's retention, extrapolation, and rejection abilities. The analysis of the experimental results were conducted in an absolute and a relative mode. The purpose of the relative mode was to improve overall emotion detection over the absolute mode by comparing all network outputs and picking a winner.

Our experiments suggest that networks tuned better on emotions that involved more pronounced motion. We also found that a trade-off existed between large and small receptive field widths. Large widths improved extrapolation, while degrading retention and rejection, while small widths had the opposite effect.

References

[1] M. Abdel-Mottaleb, R. Chellappa, and A. Rosenfeld, "Binocular motion stereo using MAP estimation", *IEEE CVPR*, 321-327, 1993.

[2] J.N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality and Social Psychology*, Vol. 37, 2049-2059, 1979.

[3] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison Wesley, 1991.

[4] K. Mase, "Recognition of facial expression from optical flow," *IEICE Transactions,* Vol. E 74, No. 10, 3474-3483, 1991.

[5] J. Moody and C. Darken 1988. "Learning with Localized Receptive Fields," *Proceedings of the 1988 Connectionist Models Summer School*, eds Touretzky, Hinton, and Sejnowski. Morgan-Kaufmann Publishers, 1988.

[6] D.A. Pomerleau, *Neural Network Perception for Mobile Robot Guidance,* Ph.D. thesis, Carnegie Mellon University, Department of Computer Science, 1992.

[7] M. Seibert and A.M. Waxman, "Adaptive 3-D Object Recognition from Multiple Views", *IEEE PAMI*, Vol. 14, No. 2, 107-124.

[8] D. Terzopoulos, and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE PAMI*, Vol. 15, No. 6, 569-579, 1993.

[9] Y. Yacoob, and L.S. Davis, "Labeling of human face components from range data," *IEEE CVPR*, 592-593, 1993.

[10] Y. Yacoob, and L.S. Davis, "Computing Spatio-Temporal Representations of Human Faces" *IEEE CVPR*, 70-75, 1994.

[11] Y. Yacoob, and L.S. Davis, *Recognizing Human Facial Expressions,* Technical Report CAR-TR-706, Center for Automation Research, University of Maryland, College Park, May 1994.