A Simple Neural Network Models Categorical Perception of Facial Expressions

Curtis Padgett and **Garrison W. Cottrell**
Computer Science & Engineering 0114
University of California, San Diego
La Jolla, CA 92093-0114
{cpadgett,gary}@cs.ucsd.edu

Abstract

The performance of a neural network that categorizes facial expressions is compared with human subjects over a set of experiments using interpolated imagery. The experiments for both the human subjects and neural networks make use of interpolations of facial expressions from the Pictures of Facial Affect Database [Ekman and Friesen, 1976]. The only difference in materials between those used in the human subjects experiments [Young et al., 1997] and our materials are the manner in which the interpolated images are constructed – image-quality morphs versus pixel averages. Nevertheless, the neural network accurately captures the categorical nature of the human responses, showing sharp transitions in labeling of images along the interpolated sequence. Crucially for a demonstration of categorical perception [Harnad, 1987], the model shows the highest discrimination between transition images at the crossover point. The model also captures the shape of the reaction time curves of the human subjects along the sequences. Finally, the network matches human subjects' judgements of which expressions are being mixed in the images. The main failing of the model is that there are intrusions of "neutral" responses in some transitions, which are not seen in the human subjects. We attribute this difference to the difference between the pixel average stimuli and the image quality morph stimuli. These results show that a simple neural network classifier, with no access to the biological constraints that are presumably imposed on the human emotion processor, and whose only access to the surrounding culture is the category labels placed by American subjects on the facial expressions, can nevertheless simulate fairly well the human responses to emotional expressions.

Introduction

Research into the nature of the perception of facial images in humans (in tasks such as identification of the subject as well as what expression is being displayed) has uncovered considerable evidence that the process is categorical [Beale and Keil, 1992, Etcoff and Magee, 1992, Young et al., 1997]. This research has focused on how human responses change over a sequence of interpolated imagery between two prototypes. The studies have consistently reported categorical transitions in the sequences. Categorical perception (CP) typically involves demonstrating a boundary region where responses by subjects change rapidly and where the subjects show a correspondingly greater ability to discriminate the stimuli [Liberman et al., 1957, Harnad, 1987].

In previous work we evaluated several types of image features in terms of their efficacy as inputs to neural network models of emotion recognition. The facial expression images we used were from the Pictures of Facial Affect (PFA) database [Ekman and Friesen, 1976]. The categorization rates of human subjects in a six-way forced choice labeling of the images [Ekman and Friesen, 1977] (provided with the PFA) were used by the model as targets for the emotion categories. The best network correctly recognized 86.2% of the expressions displayed in novel face images [Padgett and Cottrell, 1997]. We then used this model to predict human responses to constructed images that dissolve[1] from one facial expression image to another [Padgett et al., 1996, Adolphs et al., 1998]. When tested on pixel-averaged transitions between facial expressions of the same subject, the model predicted that some transitions would be less "categorical" than others, with shallower transition curves [Padgett et al., 1996]. Human responses on the same pixel-averaged stimuli show similar variations [Adolphs et al., 1998]. Human subject studies making use of the Ekman and Friesen prototypes also show categorical responses using morph sequences of line drawings extracted from the Ekman and Friesen images [Etcoff and Magee, 1992], and using image-quality morph sequences that appear as natural as the original images [Calder et al., 1996, Young et al., 1997].

In one of the most extensive studies with human subjects, Young et al. (1997, henceforth "Megamix") show that image-quality morph sequences between six emotional expressions (Happy, Sad, Afraid, Angry, Surprised, and Disgusted) and "neutral" expressions exhibit categorical behavior. In contrast to Etcoff and Magee's work, they used photo quality images instead of line drawings. In contrast to Calder et al., *all* possible transitions between emotion pairs for a single subject ("JJ") from the PFA database (including neutral) were tested. This comprehensive study of human responses to facial expressions is the inspiration for our current study. In the following sections we review the results from the Megamix study in more detail, describe the neural network model from which we develop the comparison, and present our results.

Review of "Megamix"

The Megamix study is important as it exhaustively examined the transition space between all six pairs of emotions in the PFA plus "neutral" faces. The study provided the most in-depth look at how humans classify morph stimuli and their ability to discern differences within and between class boundaries. Although the stimuli were limited to a single individual's expressions (the "JJ" images in the PFA) and a rather coarse step size between the images along the transition, the amount and kind of data collected was quite large, and is thus extremely useful.

The focus of the Megamix study was in demonstrating that two dimensional accounts of classifying emo-

[1] "Dissolve" is a term from graphics denoting a fade from one image while fading into another. We use this term to distinguish our linear pixel-average transitions from image-quality morphing, an inherently nonlinear process.

Figure 1: Example dissolve sequences of subject JJ from the Facial Affect database. All seven emotions used in the study are shown here. The image sequences are linearly interpolated between the two database images at each extreme.

tions [Russell, 1980] based on a multi-dimensional scaling (MDS) of similarity ratings of emotion categories do not adequately account for the observed boundary behavior between emotions. MDS results in a "circumplex" of emotions, a two-dimensional scaling solution where emotions are arranged around a circle in the scaling space. Accounts based on this would suggest morphing between pairs of emotions on opposite sides of the circumplex would pass through a neutral space in the center. On the contrary, all emotion pairs showed categorical behavior with few intrusions from other categories [Young et al., 1997].

For this study, we are interested in comparing the reported results of the human subject experiments in Megamix to the neural network model used in our previous study [Padgett et al., 1996]. The data used in their experiments consisted of morphed imagery from PFA. A single subject in database ("JJ") served as the endpoints for the transition sequences. In their Experiment 2, image-quality morphs were constructed between all six emotions plus neutral (Experiment 1 just used the six emotion prototypes as morph endpoints). Step sizes of 90%, 70%, 50%, 30%, and 10% were used between each pair of endpoints (105 unique images). These were presented in random order to subjects, who made a 7-way forced choice between the six emotion labels and neutral. An example of the human subject response curves is given in Figure 3 (middle). Response times (RT's) were also recorded. They found the resulting RT curves were "scalloped", with the fastest RT's near the prototype emotion, dropping off farther from the prototype.

In their Experiment 3, subjects were required to discriminate (same/different judgements) simultaneously presented images that were one step away from each other along the transitions. The subjects showed better discrimination near category boundaries than near prototypes, a standard requirement for categorical perception.

Finally, in Experiment 4 of Megamix, Young et al. tested the extent to which their subjects could tell what two emotions were represented in the morph images. This is important because, if the images are perceived categorically, one expects that subjects should be poor at judging what other emotion is mixed into the image. They asked the subjects to give three responses to an image: which emotion it was closest to, then the next closest emotion, then the next, scored as 3, 2, and 1, respectively. They included the prototype images as well, in order to be able to control for the intrinsic similarity between certain emotions. If a surprise image already looks like fear, for example, this could bias the results. By collecting the three scores for the prototypes, they could subtract off the response the prototype engendered to other emotion categories. These difference scores were then averaged across all emotions, and the average responses to the prototype being moved towards plotted. Their data is plotted in Figure 5. It clearly shows the subjects are sensitive to the secondary category in the images.

Neural Network Model

Although we were unable to obtain the Megamix morph sequences, we had previously developed *dissolve* sequences for testing on the transition behavior between emotion pairs. The dissolves are a weighted average of corresponding pixels between two expression prototype images of the same individual. The transitions are produced by varying the weights in fixed steps of 10%. This technique worked reasonably well since the images were preprocessed to align the eyes and mouth, and normalized for brightness. Some artifacts (multiple features) can occasionally be observed in the images. A sequence thus consisted of 9 dissolve images (not including the prototypes) at 10% mix intervals for the subject JJ. Figure 1 shows examples of the transitions. We used the same images of JJ from the database as were used in Megamix for the endpoints.

In previous work [Padgett and Cottrell, 1997], we determined that extracting features from the eye and mouth regions, rather than whole-face "eigenfaces" gives the best generalization performance for emotion recognition. The features we used were the principal components of 32x32 pixel patches randomly sampled from the face images. These form a set of basis images that resemble the filtering performed by some types of cells in primary visual cortex (see Figure 2, right). Overlapping patches from the eyes and mouth were projected onto these features and the resulting scores were given as inputs to a neural network model. Two patches are used for each eye, and three for the mouth, from the regions shown in the left side of Figure 2. Each patch is projected on to the top 15 principal components of the random blocks resulting in 105 dimensional input patterns.

The training patterns consisted of 89 images of eleven subjects (five male, six female) from the PFA database (not including male subject "JJ", who is used for testing).[2] These included images of all six expressions plus eleven neutral im-

[2]The PFA database is unbalanced, in that not all subjects have all expressions represented in the database, and some have multiple occurrences of some expressions. Hence the number of images is not 77.

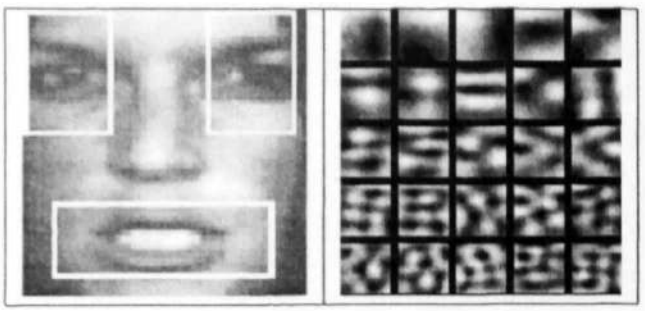

Figure 2: Left: The feature regions on a normalized test image. Right: The top 25 random block principal components, 15 of which are used as features to construct the inputs to the neural networks.

ages. In an attempt to get a balanced training set, random samples of an equal number of each emotion (10) were drawn from the 89 images and reserved for training (70 images). The remaining 19 images were used as a hold out set to stop network training. Different subsets of 70 training images were used for each network.

The network model of a "subject" consists of ensembles of 11 feed-forward, fully connected "vanilla" neural networks.[3] Each network has 105 inputs, 10 hidden units, and an output layer of 7 units, one for each emotion plus neutral. All units except the inputs are standard logistic functions. We trained the networks with back propagation and the mean-squared error cost function [Rumelhart et al., 1986]. The teaching signal was a "1" for the putative expression being portrayed, and "0" for the other six outputs. Each network in the ensemble uses different initial random weights, a different random sample of 70 training images (subject to the 10 images per emotion constraint), and thus a different hold-out set. Training was halted when the error on the hold-out set went up over three epochs. The networks took about 100 epochs to train. We trained 50 such network ensemble "subjects". All network ensembles generalized to the "JJ" images with 100% accuracy. "JJ" is a particularly good subject, in that it is easy to recognize his expressions, which is why Young et al. used his images for their human subjects.

To combine the scores of the 11 networks in the ensemble, a number of different techniques are possible: winner take all, weighted average output, voting, etc. The method that we found to consistently give the highest generalization rate is to use Z scores on a per output basis from the 11 networks. The "raw" ensemble output for emotion j is:

$$a_j = \sum_{i=1}^{11} o_{ij}$$

where o_{ij} is the output of ensemble component network i on emotion j. This is converted to a Z score:

$$z_j = \frac{a_j - \bar{a}_j}{\sigma_j}$$

where $\bar{a}_j$ and σ_j are the average and standard deviation of the raw ensemble output for emotion j over *all* training patterns. The "final" ensemble output, O_j for emotion j, is the softmax of the Z scores:

$$O_j = \frac{e^{z_j}}{e^{\sum_{i=1}^{7} z_i}}$$

[3]The number "11" is not critical here, it is used for historical reasons.

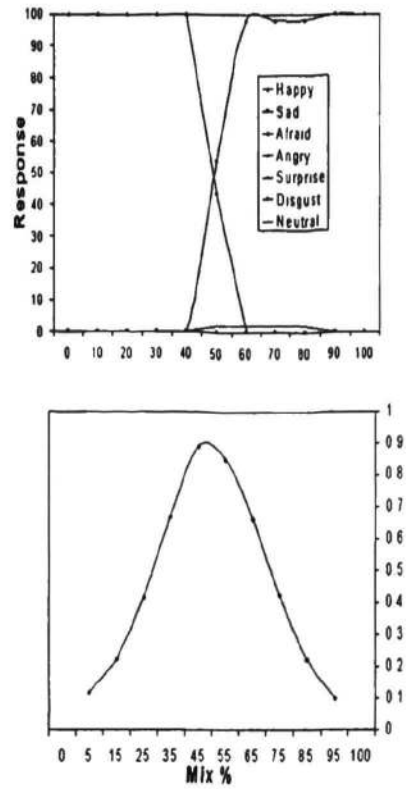

Figure 4: The top graphs show the neural network responses for two emotion transitions. The top left graph is Anger-Disgust and the top right graph shows Sad-Surprise. The corresponding bottom graphs show the associated discrimination which we model as 1 - the cosine between two consecutive output response vectors.

The output values from the ensemble networks are used to generate *responses* to a given stimulus input, corresponding to a button press in the Megamix study. The highest output value, $\max_j O_j$, for a particular input image is considered to be the emotion label of the button pressed.

We can also extract response times from our model. A standard measure of reaction time of a feed-forward neural network is to assume that it is proportional to the output error [Seidenberg and McClelland, 1989]. In our case, since there is no predetermined correct response to the dissolve imagery, we simply use the difference between the maximum output (corresponding to the network's response), and the maximum *possible* output (1.0). Thus, the more uncertain the maximum response is (the farther from 1.0), the slower the RT.

To model the discriminability between a pair of stimuli measured in Megamix Experiment 3, we suppose that each stimulus pattern is processed by the network, and the 7-dimensional output vectors (the seven O_j scores treated as a vector) are stored. This gives the overall response to the stimulus, with no decision imposed. The cosine between these two vectors gives the *similarity* of the two stimuli to the network. The more the output varies, the less similar the stimuli will be. We thus use 1 - cosine as a measure of discriminability.

Finally, to model the ranking of "closest emotions" given by the subjects in Experiment 4, we simply use their corresponding rank in the output vector.

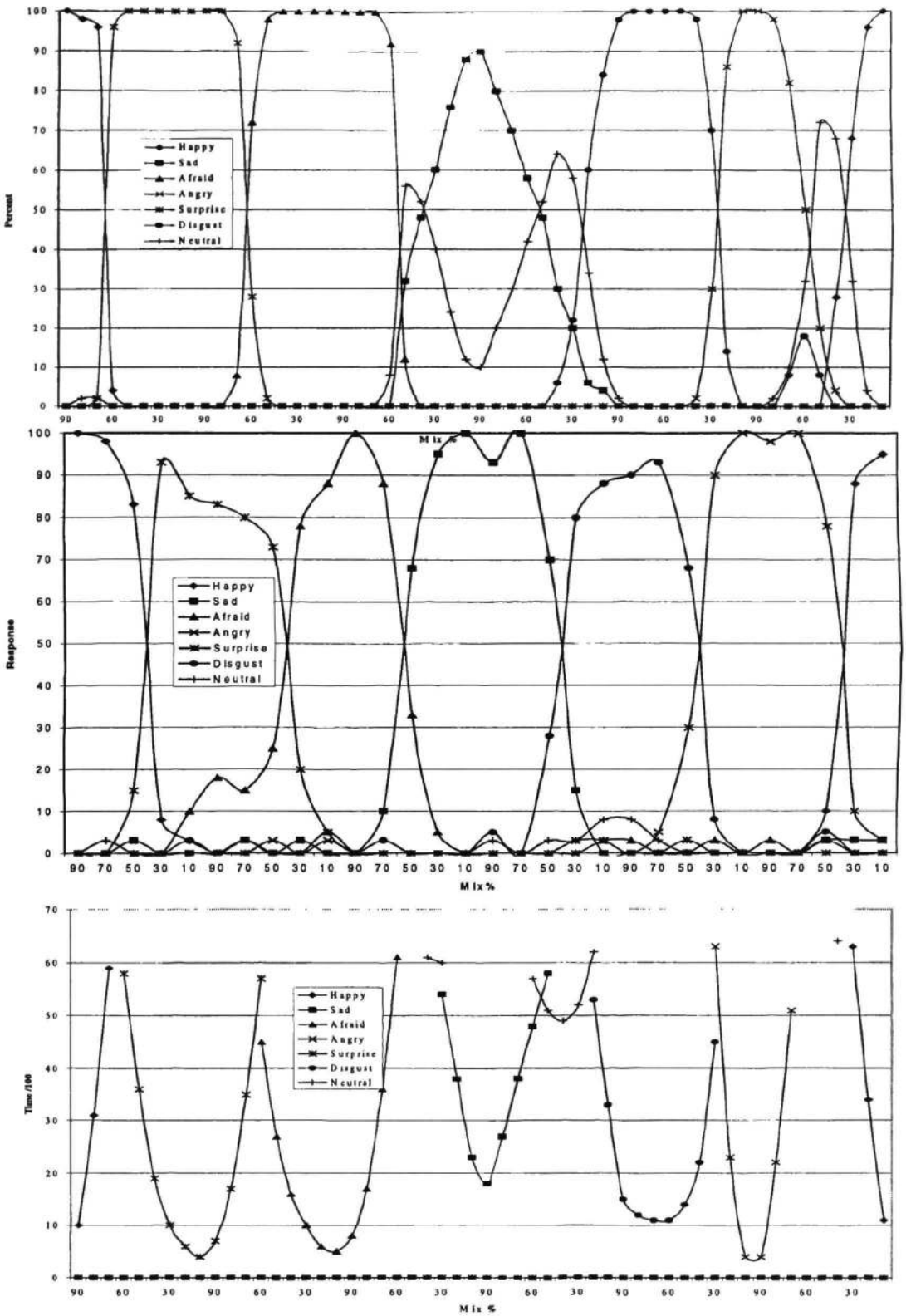

Figure 3: Example comparison of the neural network model to the human responses given in the Megamix study. The emotion sequence is as follows: Happy-Surprise-Afraid-Sad-Disgust-Angry-Happy. The top graph is the neural network models response. The middle graph shows the results from the Megamix study. The final plot gives the models response time. Each graph shows the response as the interpolation values change between the emotion prototypes.

Results

The first experiment examines the average response curves (percentage of subjects giving a particular labeling to a stimulus) as the mixture of the two emotion prototypes varies. The stimuli presented to both the neural network model and the human subjects were novel transitional faces. An example of the average responses for the 50 ensemble networks are presented at the top of Figure 3. The average response of 40 human subjects to the same sequence of emotion transitions are reproduced from the Megamix study in the middle graph [Young et al., 1997].

The most striking feature found in both the ensemble model and the subjects' responses is very sharp transition regions from emotion to emotion across the sequence. This is true for all human transitions including those not shown. For the model, the transition behavior was also sharp. However, in nearly half of the instances (7/15) of emotion-emotion transitions (not involving neutral) the neutral response is stronger than that of one of the endpoint emotions near the transition.[4] In Megamix, they did find intrusions of other emotions in 2/15 of these cases (using a binomial test), but the responses were lower than the endpoint emotions. We get similar intrusions of this type, for example, fear intrudes on surprise in the Megamix data and in our network. These neutral intrusions cannot be accounted for by the finer grain of our transitions, as the average number of images for which neutral is the highest response near a transition is 3 (i.e., these intrusions span 3 10% step sizes). More likely, the model's behavior is due to the use of dissolve faces instead of morphs. Since dissolves are pure weighted averages, and morphs are inherently nonlinear, it makes sense that some mixes may actually resemble the neutral prototype more than a morph would, as neutral is probably the average in pixel space of all emotions. We plan on rerunning the experiment with image-quality morphs (currently under construction) to eliminate this possible confound.

The final graph in Figure 3 presents the ensembles' simulated reaction time (RT) for the same emotion sequence. Young et al. found the resulting RT curves were *scalloped*, with the fastest RT's near the prototype emotion and dropping off farther from the prototype. In the lower panel, we show the network RT's for the responses for any emotion for which the model subject response curve was over 23% (the cutoff they used for their plots; unfortunately, we were unable to obtain the human data to plot here). These curves show the same scallop shape as in the Megamix paper. Figure 4 presents two examples of the ensemble models' stimulus discrimination. The top graphs show the subject response curves for two pairs of emotions and the bottom graphs depicts the discrimination score (1 - cosine of the two output vectors). The curves demonstrate that the model is most sensitive to stimulus changes near the boundary, which was also true for discrimination tests on human subjects. The model also showed that the mean discriminability score of 0.45 (0.31) for transitions (90-70,30-10) near prototypes was significantly different ($z=26.0, p<.01$) using a normal test for different means [Keeping, 1995] than the score of 0.69 (0.29), for transitions far from the prototype (70-50,50-30). This too was significant in the Megamix study [Young et al., 1997].

[4] In fact, in some sense, we are putting our worst foot forward here, in that we also modeled their Experiment 1 (data not shown) which was a six-way forced choice not including neutral. There, only three of fifteen (3/15) transitions resulted in an emotion response more prominent than either end point, and these were restricted to one image in the sequence.

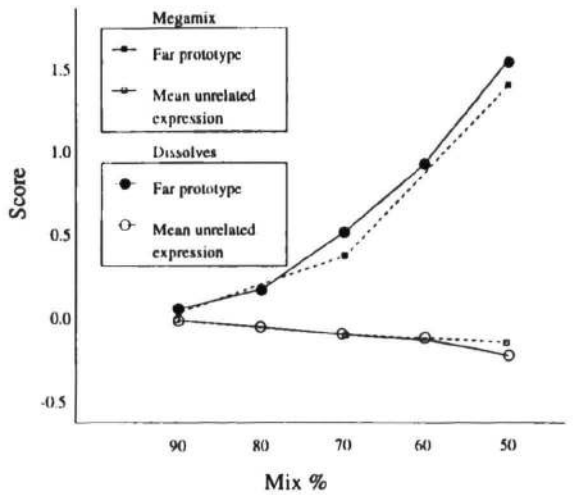

Figure 5: This graph compares the neural network model and the actual human scores from the Megamix study, computed by the same method. The plots represent the average rating subjects give to the emotion in a face as it falls further from a given anchor prototype (see text for details). The emotions are lumped into two classes, the related emotion (the one being mixed with the prototype) and the unrelated emotions. Both the neural network model and the human subjects exhibit a steep rise in prominence for the related emotion with no detectable increase for unrelated emotions.

Experiments 1-3 of Megamix strongly support categorical perception of emotion categories. In Experiment 4, Young et al. considered to what extent subjects were nevertheless sensitive to the other category being mixed into the image. For example, can the subjects perceive the anger in a 90% happy/10% angry morph, even though they respond "happy" to the image? As described earlier, they asked the subjects to give three responses to an image: which emotion it was closest to, then the next closest emotion, then the next. They scored the three responses as 3, 2, and 1, and subtracted off the average score for the dominant prototype as described earlier. These difference scores were then averaged across subjects for the "prototype being moved towards" (they call this the "far prototype"). The scores for the other four unrelated emotion categories (those not represented in the morph) were averaged together as well. These two scores were then averaged across *all* transitions, and plotted. Their data is shown as the dashed lines in Figure 5.

We used the same methodology for our networks, using the rank order of the network outputs to extract scores. The results are shown in Figure 5 as the solid lines. The unfilled circles and squares show the difference scores for the four emotions *not* represented in the dissolves or morphs. As can be seen from the Figure, the network data lies right on top of the human data in this case.

Discussion

We have shown that a feed forward neural network model using a feature based representation of the face (projections of feature regions on a fixed filter set) accounts for the observations found in the human study. Specifically, the models exhibit categorical responses: sharp transitions in the response curves and higher discrimination across category boundaries. The scallop shape in the human RT's was also modeled by the same network. In addition, the models show a very good match to the human subjects' sensitivity to the non-dominant

prototype being mixed into the images. Unlike the classical account of categorical perception, humans were able to make intra-categorical distinctions, and these results were reflected in the model as well.

The point of departure between our model and the data is the intrusion of neutral responses in our replication of their Experiment 2 (not as prevalent in our Experiment 1, see footnote4). We believe that this difference is due to the way in which we constructed our faces; simple pixel averages are more likely to be like neutral images than true morphs, which do not fall on straight lines between the endpoints in perceptual space [Busey, 1997]. We plan to verify this conjecture by applying our model to image-quality morphs in future work.

These results show that a neural network classifier, with no access to the biological constraints that are presumably imposed on the human emotion processor, and no access to the surrounding culture except to the extent that the network is instructed to carve up the input space into the same categories, can nevertheless simulate fairly well the human responses to emotional expressions.

Neural network modelers may object that, given the way we extracted the various response variables, *of course* the results would come out this way. For example, because the output vector is changing the most at category boundaries, our measure of discrimination will be highest there. In other words, it is "embarrassingly easy" to account for these results. Rather than an embarrassment, we suggest that the model is therefore a *natural* explanation of the phenomenon of categorical perception.

The reason that the neural network shows categorical perception is simple. Early in training, the network does not show steep boundaries between the classes, so the change in responses along a transition is more shallow. As learning progresses, reducing the error corresponds to sharpening the boundaries between the categories. Thus the region of ambiguity is shortened. However, different exemplars give different results. Easily identified emotions, as in the JJ images, give rise to steeper response changes than morphs between other subjects whose portrayals are not as pronounced. This is in agreement with other studies [Beale and Keil, 1995] that show familiarity with the endpoints determines the steepness of the transition in human subjects.

In future work, we intend to show that our model provides a nearly complete account of the perception-classification process in that it *learns* to classify emotions. This is of interest because recent work has shown that, in the case of identity [Beale and Keil, 1995] (but not shown so far for emotions) perception of identity is non-categorical for unfamiliar stimuli, but is categorical for familiar stimuli. This suggests that categorical perception is a phenomenon that naturally falls out of a learning process that puts increasingly sharper boundaries between stimulus categories as they become more familiar.

Acknowledgements

We would like to thank Andy Young for providing us with some of the data from the Megamix study, which is reproduced in Figures 3 and 5. We would also like to thank the members of Gary's Unbelievable Research Unit (GURU) for helpful comments on this work.

References

[Adolphs et al., 1998] Adolphs, R., Padgett, C., Logan, C., and Cottrell, G. (1998). Categorical perception of emotional facial expressions: Computer models and human performance. In Preparation.

[Anderson and Rosenfeld, 1988] Anderson, J. and Rosenfeld, E., editors (1988). *Neurocomputing: Foundations of Research*. MIT Press, Cambridge.

[Beale and Keil, 1992] Beale, J. and Keil, F. (1992). Categorical effects in the perception of faces. *Cognition*, 57:217–239.

[Beale and Keil, 1995] Beale, J. and Keil, F. (1995). Categorical perception as an acquired phenomenon: What are the implications? In Smith, L. and Hancock, P., editors, *Neural Computation and Psychology: Workshops in Computing Series*, pages 176–187, London. Springer-Verlag.

[Busey, 1997] Busey, T. (1997). Where are morphed faces in multi-dimensional face space? under review.

[Calder et al., 1996] Calder, A., Young, A., Perrett, D., Etcoff, N., and Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, 3:81–117.

[Ekman and Friesen, 1976] Ekman, P. and Friesen, W. (1976). Pictures of facial affect.

[Ekman and Friesen, 1977] Ekman, P. and Friesen, W. (1977). *Facial Action Coding System*. Consulting Psychologists, Palo Alto, CA.

[Etcoff and Magee, 1992] Etcoff, N. and Magee, J. (1992). Categorical perception of facial expressions. *Cognition*, 44:227–240.

[Harnad, 1987] Harnad, S. R. (1987). *Categorical perception: the groundwork of cognition*. Cambridge University Press, Cambridge, NY.

[Keeping, 1995] Keeping, E. S. (1995). *Introduction to Statistical Inference*. Dover Publications, New York.

[Liberman et al., 1957] Liberman, A., Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–368.

[Padgett and Cottrell, 1997] Padgett, C. and Cottrell, G. (1997). Representing face images for classifying emotions. In *Advances in Neural Information Processing Systems 9*, Cambridge, MA. MIT Press.

[Padgett et al., 1996] Padgett, C., Cottrell, G., and Adolphs, R. (1996). Categorical perception in facial emotion classification. In *Proceedings of Cognitive Science Conference*.

[Rumelhart et al., 1986] Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536. Reprinted in [Anderson and Rosenfeld, 1988].

[Russell, 1980] Russell, J. A. (1980). A circomplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

[Seidenberg and McClelland, 1989] Seidenberg, M. and McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523–568.

[Young et al., 1997] Young, A., Rowland, D., Calder, A., Etcoff, N., Seth, A., and Perrett, D. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, 63:271–313.