Categorical Perception in Facial Emotion Classification

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

Ralph Adolphs
Department of Neurology
University of Iowa, Iowa City
radolphs@blue.weeg.uiowa.edu

Abstract

We present an automated emotion recognition system that is capable of identifying six basic emotions (happy, surprise, sad, angry, fear, disgust) in novel face images. An ensemble of simple feed-forward neural networks are used to rate each of the images. The outputs of these networks are then combined to generate a score for each emotion. The networks were trained on a database of face images that human subjects consistently rated as portraying a single emotion. Such a system achieves 86% generalization on novel face images (individuals the networks were not trained on) drawn from the same database.

The neural network model exhibits categorical perception between some emotion pairs. A linear sequence of morph images is created between two expressions of an individual's face and this sequence is analyzed by the model. Sharp transitions in the output response vector occur in a single step in the sequence for some emotion pairs and not for others. We plan to us the model's response to limit and direct testing in determining if human subjects exhibit categorical perception in morph image sequences.

Introduction

In this paper, we describe a neural network model that classifies static face images based on their emotional content and examine the behavior of the model over a sequence of linearly interpolated images between two differing emotions of the same face. We are specifically looking for emotion pairs where the transition in the output response of the network is abrupt. That is, prior to the transition, the model classifies all the images in the sequence as examples of the first category and all the subsequent images as examples of the second emotion. Such transitions are known as categorical perception and are known to occur in many perceptual tasks [9]. The model's predictions can then be compared with a similar set of tasks performed on human subjects. From this interaction we hope to discern the functional organization of the visual emotion recognition system.

The neural network model consists of an ensemble of two layer, feed-forward networks trained with back propagation. The faces are represented to the network as projections of seven 32x32 pixel blocks from feature regions (both eyes and mouth) onto the principal component space generated from randomly located blocks in the image data set (see Figure 1). This technique is similar to the eigen-face/feature recognition work of Turk and Pentland [14] and Pentland et. al. [13] where projections of faces or features are used to reduce the dimensionality of the image. However, where their work uses fixed locations on the face to generate the eigen-space (resulting in a face- or feature-like appearance of the templates), we use image areas drawn randomly so that our templates are of a more non-specific nature (see Figure 2).

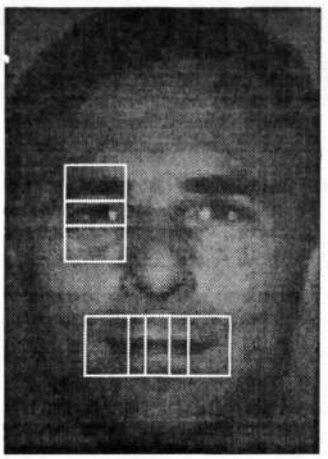

Figure 1: The figure shows the location of the 32x32 pixel blocks in the feature regions (only one eye region is shown, both are used). Each block is projected onto the top 15 eigenvectors, resulting in a 105 dimensional vector for each face.

In previous work we have shown that the generalization obtained with this representation is superior to those obtained using the eigen-face/feature strategy [12]. The expected generalization rate on novel individuals presented to the network making use of the random block representation is 86% while humans do nearly 92% on the same database. These results are comparable with emotion recognition rates obtained by other automated vision systems which require a *neutral to emotion* temporal sequences for training and evaluation [11, 15, 1].

Face Data

In working with emotions in face images, care must be taken to insure that the particular emotion being portrayed is correct. Feigned emotions by untrained individuals exhibit significant differences with the prototypical face expression [7]. These differences often result in disagreement between the observed emotion and the expression the actor is attempting to feign. In previous work, Cottrell and Metcalfe had undergraduates feign emotions. While their network performed well on identity and gender classification, it never did well on emotion. Cottrell and Metcalfe speculated that their results were due to poor portrayal of the emotions by their subjects [5].

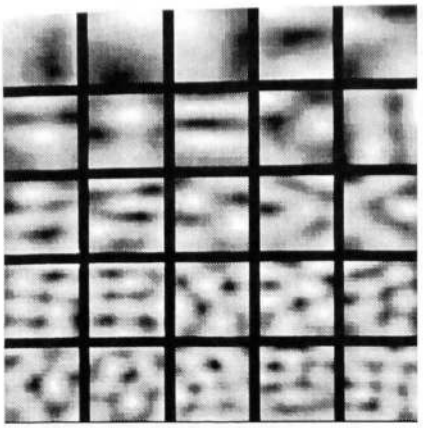

Figure 2: The top 25 eigenvectors from PCA of 32x32 pixel patches drawn randomly over the face database.

To reduce this possibility, we make use of a validated facial emotion database (Pictures of Facial Affect) assembled by Ekman and Friesen [6]. Each of the face images in this set exhibits a substantial agreement between the labeled emotion and the observed response of human subjects. The actors used in this database were trained to reliably produce emotions using FACS [7] and their images were presented to undergraduates for testing. The agreement between the emotion the actor was required to express and the students' observations was at least 70% on all the images incorporated into the database.

Twelve of the fourteen individuals contained in the Pictures of Facial Affect database were used in this study, 6 male and 6 female (the two remaining set of images were inadvertently corrupted during image capture). A total of 97 images each portrays one of 7 emotions– happy, sad, fear, anger, surprise, disgust or neutral. With the exception of the neutral faces, each image in the set is labeled with a response vector of the remaining six emotions indicating the fraction of total respondents classifying the image with a particular emotion.

Although care was taken in collecting the original images, natural variations in lighting, head size and the mouth's expression must be accounted for. The original images exhibited significant variation in the distance between the eyes (2.7 pixels) and in the vertical distance from the eyes to the mouth (5.0 pixels). To achieve scale invariance, each image was scaled so that prominent facial features were located in the same image region. Eye and mouth templates were constructed from a number of images and the most correlated template was used to localize the respective feature. Illumination variances were minimized by individually stretching each of the images to encompass the full grey scale range. Similar techniques have been employed in previous work on faces [3, 4, 14, 13]. Figure 3 shows examples of some of the normalized face images used in the study.

Classifier design and training

The models used to conduct this study consist of ensembles of feed-forward, fully connected neural networks each containing a single hidden layer with 10 nodes. Each network is trained independently using on-line back propagation with the response vectors from the Pictures of Facial Affect database

Figure 3: Examples from the Pictures of Facial Affect database normalized and cropped.

serving as the target. The input to each network is the projection of the seven locations of the faces shown in Figure 1 onto the top 15 eigenvectors of 900 random 32X32 blocks (the top 3 rows of Figure 2. The projections onto each eigenvector are normalized into Z scores on a per-eigenvalue basis.

Since we had a small dataset of twelve individuals, for each individual, we trained an ensemble of networks on the remaining 11 individuals, and then combined the scores of the ensemble members to get a generalization score on the one individual the ensemble had not seen during training. To minimize the impact of choosing a poor hold out set from the training set, each of the 11 individuals in the training set was in turn used as a hold out. If the error on the hold out set went up over three training epochs, training was stopped. This procedure is illustrated in Figure 4. Thus we end up with 12 independent ensemble network models.

To combine the scores of the 11 networks, 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 involved using Z scores from the 11 networks for each individual. The average output for each possible emotion across all the networks was calculated along with its deviation over the entire training set. These values were used to normalize the summed output of the 11 networks. The highest output Z score for a particular input was considered to be the emotion found by the ensemble. For any input pattern, we calculate the average of all 11 network outputs for emotion j:

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

where o_{ij} is the output of net_i on emotion j for that pattern. Then we convert this to a Z score:

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

where $\bar{a}_j$ and σ_j are the average and deviation for output unit j across all the pattern outputs over the entire 11 networks. The average generalization rate achieved by the classifiers is 86% (± 0.2).

Morph Sequence

A large body of literature in cognitive psychology has demonstrated that certain stimuli, such as phonemes, are perceived categorically by human subjects [10, 9]. Categorical perception is said to occur when stimuli can be discriminated no better than they can be labeled, although in practice somewhat more relaxed criteria are often taken as evidence of categorical

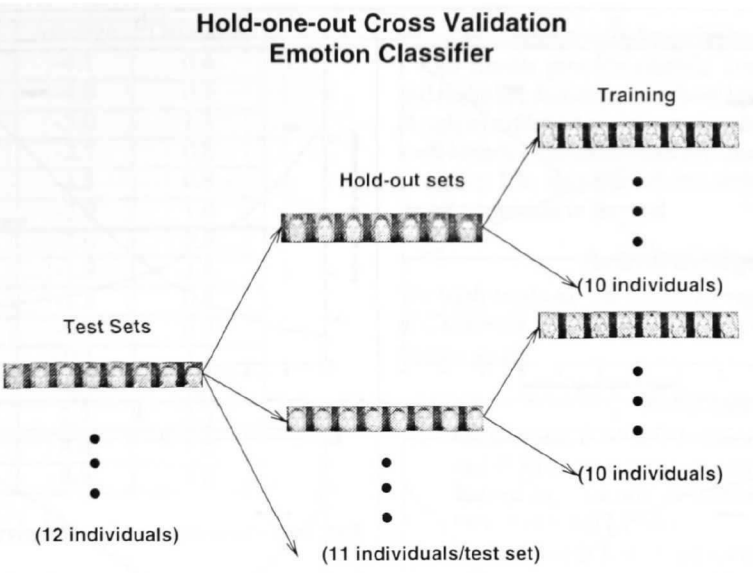

Figure 4: Structure of training sets used for creating ensemble network.

Figure 5: Typical examples of the morphing process. The original expressions are at the extremes of the sequences while the interior images are linear interpolations of the two images.

perception. Although the evidence for certain "low level" categorical perception is strong (e.g. phonemes, colors), much less is known about how we categorize more complex stimuli. Two recent studies suggest that some of the information signaled by faces, notably their emotional expression [8] and their unique identity [2], is perceived categorically.

The present study concentrates on the perception of emotional facial expressions. We are aware of only a single study that has provided evidence for categorical perception of emotion in facial expressions [8]. That study used line drawings of faces, and morphs of those line drawings, as the stimuli. Transitions between certain emotional expressions appeared to be perceived categorically, while other transitions did not show such an effect. Given these intriguing findings obtained with line drawings of faces, we wanted to approach the issue using images of actual facial expressions of emotion.

To determine the type of transitions that the neural network model exhibits, we linearly transform a face image of an individual expressing one emotion to the same individual expressing another at fixed intervals. The resultant morphed image sequence can then be transformed into the input representation and presented to the classifier in a normal manner. Figure 5 shows typical image sequences generated by this process.

For the network model, three distinct types of response vectors are generated over the course of the transition sequences. At either end of the morph transition sequence, the network model responds correctly to the original image. When the maximum output response changes from the first to the second emotion, this is termed the *crossover point*.

A *Type 1 transition* is where the network response at the crossover point is high with respect to a threshold response – that is, *both* emotions are above threshold. We set the threshold at 0.5 standard deviations above the average response, which is the maximum value that maintains the 86% generalization rate of the network (in the original work, a correct response was taken as the maximum Z score [12]). For Type 1 transitions, both emotions elicit high responses over a large portion of the sequence indicating similarity between the two emotions in the transition sequence.

A *Type 2 transition* is when both emotion responses are below threshold at the crossover point. This type of transition has a large portion of the morph sequence without any prominent emotions. This indicates that the categories are quite

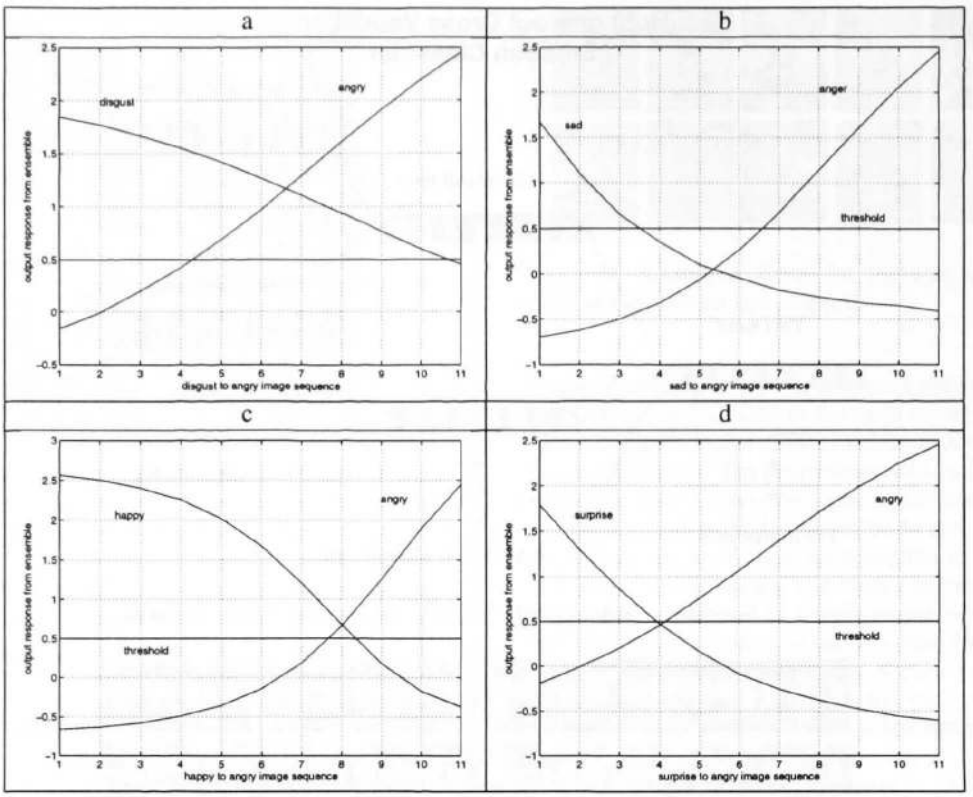

Figure 6: The graphs show the output response of the two emotions of the neural network model for the morph image sequences. Graphs **a** and **b** are examples of Type 1 and 2 transitions respectively, while **c** and **d** illustrate Type 3.

distant in the model space.

A *Type 3 transition* occurs when the crossover point is near the threshold. This indicates a sharp transition in classification of the morph images so that all (or most) images prior to the crossover are classified the same as the original image and those subsequent to it, the same as the morphed–to image. It is this type of transition we associate with categorical perception. Figure 6 presents examples of the output responses of the associated emotions that illustrate the three types of transitions.

Results

To examine the type of transitions between emotion pairs, morph sequences between differing emotions in each of the 12 ensemble network's test set were generated and presented to the appropriate ensemble network for evaluation. A morph sequence of 9 images (plus the 2 originals) was constructed for each distinct emotion pair of the individual (fear to fear morphs for example, were not examined). The total number of morph sequences was 250, approximately 17 sequences for each of the 15 possible emotion combinations.

A simple score was used to evaluate the type of the transition for each of the morph sequences. As the change in the ensemble network's output across the sequence was either monotonically increasing or decreasing, simply counting the number of outputs greater than the threshold of +0.5 standard deviations is sufficient in determining the type of the transition. If we simply add the number of instances where emotion 1 is above the threshold to the number of instances emotion 2 is above the threshold and subtract the length of the sequence from the total, the sign and magnitude of the resultant value provides a simple score indicating the transition type. High positive scores indicate that the ensemble network output was responsive to both emotions over a large number of images (Type 1 transition) while large negative scores indicate that the response vectors of both emotions were below the threshold for a number of images (Type 2 transition). Scores near zero are the Type 3 transitions that indicate categorical perception. Figure 7 presents the experimental results from lowest to highest scores.

The ensemble networks predictions about the type of relationships between the emotion categories that exist in the face image data is presented in Figure 8. The entries in the table are arranged so that the emotion pairs at the top of the columns have the scores most representative of that type. A cut off of ± 1.5 steps was used to delimit the range of values associated with Type 3 behavior (i.e. the transition area scores were within 1.5 steps of 0).

The ensemble models' predictions seem reasonable given the nature of the categories. For instance, happy faces, the only positive emotion examined, do not seem particularly close to any of the other five emotions. Happy has no positive scores and the transitions between it and the other emotions consist solely of Types 2 and 3. Type 1 transitions are exhibited by the emotion pairs that are neighbors when the human data from the Pictures of Facial Affects database is subjected

Emotion Pair	Average	Deviation
Fear-Disgust	-4.1	0.4
Surprise-Disgust	-4.0	0.3
Happy-Sad	-3.0	1.3
Fear-Anger	-2.7	0.8
Happy-Anger	-2.5	0.9
Sad-Fear	-2.0	1.0
Happy-Surprise	-2.0	0.5
Sad-Surprise	-1.9	1.8
Happy-Fear	-1.3	0.8
Sad-Anger	-0.4	0.9
Anger-Surprise	-0.1	1.3
Happy-Disgust	-0.1	1.0
Sad-Disgust	2.7	2.3
Anger-Disgust	4.4	1.6
Fear-Surprise	8.9	1.0

Figure 7: Averages and deviations of emotion pair scores of the neural network model.

Type 1	Type 2	Type 3
Fear-Surpr	Fear-Disg	Happy-Disg
Anger-Disg	Surprise-Disg	Anger-Surpr
Sad-Disg	Happy-Sad	Sad-Anger
	Fear-Anger	Happy-Fear
	Happy-Anger	
	Sad-Fear	
	Happy-Surpr	
	Sad-Surpr	

Figure 8: The ensemble network's predictions of the type classifications for the various emotion pairs.

to Multi-Dimensional Scaling (MDS) (data not shown). Type 2 transitions are the most common, indicating significant separation between most emotion pairs (again similar to the circular arrangement found using MDS). Finally, the Type 3 transitions are from emotion pairs that are quite opposite.

Conclusion

We have demonstrated that a relatively simple neural network model is able to recognize emotions and make predictions about how visual categories are related to one another. We intend to use these results to guide us in testing human subjects in order to determine if categorical perception occurs in morph image sequences of actual images. An exhaustive search of all possible morph combinations is difficult (time consuming and costly) when testing humans but the models' predictions should be able to significantly reduce the number and type of combinations tested. We have begun human tests to validate the predictions on the Type 1 and Type 3 transitions. Subjects are randomly shown each image 10 times and make a forced choice between the endpoint emotions. Although our n of 2 is not large enough to be reliable, both subjects showed very sharp (sigmoidal) transitions for Type 3, and very linear trends for Type 1, as predicted by the model. If a comparison of the model and human data is favorable, we can begin to use this type of model to investigate the performance changes encountered in emotion recognition (and other static recognition tasks) for patients with brain lesions.

Our results provide specific avenues for further research, and make predictions about how human subjects may perceive blends of different emotions signaled by a face. Since it is just such blends of emotion that are most typically encountered in everyday life, this line of research will contribute to human social cognition in general.

Acknowledgements

We wish to thank the GEURU research group at University of California, San Diego and reviewers for helpful comments on this work.

References

[1] M. Bartlett, P. Viola, T. Sejnowski, J. Larsen, J. Hager, and P. Ekman. Classifying facial action. In *Advances in Neural Information Processing Systems 8*, Cambridge, MA, 1996. MIT Press.

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

[3] D. Beymer. Face recognition under varying pose. Technical Report AI Memo No. 1461, MIT Artificial Intelligence Lab, 1993.

[4] R. Brunelli and T. Poggio. Face recognition: Feature versus templates. *IEEE Trans. Patt. Anal. Machine Intell.*, 15(10), October 1993.

[5] Garrison W. Cottrell and Janet Metcalfe. Empath: Face, gender and emotion recognition using holons. In R.P. Lippman, J. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 564–571, San Mateo, 1991. Morgan Kaufmann.

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

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

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

[9] Stevan R. Harnad. *Categorical perception: the groundwork of cognition*. Cambridge University Press, Cambridge, NY, 1987.

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

[11] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, 74(10):3474–3483, 1991.

[12] C. Padgett and G. Cottrell. Identifying emotion in static face images. In *Proceedings of the 2nd Joint Symposium on Neural Computation*, volume 5, pages 91–101, La Jolla, CA, 1995. University of California, San Diego.

[13] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*, 1994.

[14] Matthew Turk and Alexander Pentland. Eigenfaces for recognition. *The Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[15] Yacoob and Davis. Recognizing human facial expression. Technical Report CAR-TR-706, University of Maryland Center for Automation Research, 1994.