A Six-Unit Network is All You Need to Discover Happiness

Matthew N. Dailey **Garrison W. Cottrell**
{mdailey,gary}@cs.ucsd.edu
UCSD Computer Science and Engineering
9500 Gilman Dr., La Jolla, CA 92093-0114 USA

Ralph Adolphs
ralph-adolphs@uiowa.edu
University of Iowa Department of Neurology
220 Hawkins Dr., Iowa City, IA 52242 USA

Abstract

In this paper, we build upon previous results to show that our facial expression recognition system, an extremely simple neural network containing six units, trained by backpropagation, is a surprisingly good computational model that obtains a *natural* fit to human data from experiments that utilize a forced-choice classification paradigm. The model begins by computing a biologically plausible representation of its input, which is a static image of an actor portraying a prototypical expression of either Happiness, Sadness, Fear, Anger, Surprise, Disgust, or Neutrality. This representation of the input is fed to a single-layer neural network containing six units, one for each non-neutral facial expression. Once trained, the network's response to face stimuli can be subjected to a variety of "cognitive" measures and compared to human performance in analogous tasks. In some cases, the fit is even better than one might expect from an impoverished network that has no knowledge of culture or social interaction. The results provide insights into some of the perceptual mechanisms that may underlie human social behavior, and we suggest that the system is a good model for one of the ways in which the brain utilizes information in the early visual system to help guide high-level decisions.

Introduction

In this paper, we report on recent progress in understanding human facial expression perception via computational modeling. Our research has resulted in a facial expression recognition system that is capable of discriminating prototypical displays of Happiness, Sadness, Fear, Anger, Surprise, and Disgust at roughly the level of an untrained human. We propose that the system provides a good model of the perceptual mechanisms and decision making processes involved in a human's ability to perform forced-choice identification of the same facial expressions. The present series of experiments provides significant evidence for this claim.

One of the ongoing debates in the psychological literature on emotion centers on the structure of emotion space. On one view, there is a set of discrete basic emotions that are fundamentally different in terms of physiology, means of appraisal, typical behavioral response, etc. (Ekman, 1999). Facial expressions, according to this categorical view, are universal signals of these basic emotions. Another prominent view is that emotion concepts are best thought of as prototypes in a continuous, low-dimensional space of possible emotional states, and that facial expressions are mere clues that allow an observer to locate an approximate region in this space (e.g. Russell, 1980; Carroll and Russell, 1996).

One type of evidence sometimes taken as support for categorical theories of emotion involves experiments that show "categorical perception" of facial expressions (Etcoff and Magee, 1992; Young et al., 1997). Categorical perception is a discontinuity characterized by sharp perceptual category boundaries and better discrimination near those boundaries, as in the bands of color in a rainbow. But as research in the classification literature has shown (e.g. Ellison and Massaro, 1997), seemingly categorical effects naturally arise when an observer is asked to employ a decision criterion based on continuous information. Neural networks also possess this dual nature; many networks trained at classification tasks map continuous input features into a continuous output space, but when we apply a decision criterion (such as "choose the biggest output") we may obtain the *appearance* of sharp category boundaries and high discrimination near those boundaries, as in categorical perception.

Our model, which combines a biologically plausible input representation with a simple form of categorization (a six-unit softmax neural network), is able to account for several types of data from human forced-choice expression recognition experiments. Though we would not actually propose a localist representation of the facial expression category decision (we of course imagine a more distributed representation), the evidence leads us to propose 1) that the model's input representation bears a close relationship to the representation employed by the human visual system for the expression recognition task, and 2) that a dual continuous/categorical model, in which a continuous representation of facial expressions coexists with a discrete decision process (either of which could be tapped by appropriate tasks), may be a more appropriate way to frame human facial expression recognition than either a strictly categorical or strictly continuous model.

The Expression Classification Model

For an overview of our computational model, refer to Figure 1. The system takes a grayscale image as input, computes responses to a lattice of localized, oriented spatial filters (Gabor filters) and reduces the resulting high dimensional input by unsupervised dimensionality reduction (Principal Components Analysis). The resulting low-dimensional representation is then fed to a single-layer neural network with six softmax units (whose sum is constrained to be 1.0), each corresponding to one expression category. We now describe each of the components of the model in more detail.

The Training Set: Pictures of Facial Affect

The model's training set is Ekman and Friesen's Pictures of Facial Affect (POFA, 1976). This database is a good

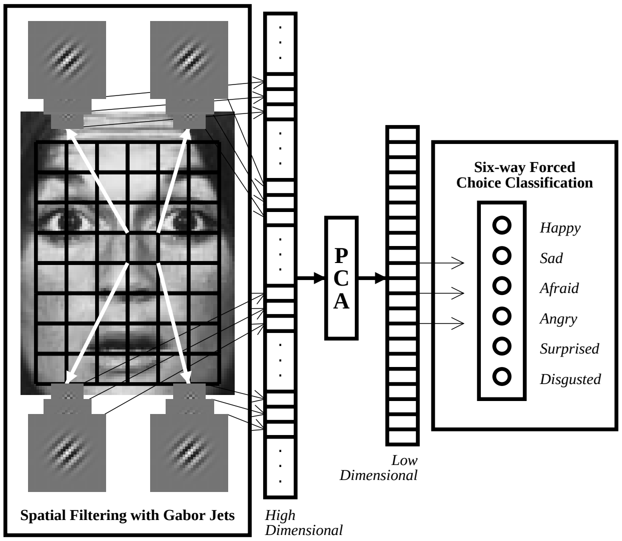


Figure 1: Facial Expression Classification Model.

training set because the face images are reliably identified as expressing the given emotion by human subjects (at least 70% agreement), and the images are commonly used in psychological experiments. We digitized the 110 POFA slides by scanning them at 520x800 pixels, performing a histogram equalization, aligning the eyes and mouths to the same location in every image by a linear transformation, and cropping off most of the background. The result is a set of 110 240x320 grayscale images of 14 actors portraying prototypical expressions of six basic emotions and neutral.

Feature Extraction: The Gabor Jet Lattice

The system represents input stimuli using a lattice of responses of 2-D Gabor wavelet filters (Daugman, 1985). The Gabor filter, essentially a sinusoidal grating localized by a Gaussian envelope, is a good model of simple cell receptive fields in cat striate cortex (Jones and Palmer, 1987). It provides an excellent basis for recognition of facial identity (Wiskott et al., 1997), individual facial actions (Donato et al., 1999), and facial expressions (Dailey and Cottrell, 1999; Lyons et al., 1999). We use phase-invariant Gabor magnitudes with a parameterization of the filter at five scales ranging from 16–96 pixels in width and eight orientations ranging from 0 to $\frac{7\pi}{8}$ as described by Donato et al. (1999). Thus, at each point in the lattice (in our representation a 29×36 grid of filter locations placed at regular 8-pixel intervals over the face), we extract a 40-element vector of Gabor magnitudes (sometimes called a "jet") that characterizes a localized region of the face. A few of the filters are displayed graphically in Figure 1. To extract the $29 \times 36 \times 40 = 41,760$ filter responses, we first convolve the entire image with each filter and take the magnitude of each complex valued response. We then (globally) divisively normalize the vector of responses at each filter scale to unit length. By equalizing the contribution of each filter size to the final representation, we overcome the problem that most of an image's power lies in lower spatial frequency ranges, without destroying information possibly present in the relative magnitude of response at each orientation. Since even the smallest filters in our representation overlap with their neighbors, and Gabor magnitudes are mildly invariant to slight translation, we lose very little of the information in the higher spatial frequency ranges, with a small price paid (due to ignoring phase information) in loss of precise feature localization and a larger price paid in that the resulting representation is very high dimensional (41,760 elements).

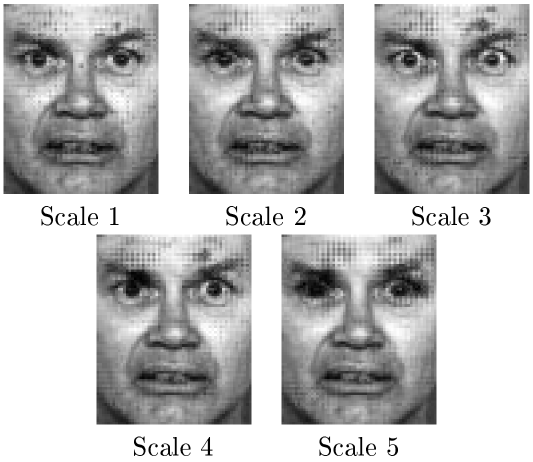
Evaluation of the representation In this section, we examine the representation's utility and plausibility.

Donato et al. (1999) found that a nearest neighbor classifier with a cosine similarity metric applied directly to a Gabor grid-based representation achieved 95.5% correct classification of image *sequences* containing *individual facial actions* (Ekman and Friesen, 1978), e.g. facial action 1, the inner brow raiser. We evaluated this type of classifier on our task, classification of full-face expressions in static images. Nearest neighbor classification of the 96 expressive faces in POFA using leave-one-actor-out cross validation and a cosine similarity metric achieves an expected generalization accuracy of 74.0%. There are several possible reasons for this sub-par performance: the need to simultaneously integrate information from multiple facial actions, the small size of the POFA database, and/or the lack of information on the dynamics of facial movement. But the simple system's performance is well above chance (16.7% correct), giving an indication that a more complicated (and more psychologically plausible) model such as a neural network could do much better.

One way of visualizing the effectiveness of a representation, and gaining insight into how an agent might use the representation to support decision-making, is to apply discriminant analysis.[1] For the Gabor magnitude components at a given location and spatial frequency, we find Fisher's Linear Discriminant (Bishop, 1995), the projection axis $\vec{w}$ that maximizes the criterion $J(\vec{w})$, the ratio of between-class to within-class scatter along $\vec{w}$. $J(\vec{w})$ is a measure (invariant to linear transformations) of the diagnosticity of that portion of the representation for determining the class of the stimulus. That is, we can determine exactly how well (in the linear sense) the representation separates individual facial expressions.

We applied this method to the 85 expressive faces of a 12-actor subset of the POFA database The results for Fear, the most difficult to recognize expression in POFA (for both humans and machines), are shown in Figure 2. The size of the dots placed over each grid location in the face is proportional to how easy it is to separate Fear from all of the other expressions based on the 8 Gabor filter responses extracted at that position of the grid. There are two interesting aspects to the result. First, the lowest spatial frequency channel (using filters about

[1] We introduced this visualization method for the Gabor representation in a recent technical report (Dailey and Cottrell, 1999), and Lyons et al. (1999) have independently introduced a similar technique.



Scale 1 Scale 2 Scale 3

Scale 4 Scale 5

Figure 2: Diagnosticity of Gabor filter locations for Fear discrimination, separated by filter spatial frequency, from scale 1 (highest SF) to scale 5 (lowest).

96 pixels in width, compared to the total image width of 240) is best for this expression, implying that improvement might be obtained by dropping the smaller scales from the representation and even increasing the filter size. Second, the technique hints at which facial actions are most reliable for distinguishing expressions from one another, readily making predictions for psychological experiments. According to Ekman and Friesen (1978), prototypical displays of Fear include facial action 1 (inner brow raise), 2 (outer brow raise), 4 (scrunching together of the eyebrows), and 5 (upper eyelid raise) in the upper face, along with 25 (lips part) and some combination of 20 (lip stretch), 26 (jaw drop), or 27 (mouth stretch) in the lower face. Although some discriminability can be obtained in the higher spatial frequencies in the region of the mouth (presumably detecting facial action 25), our model finds that the best regions are in the lower spatial frequencies around the eyes, especially around the upper eyelids.

Principal Components Analysis for Dimensionality Reduction

We use Principal Components Analysis (PCA) as a simple, unsupervised, linear method to reduce the dimensionality of the network's input patterns by projecting each 41,760-element pattern onto the top k eigenvectors of the training set's covariance matrix. This speeds up classifier training and improves generalization. We experimented with various values of k and achieved the best generalization results with $k = 35$, so in all experiments reported here we project training and test patterns onto the top 35 principal component eigenvectors of the training set, then use the standard technique of "z-scoring" each input to a mean of 0 and a standard deviation of 1.0 (Bishop, 1995).

Classification by a Six Unit Network

The classification portion of the model is a six-unit neural network. Each unit in the network first computes its net input, a weighted sum of the input patter $\vec{x}$: $a_i = b_i + \sum_j w_{ij}x_j$. Then the softmax function $y_i = e^{a_i} / \sum_k e^{a_k}$ is applied to the net inputs to produce a 6-element output vector $\vec{y}$. The network is trained with the relative entropy error function (Bishop, 1995). Since the outputs of this network must sum to 1.0, we use a constant target vector of $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})^T$ for the neutral training stimuli.

With no hidden layer and just 35 elements in its input, the network is very small, but its number of parameters, 216, is still large compared to the number of training examples (88-99). Therefore, we must avoid overtraining the network; we have found that too-fast optimization techniques lead to poor generalization. We have obtained the best results using stochastic gradient, momentum, weight decay, and early stopping using a holdout set. For the experiments reported here, we used a learning rate $\eta = 0.0017$ (the number of units divided by the number of inputs times 0.01), a momentum $\alpha = 0.9$, and weight decay rate $\nu = 0.01$.

The early stopping technique bears some explanation. We obtain expected generalization results by leave-one-actor-out cross validation. For POFA, this means a network is trained on the images of 13 actors and tested on generalization to the 14th. Rather than training on the full 13 actors, we leave one out as a holdout set to help determine when to stop training. After each epoch of training on the remaining 12 actors' faces, we test the network's performance on the 13th actor (the holdout set). If classification accuracy on the holdout set has not improved in 6 epochs, we stop training and restore the weights from the best epoch. Training time under this paradigm varies greatly; it ranges anywhere from 60 to 300 epochs depending on which partition into training, holdout, and test set is used.

Evaluation of the Network's Performance

How does the network perform the expression recognition task? An examination of the trained network's representation provides some insight. The idea is to project each unit's weight vector back into image space in order to visualize what the network is sensitive to in an image. But this is not a trivial task; though PCA is linear and easily inverted, the Gabor magnitude representation, besides being subsampled, throws away important phase information. Normalization of the power in each spatial frequency channel could also be problematic for inversion. Current techniques for inverting Gabor magnitude representations (C. von der Malsburg, personal communication) are computationally intensive and make several assumptions that do not apply here. So we instead take a simpler approach: learning the function from the 35-element input space into facial image space with linear regression, then using the regression formula to produce an image that visualizes each network unit's weight vector.

The results for one network trained on an arbitrary 12-actor subset of POFA are shown in Figure 3. In each image, each pixel value is the result of applying the regression formula predicting the value of the pixel at that

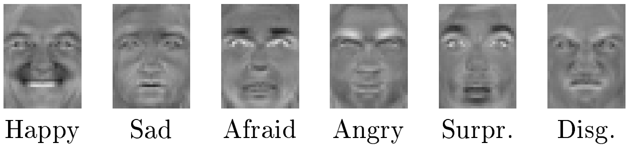


Figure 3: Images reconstructed by linear regression from a trained network's weight vectors.

location as a linear function of the 35-element weight vector for the given network output unit. Dark and bright spots indicate the features that excite or inhibit a given output unit depending on the relative gray values in the region of that feature. Note that the representations are very much like one might predict given the linear discriminant analysis described earlier: each unit combines evidence based upon the presence or absence of a few local features; for Fear, the salient criteria appear to be the eyebrow raise and the eyelid raise, with a smaller contribution of parted lips.

An important factor not shown in Figure 3 is the effect output units have on each other. Due to the divisive normalization of the softmax function, an active output unit can effectively inhibit other units that are only mildly activated. Nevertheless, it seems clear from the reconstructions that the network's effective strategy is to learn how the combination of facial actions involved in each prototypical expression can be reliably detected in a static image. We hypothesize that, when faced with a forced choice expression recognition task, humans must use similar representations and classification strategies. In the next two sections, we provide some indirect support for this hypothesis with both qualitative and quantitative comparisons between the model's performance and human performance on the same stimuli.

Modeling Forced-Choice Classification

Ekman and Friesen (1976) presented subjects with the task of 6-way forced choice classification of the expressive stimuli in POFA and provide the results of their experiment with the dataset. Their criterion for admission into the final database was that at least 70% of subjects should agree on each face's classification into one of the six POFA expression categories. On average, the proportion of agreement (or chance of correct classification) was 91.7%.

Classification accuracy comparison

We trained $14 \times 13 = 182$ networks, one for each of the possible partitions of the database into a training set of 12 actors, a holdout set of one actor, and a test set of one actor. After training using the method described earlier, we tested each network's classification accuracy on its generalization (test) set and averaged their performance. The 182 networks, on average, obtain a classification accuracy of 85.9% (compared to a human accuracy of 91.7%), and interestingly, the rank order of expression category difficulty, Happy – Disgusted – Surprised – Sad – Angry – Afraid, is *identical to that of the humans*. We also find that the humans and networks show the same rank order We have also found that it is possible to boost classifier accuracy on this task if the classifier is given the opportunity to "peek" at the test set (without labels) before actually classifying it. This "batch mode" classification technique is a plausible model for familiarizing subjects with the stimuli in an experiment prior to testing them. It boosts classifier accuracy to up to 95%; details are available in a technical report (Dailey and Cottrell, 1999).

Visualization with Multidimensional Scaling

Multidimensional Scaling (MDS) is a frequently-used technique for visualizing relationships in high-dimensional data. It aims to embed stimuli in a low dimensional space (usually two or three dimensions) while preserving, as best possible, observed distances or similarities between each pair of stimuli. MDS has long been used as a tool for exploring the psychological structure of emotion. Russell has proposed a "circumplex" model of affect (Russell, 1980) that describes the range of human affective states along two axes, pleasure and arousal. Russell and colleagues have found support for their theory in a wide range of studies for which MDS consistently yields two-dimensional solutions whose axes resemble pleasure and arousal.

A similar technique can be applied to Ekman and Friesen's forced-choice data. We computed a 96×96 Euclidean distance matrix from the 6-dimensional response vectors supplied by Ekman and Friesen and used non-metric MDS[2] to find a 2-dimensional configuration of the 96 stimuli. This configuration, shown in the first graph of Figure 4, yielded a Kruskal stress $S = 0.205$. The circumplex embedded in Ekman and Friesen's data, Happiness – Surprise – Fear – Sadness – Anger – Disgust, or HSFMAD (using M for Maudlin in place of Sadness to distinguish it from Surprise), is different from that typically reported by Russell and colleagues. This is not surprising, however, because a large portion of Russell's circumplex (affective states that are negative on the arousal dimension and positive or neutral on the pleasure dimension, such as sleepiness, content, and relaxation) is simply not represented in POFA. The HSFMAD circumplex *is* the same, however, reported by Katsikitis (1997), who used the same set of expressions, a similar forced-choice arrangement, but an entirely different set of photographs in which the actors were not instructed on how to portray each expression.

Does the facial expression similarity structure induced by the network resemble the human psychological similarity structure in any way? We have performed MDS analyses at three levels in this network: at the input layer (on the Gabor/PCA representation), at the net inputs to the network's output units (the units' un-softmaxed activations a_i), and at the softmax output layer. As one might expect, at the input layer, the patterns form

[2] There are many varieties of MDS; we implemented the Guttman-Lingoes SSA-1 algorithm as described in Borg and Lingoes (1987). Put briefly, the algorithm iteratively derives a configuration $\mathbf{X}$ that minimizes Kruskal's stress S, which is the proportion of variance in a monotonic regression unexplained by $\mathbf{X}$.

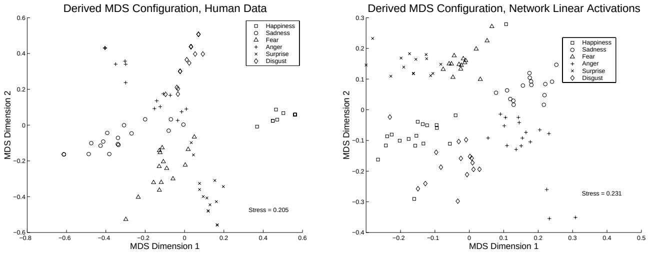


Figure 4: MDS configurations derived from human classification data and the linear activations of the units in the network model. The circumplex (order of stimuli around the graph) is the same: H-S-F-M-A-D (M=Maudlin/Sadness).

a cloud in the plane with little structure. At the network's output, the responses on the training set tend to be so nearly binary that there is very little similarity structure. But using the net inputs to the softmax units, averaged over all 182 networks, we obtain a solution (stress = 0.231) that orders the expressions in the same way as the human circumplex, as shown in the second graph of Figure 4.

With the caveat that this only occurs in the linear part of the network, the fact that the human and network MDS solutions contain the same ordering is striking. It is very unlikely ($p = 0.017$ for a single trial and $p = 0.033$ for two trials) that we would obtain the same ordering if the human and network similarity structure were in fact unrelated.

Correlation of network and human errors

MDS analysis is useful as a visualization tool, but the correspondence between the human circumplex and network circumplex is not a formal test of the model. Is the correspondence between the human and network MDS solutions simply a fortuitous coincidence? One way to address this concern is with a direct comparison of the confusion matrices for the humans and networks. For the humans and networks, we computed the 6×6 confusion matrix whose ij-th entry gives the probability that when a face from class i is present, the humans or networks (on the training set) respond with expression j. Since the network was explicitly trained to produce label i for members of class i, we removed the diagonal elements from each confusion matrix and compared the network and human *error patterns*, i.e. the 30 off-diagonal terms of the confusion matrices. Note that it is not "cheating" to use the network's responses on the training set here; the network was never biased in any way to make errors similar to humans. We found that the correlation between the off-diagonal elements of the confusion matrices for the humans and networks is $r = 0.567$. An F-test ($F(1, 28) = 13.3; p = 0.0011$) confirms the significance of this result. These results lead us to claim that much of the facial expression similarity structure observable in forced-choice experiments is due to direct perceptual similarity, and that our model does an excellent job of capturing that structure.

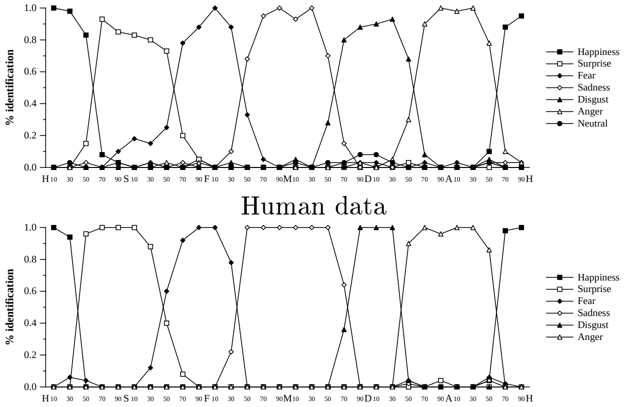
Modeling Perception of Morphs

Beyond the forced-choice classification data provided by Ekman and Friesen, the literature on categorical perception of facial expressions transitions is a treasure trove of data for modeling. Previous work (Padgett and Cottrell, 1998) compared a somewhat different facial expression recognition model to human behavior in a large study by Young et al. (1997) (henceforth referred to as "Megamix"). In the Megamix study, the researchers created morph stimuli interpolating each of the 21 possible transitions between six expressive images and one neutral image of POFA actor "JJ." They then tested subjects on forced-choice identification of the perceived expression in the morphs (they also measured response times, discrimination, and the subjects' ability to detect mixed-in expressions in the morph stimuli). Padgett and Cottrell (1998) simulated the Megamix morph stimuli with *dissolves*, or linear combinations of each source image and target image. Their linear feature extraction technique (projection of eye and mouth regions onto a Local PCA basis) and neural network classifier applied to the linear dissolves produced good results. However, when we created true morphs and attempted to apply the same techniques, we found that the model no longer fit the human data — there were large intrusions of unrelated expressions along the morph transitions, indicating that linear feature extraction is unable to produce a smooth response to nonlinear changes in the image. One might expect that the Gabor magnitude representation, with its built-in invariance to phase, might better capture the smooth, categorical transitions observed in the Megamix study on nonlinear morphs. In this section, we very briefly show that this is indeed the case: the Gabor/PCA-based model does produce smooth transitions between expression categories without intrusions and a very good fit to the human identification data without any free parameters.

Network training

We used a slightly different methodology for modeling this data because we wanted to model each human subject with one trained network. This requires as much between-subject variability as possible (although variability is difficult to achieve given POFA's small size). We trained 50 networks on different random partitions of the 13 non-JJ actors' images into training and holdout sets. Each network's training set consisted of 7 examples of each expression plus neutrality, with the remaining data used as a holdout set. As before, neutral stimuli were assigned the uniform target vector $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]^T$ and the expressive faces were assigned binary target vectors.

After training each network until holdout set classification error was minimized, we tested its performance on JJ's prototypes as well as all morphs between them. We then extracted identification, response time, discrimination, and faint morph detection response variables from the model.



Human data

Model predictions

Figure 5: Human and network responses to JJ morphs along the transitions HSFMDA.

Model fit

Using the same response variable measurements as Padgett and Cottrell (1998), we do find the Megamix pattern of sharp categorical transitions, scallop-shaped response time curves, improved discrimination near category boundaries, and a close correspondence between humans and networks on detection of the secondary expression in morph transitions. Due to space limitations, we cannot report all of the Megamix modeling results here, but we do show the model's fit to the human responses on one series of morph transitions. Forced-choice identification results for the Happy – Surprised – Afraid – Sad – Disgusted – Angry – Happy transition series are shown in Figure 5. The human data and model prediction are quite similar, but the networks appear to place slightly sharper boundaries between expressions; this is because there is not as much variation in our population of network "subjects" as that occurring in the Megamix data. Nevertheless, the correspondence ($r^2 = 0.846$) is remarkable considering that the networks were never trained on images of JJ or morph stimuli and that there are absolutely no free parameters involved in fitting the model to the data.

Discussion

We have shown that a simple, mechanistic computational model obtains a natural fit to data from several psychological studies on classification of human facial expressions. Exploring the space of possible expression classification models has led us to reject several alternative models (including local PCA-based input representations and more complicated ensembles of networks containing hidden layers). Since one simple model, despite its lack of culture and social experience, explains so much data without any free parameter fitting, we claim that it is a strong model for how the human visual system perceives facial expressions in static images. To the extent that performance in the controlled forced-choice psychological experiments cited here generalizes to more naturalistic social situations (an admittedly big assumption to make), we suggest that the model captures the essentials of the visual processing used to make many social judgments.

Acknowledgments

We thank Gary's Unbelievable Research Unit (GURU) for valuable comments on this research, Curtis Padgett for laying the foundation for the work, and Andrew Young for data obtained in his "Megamix" study. The research was funded by NIH grant MH57075 to GWC.

References

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford University Press, Oxford.

Carroll, J. M. and Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70(2):205–218.

Dailey, M. N. and Cottrell, G. W. (1999). PCA = Gabor for expression recognition. UCSD CSE TR CS-629.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Optical Society America A*, 2:1160–1169.

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Classifying facial actions. *IEEE PAMI*, 21(10):974–989.

Ekman, P. (1999). Basic emotions. In Dagleish, T. and Power, M., editors, *Handbook of Cognition and Emotion.* Wiley, New York.

Ekman, P. and Friesen, W. (1976). *Pictures of Facial Affect.* Consulting Psychologists, Palo Alto, CA.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System.* Consulting Psychologists, Palo Alto, CA.

Ellison, J. W. and Massaro, D. W. (1997). Featural evaluation, integration, and judgment of facial affect. *JEP: HPP*, 23:213–226.

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

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.

Katsikitis, M. (1997). The classification of facial expressions of emotion: A multidimensional scaling approach. *Perception*, 26:613–626.

Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE PAMI*, 21(12):1357–1362.

Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorical perception of facial expressions. In *Proc. 20th Cognitive Science Conference*, pages 806–807, Mahwah, NJ. Erlbaum.

Russell, J. A. (1980). A circumplex model of affect. *J. Personality and Social Psych.*, 39:1161–1178.

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE PAMI*, 19(7):775–779.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N., Seth, A., and Perrett, D. I. (1997). Facial expression megamix. *Cognition*, 63:271–313.