EMPATH: A Neural Network that Categorizes Facial Expressions

Matthew N. Dailey[1], Garrison W. Cottrell[1], Curtis Padgett[2], and Ralph Adolphs[3]

Abstract

■ There are two competing theories of facial expression recognition. Some researchers have suggested that it is an example of ''categorical perception.'' In this view, expression categories are considered to be discrete entities with sharp boundaries, and discrimination of nearby pairs of expressive faces is enhanced near those boundaries. Other researchers, however, suggest that facial expression perception is more *graded* and that facial expressions are best thought of as points in a continuous, low-dimensional space, where, for instance, ''surprise'' expressions lie between ''happiness'' and ''fear'' expressions due to their perceptual similarity. In this article, we show that a simple yet biologically plausible neural network model, trained to classify facial expressions into six basic emotions, predicts data used to support both of these theories. Without any parameter tuning, the model matches a variety of psychological data on categorization, similarity, reaction times, discrimination, and recognition difficulty, both qualitatively and quantitatively. We thus explain many of the seemingly complex psychological phenomena related to facial expression perception as natural consequences of the tasks' implementations in the brain. ■

INTRODUCTION

How do we see emotions in facial expressions? Are they perceived as discrete entities, like islands jutting out of the sea, or are they more continuous, reflecting the structure beneath the surface? We believe that computational models of the process can shed light on these questions. Automatic facial expression analysis is an active area of computer vision research (Lien, Kanade, Cohn, & Li, 2000; Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999; Lyons, Budynek, & Akamatsu, 1999; Rosenblum, Yacoob, & Davis, 1996). However, there has only been limited work in applying computational models to the understanding of human facial expression processing (Calder, Burton, Miller, Young, & Akamatsu, 2001; Lyons, Akamatsu, Kamachi, & Gyoba, 1998). In particular, the relationship between categorization and perception is controversial, and a computational model may help elucidate the connection between them.

Basic Emotions and Discrete Facial Expression Categories

Although the details of his theory have evolved substantially since the 1960s, Ekman remains the most vocal proponent of the idea that emotions are discrete entities. In a recent essay, he outlined his theory of basic emotions and their relationship with facial expressions (Ekman, 1999). ''Basic'' emotions are distinct families of affective states characterized by different signals, physiology, appraisal mechanisms, and antecedent events. Ekman cites early evidence suggesting that each emotion is accompanied by distinctive physiological changes that prepare an organism to respond appropriately. For instance, blood flow to the hands increases during anger, possibly in preparation for a fight. In addition to physiological changes, according to the theory, each basic emotion family is also accompanied by a fast appraisal mechanism that attends to relevant stimuli and a set of universal antecedent events (e.g., physical or psychological harm normally leads to a state of fear, and loss of a significant other normally leads to a state of sadness). Finally, and most importantly, Ekman believes that emotions evolved to ''inform conspecifics, without choice or consideration, about what is occurring: inside the person . . . , what most likely occurred . . . , and what is most likely to occur next'' (p. 47). Thus, every basic emotion family is necessarily accompanied by one (or perhaps a few for some families) distinctive prototypical signals, including a set of facial muscle movements and body movements (e.g., approach or withdrawal). The signals are not entirely automatic; they may be attenuated, masked, or faked in certain circumstances. Furthermore, within emotion families, individual differences and situational context allow for small variations on

[1]University of California, San Diego, [2]Jet Propulsion Laboratory, [3]University of Iowa

Note: EMPATH stands for ''EMotion PATtern recognition using Holons'', the name for a system developed by Cottrell & Metcalfe (1991).

the emotion's theme. But between families, the physiology, appraisal mechanisms, antecedents, and signals differ in fundamental ways. Based on these criteria, Ekman proposes that there are 15 basic emotion families: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame. The two crucial components of the theory, which distinguish it from other theorists' approaches, are that emotions are fundamentally separate from one another and that they evolved to help organisms deal with fundamental life tasks.

On this view, since emotions are distinct, and each emotion family is accompanied by a small set of distinctive signals (facial expressions), we might expect subjects' facial expression categorization behavior to exhibit the characteristics of discrete, clear-cut decisions, not smooth, graded, fuzzy categorization. Evidence that facial expressions are perceived as discrete entities, then, would be further evidence for the theory of basic emotions and a deterministic expression/emotion mapping. Indeed, evidence of ''categorical perception'' (CP) of facial expressions has recently emerged in the literature.

In some domains, it appears that sensory systems adapt to impose discontinuous category boundaries in continuous stimulus spaces. For instance, in a rainbow, we perceive bands of discrete colors even though the light's wavelength varies smoothly. In psychophysical experiments, subjects have difficulty discriminating between two shades of green differing by a small constant wavelength distance, but find it easier to distinguish between two stimuli the same distance apart but closer to the green/yellow boundary. This phenomenon is called ''categorical perception'' (Harnad, 1987). It also occurs in auditory perception of phonemes. When we listen to utterances varying continuously from a /ba/ sound to a /pa/ sound, we perceive a sudden shift from /ba/ to /pa/, not a mixture of the two. As with colors, we can also discriminate pairs of equidistant phonemes better when they are closer to the perceived /ba/–/pa/ boundary. In general, CP is assessed operationally in terms of two behavioral measures, categorization judgments and discrimination (same/different) judgments. Categorization measures typically use a forced-choice task, for example, selection of the /ba/ category or the /pa/ category. The stimuli are randomly sampled from smoothly varying continua such as a step-by-step transition between /ba/ and /pa/ prototypes. Even though subjects are not told that the data come from such continua, they nevertheless label all stimuli on one side of some boundary as /ba/, and all stimuli on the other side as /pa/, suggesting a sharp category boundary. For the second behavioral measure of CP, discrimination, subjects are asked to make a ''same/different'' response to a pair of stimuli that are nearby on the continuum (simultaneous discrimination), or perform a sequential (ABX) discrimination task in which stimulus ''A'' is shown, stimulus ''B'' is shown, then either ''A'' or ''B'' is shown and subjects are asked which of the first two stimuli the third matches. For categorically perceived stimuli, subjects show better discrimination when the two stimuli are near the category boundary defined by their labeling behavior, compared with two stimuli further from the boundary.

In some cases, such as the color example, CP is thought to be an innate property of the perceptual system. But in other cases, perceptual discontinuities at category boundaries are clearly acquired through learning. For instance, Beale and Keil (1995) created morph sequences between pairs of famous faces (e.g., Clinton–Kennedy) and unfamiliar faces then had subjects discriminate or categorize neighboring pairs of faces along the morph sequence. They found that famous face pairs exhibited category effects (increased discrimination near the boundaries), but unfamiliar face pairs did not. Their result showed that CP can be acquired through learning and is not limited to low-level perceptual stimuli.

Etcoff and Magee (1992) were the first to raise the question of whether the perceptual mechanisms responsible for facial expression recognition are actually tuned to emotion categories, or whether perception is continuous, with category membership ''assigned by higher conceptual and linguistic systems'' (p. 229). The authors created caricatures (line drawings) of the Ekman and Friesen (1976) photos and 10-step morphs between pairs of those caricatures. They included happy–sad, angry–sad, and angry–afraid continua as easily discriminated category pairs. Surprised–afraid and angry–disgusted continua were included as less easily discriminated pairs. Happy–neutral and sad–neutral continua were included to test for category boundaries along the dimension of presence or nonpresence of an emotion, and finally, happy–surprised continua were added to include a transition between positive emotions. The authors found that all expressions except surprise were perceived categorically: In an ABX task, morph pairs straddling the 50% category boundary were significantly better discriminated than those closer to the prototypes, and in an identification task, subjects placed sharp boundaries between categories, with significantly nonlinear category membership functions. Etcoff and Magee interpreted these results as evidence for mandatory category assignment: ''people cannot help but see the face as showing one or another kind of emotion'' (p. 229). Their results therefore pose a serious challenge for advocates of a continuous space of emotions and rough emotion expression correspondence.

Etcoff and Magee's (1992) provocative results led to further research exploring CP in facial expression recognition. Some of the potential limitations of their study were that the stimuli were line drawings, not image-quality faces, that each subject was only exposed

to a single continuum, and that the discrimination results, being from a sequential (ABX) task, might reflect a short-term memory phenomenon rather than a perceptual phenomenon. In view of these limitations, Calder, Young, Perrett, Etcoff, and Rowland (1996) extended and replicated the earlier experiments with image-quality morphs. They produced several continua using the Ekman and Friesen (1976) photos (e.g., Ekman and Friesen prototypes and image-quality morph sequences produced in R. A.'s lab, see Figure 1). The authors first replicated Etcoff and Magee's experiments with new stimuli: image-quality happy–sad, angry–sad, and angry–afraid sequences, each using a different actor. A second experiment followed the same procedure except that each subject's stimuli included morphs from four different actors. In a third experiment, they had subjects categorize stimuli from three different expression continua employing a single actor (''J. J.'' afraid–happy, happy–angry, and angry–afraid sequences). Finally, they had subjects perform a simultaneous discrimination (same/different) task. In the second experiment, for the happy–sad transitions, the authors found that artifacts in morphing between a face with teeth and one without (a ''graying'' of the teeth as the morph moves away from the happy prototype) helped subjects in the discrimination task. However, on the whole, the results were consistent with CP: sharp category boundaries and enhanced discrimination near the boundaries, regardless of whether there was a single or several continua present in the experiment and regardless of whether the sequential or simultaneous discrimination task was used.

In the psychological literature on categorization, CP effects are usually taken as evidence that (1) object representations are altered during the course of category learning, or (2) that subjects automatically label stimuli even when they are making a purely perceptual discrimination (Goldstone, 2000; Goldstone, Lippa, & Shiffrin, 2001; Pevtzow & Harnad, 1997; Tijsseling & Harnad, 1997). Findings of CP effects in facial expression stimuli raise the possibility that perception of facial expressions is discrete in the same way that category labeling is. Calder et al.'s experiments strengthened the argument for CP of facial expressions, but the authors shy away from Etcoff and Magee's strong interpretation that facial expression category assignment is mandatory. They propose instead that perhaps CP is ''an emergent property of population coding in the nervous system'' occurring whenever ''populations of cells become tuned to distinct categories''(p. 116). In this article, we will show precisely how this can occur, suggesting that Calder et al.'s hypothesis may indeed be correct.

Continuous Emotion Space and Fuzzy Facial Expression Categories

Other theorists hold that the relationship between emotions and facial expressions is not so categorical, discrete, or deterministic as the theory of basic emotions and the evidence for CP suggest. The notion that facial expression perception is discrete is challenged by data showing that similarity judgments of these expressions exhibit a graded, continuous structure.

Schlosberg (1952), following up on the work of his advisor (Woodworth, 1938), found that emotion category ratings and subjects' ''errors'' (e.g., the likelihood of their labeling a putatively disgusted expression as contempt) could be predicted fairly accurately by arranging the emotion categories around an ellipse whose major

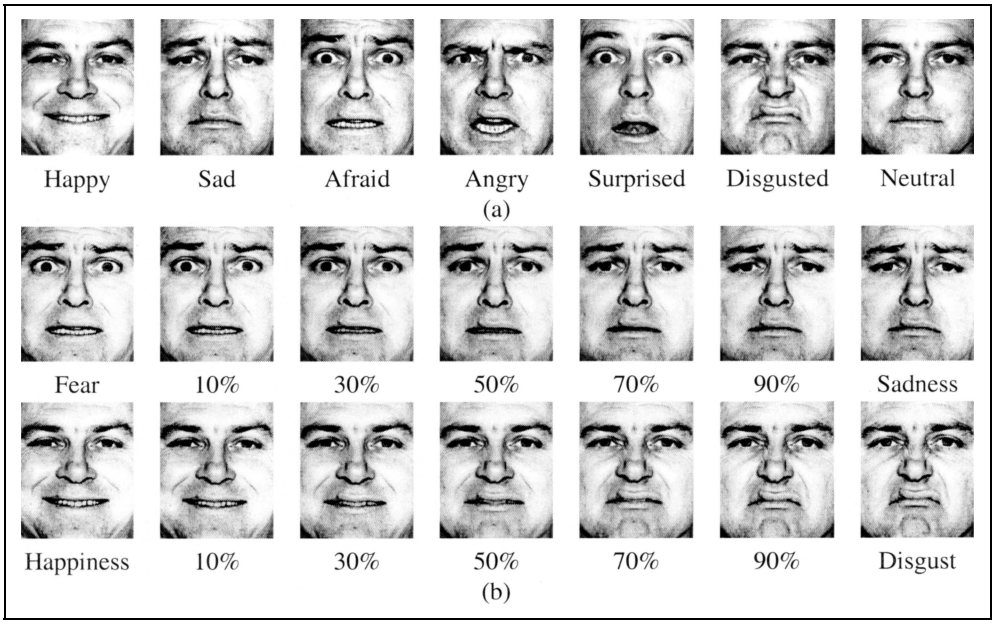

Figure 1. (a) Example prototypical expressions of six basic emotions and a neutral face for actor ''J. J.'' in Ekman and Friesen's POFA (Ekman & Friesen, 1976). (b) Morphs from fear to sadness and happiness to disgust, generated from the corresponding prototypes.

axis was pleasantness versus unpleasantness (exemplified by happy and sad expressions) and whose minor axis was attention versus rejection (exemplified by surprise and disgust expressions).

More recently, Russell (1980) proposed a structural theory of emotion concepts with two dimensions, pleasure and arousal. Russell and Bullock (1986) then proposed that emotion categories are best thought of as fuzzy sets. A few facial expressions might have a membership of 1.0 (100%) in one particular category, and others might have intermediate degrees of membership in more than one category. On their view, the facial expression confusion data supporting structural theories like Schlosberg's (1952) reflected the overlap of these fuzzy categories. To test this concept, Russell and Bullock had subjects rate a variety of facial expressions for how well they exemplify categories like ''excited,'' ''happy,'' ''sleepy,'' ''mad,'' and so forth. They found that the categories did indeed overlap, with peak levels of membership for Ekman's basic emotion categories occurring at Ekman's prototypical expressions. A similarity structure analysis (multidimensional scaling [MDS]—see Figure 9 for an example) performed on the subjects' ratings produced two dimensions highly correlated with other subjects' pleasure and arousal ratings. When asked to verify (yes or no) membership of facial expression stimuli in various emotion concept categories, there was a high level of consensus for prototypical expressions and less consensus for boundary cases. Asking subjects to choose exemplars for a set of categories also revealed graded membership functions for the categories. The data thus showed that facial expression categories have systematically graded overlapping membership in various emotion categories. On the basis of this evidence, Russell and Bullock proposed that facial expression interpretation first involves appraising the expression in terms of pleasure and arousal. Then the interpreter may optionally choose a label for the expression. Following Schlosberg, they proposed that finer, more confident judgments require contextual information.

This and additional recent research (Schiano, Ehrlich, Sheridan, & Beck, 2000; Katsikitis, 1997; Carroll & Russell, 1996; Russell, 1980; Russell, Lewicka, & Niit, 1989) suggests that there is a continuous, multidimensional perceptual space underlying facial expression perception in which some expression categories are more similar to each other than others.

Young et al.'s (1997) ''Megamix'' Experiments

Young et al. (1997) set out to further test the predictions of multidimensional and categorical accounts of facial expression perception and recognition. They pointed out that 2-D structural theories like Schlosberg's or Russell's predict that some morph transitions between expression pairs should pass near other categories. For instance, if the 2-D representation in Figure 9a adequately characterizes our perception of emotional facial expression stimuli, the midpoint of a morph between happiness and fear should be perceived as a surprised expression. Categorical views of emotional facial expressions, on the other hand, do not necessarily predict confusion along morph transitions; one would expect either sharp transitions between all pairs of categories or perhaps indeterminate regions between categories where no emotion is perceived. Furthermore, if the categorical view is correct, we might expect CP, in which subjects find it difficult to discriminate between members of the category and easy to discriminate pairs of stimuli near category boundaries, as found in previous studies (Calder et al., 1996; Etcoff & Magee, 1992). To test these contrasting predictions with an exhaustive set of expression transitions, Young et al. constructed image-quality morph sequences between all pairs of the emotional expression prototypes shown in Figure 1a. (Figure 1b shows example morph sequences produced in R. A.'s lab.)

In Experiment 1, the authors had subjects identify the emotion category in 10%, 30%, 50%, 70%, and 90% morphs between all pairs of the six prototypical expressions in Figure 1a. The stimuli were presented in random order, and subjects were asked to perform a six-way forced-choice identification. Experiment 2 was identical except that morphs between the emotional expressions and the neutral prototype were added to the pool, and ''neutral'' was one of the subjects' choices in a seven-way forced-choice procedure. The results of Experiment 2 for 6 of the 21 possible transitions are reprinted in Figure 2. In both experiments, along every morph transition, subjects' modal response to the stimuli abruptly shifted from one emotion to the other with no indeterminate regions or between. Consistent with categorical theories of facial expressions, the subjects' modal response was always one of the endpoints of the morph sequence, never a nearby category. Subjects' response times (RTs), however, were more consistent with a multidimensional or weak category account of facial expression perception: As distance from the prototypes increased, RTs increased significantly, presumably reflecting increased uncertainty about category membership near category boundaries.

In Experiment 3, Young et al. explored the extent to which subjects could discriminate pairs of stimuli along the six transitions: happiness–surprise, surprise–fear, fear–sadness, sadness–disgust, disgust–anger, anger–happiness. They had subjects do both a sequential discrimination task (ABX) and a simultaneous discrimination task (same–different). Only the data from the sequential experiments are now available, but the authors report a very strong correlation between the two types of data ($r = .82$). The sequential data are reprinted in Figure 3. The main finding of the experiment was that subjects had significantly better discrimination

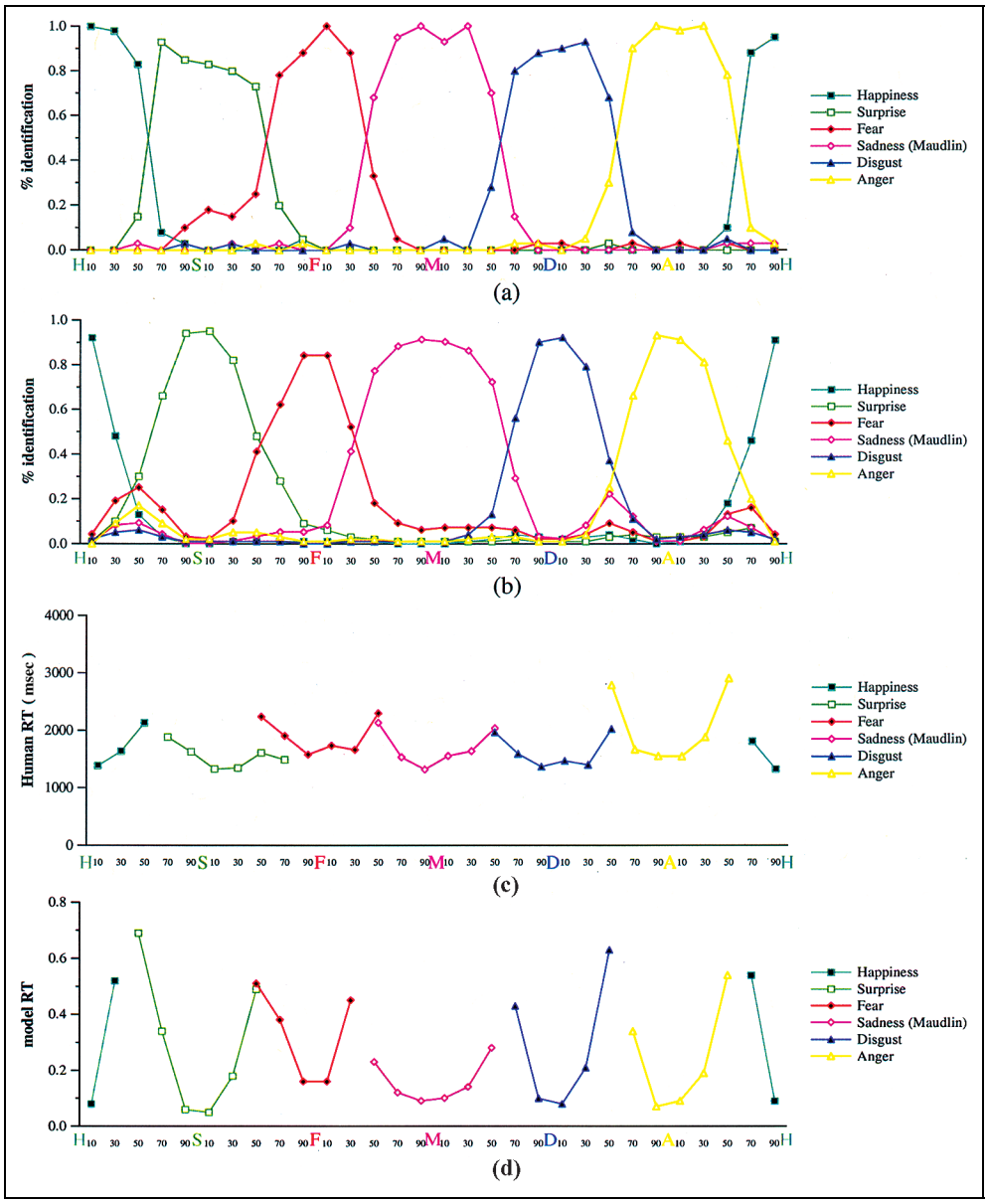

Figure 2. Selected results of Young et al.'s (1997) Experiment 2. (a) Human responses (% identification in a seven-way forced-choice task) to six morph sequences: happy–surprised, surprised–afraid, afraid–sad, sad–disgusted, disgusted–angry, and angry–happy. Every transition exhibited a sharp category boundary. (b) Modeling subjects' choice as a random variable distributed according to the pattern of activation at the network's output layer, which entails averaging across the 13 networks' outputs. The model has a correlation (over all 15 transitions) of $r = .942$ with the human subjects. (c) RTs for the same transitions in (a). RTs increased significantly with distance from the prototype. (M = sad; only 6 out of 21 possible transitions are shown). (d) Predictions of network "uncertainty" model for the same morph transitions.

performance near the category boundaries than near the expression prototypes, a necessary condition to claim CP. Experiment 3 therefore best supports the categorical view of facial expression perception.

Finally, in Experiment 4, Young et al. set out to determine whether subjects could determine what expression is "mixed-in" to a faint morph. Again, a strong categorical view of facial expressions would predict that subjects should not be able to discern what expression a given morph sequence is moving toward until the sequence gets near the category boundary. However, according to continuous theories, subjects should be able to determine what prototype a sequence is moving toward fairly early in the sequence. In the experiment, subjects were asked to decide, given a morph or prototype stimulus, the most apparent emotion, the second-most apparent emotion, and the third-most apparent emotion, using a button box with a button for each of the six emotion categories. After correcting for intrinsic confusability of the emotion categories, the authors found that subjects were significantly above chance at detecting the mixed-in emotion at the 30% level. As opposed to the identification and discrimination data from Experiments 1–3, this result best supports continuous, dimensional accounts of facial expression perception.

In summary, Young et al.'s experiments, rather than settling the issue of categorical versus continuous theories of facial expression perception, found evidence supporting both types of theory. The authors argue that a 2-D account of facial expression perception is unable to account for all of the data, but they also argue that the strong mandatory categorization view is likewise deficient.

Until now, despite several years of research on automatic recognition of facial expressions, no computational model has been shown to simultaneously explain all of these seemingly contradictory data. In the next section, we review the recent progress in computational modeling of facial expression recognition, then introduce a new model that does succeed in explaining the available data.

Computational Modeling of Facial Expression Recognition

Padgett and colleagues were the first to apply computational models toward an understanding of facial expression perception (Cottrell, Dailey, Padgett, & Adolphs, 2000; Padgett, Cottrell, & Adolphs, 1996; Padgett & Cottrell, 1998). Their system employed linear filtering of regions around the eyes and mouth followed by a multilayer perception for classification into emotion categories. The system achieved good recognition performance and was able to explain some of the results on CP of facial expressions with linear ''dissolve'' sequences. However, the system was unable to account for the sharp category boundaries humans place along image-quality morph transitions, because the linear filters' responses varied too quickly in the presence of the nonlinear changes in morph transitions. It also would have been incapable of predicting recently observed evidence of holistic effects in facial expression recognition (Calder, Young, Keane, & Dean, 2000).

Lyons et al. (1998) created a database of Japanese female models portraying facial expressions of happiness, surprise, fear, anger, sadness, and disgust. They then had subjects rate the degree to which each face portrayed each basic emotion on a 1–5 scale, and used Euclidean distance between the resulting ''semantic rating'' vectors for each face pair as a measure of dissimilarity. They then used a system inspired by the Dynamic Link Architecture (Lades et al., 1993) to explain the human similarity matrix. Similarities obtained from their Gabor wavelet-based representation of faces were highly correlated with similarities obtained from the human data, and nonmetric multidimensional scaling (MDS) revealed similar underlying 2-D configurations of the stimuli. The authors suggest in conclusion that the high-level circumplex constructs proposed by authors like Schlosberg and Russell may in part reflect similarity at low levels in the visual system.

In a recent work, Calder et al. (2001) applied a different computational model to the task of understanding human facial expression perception and recognition. The idea of their system, originally proposed by Craw and Cameron (1991), is to first encode the positions of facial features relative to the average face then warp each face to the average shape (thus preserving texture but removing between-subject face shape variations). The shape information (the positions of the facial features prior to warping) and the shape-free information (the pixel values after warping) can each be submitted to a separate principal components analysis (PCA) for linear dimensionality reduction, producing separate low-dimensional descriptions of the face's shape and texture. Models based on this approach have been successful in explaining psychological effects in face recognition (Hancock, Burton, & Bruce, 1996, 1998). However, warping the features in an expressive face to the average face shape would seem to destroy some of the information crucial to recognition of that expression, so prior to Calder et al.'s work, it was an open question as to whether such a system could support effective classification of facial expressions. The authors applied the Craw and Cameron PCA method to Ekman and Friesen's (1976) Pictures of Facial Affect (POFA) then used linear discriminant analysis to classify the faces into happy, sad, afraid, angry, surprised, and disgusted categories. The authors found that a representation incorporating both the shape information (feature location PCA) and the shape-free information (warped pixel PCA) best supported facial expression classification (83% accuracy

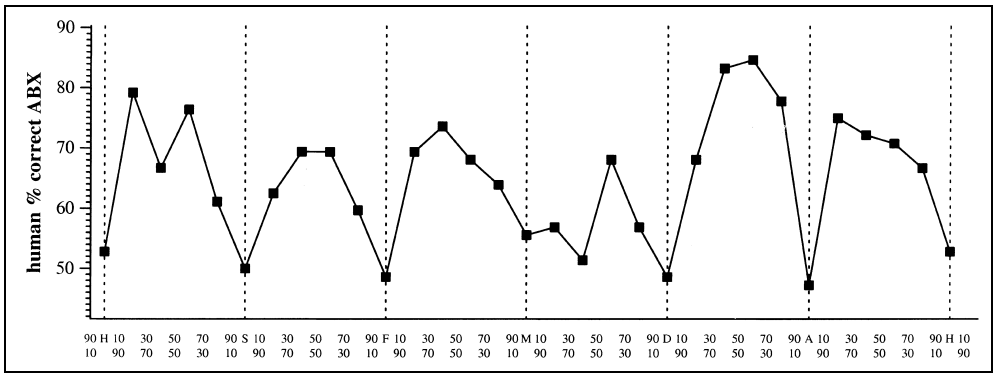

Figure 3. Results of Young et al.'s (1997) Experiment 3 for the sequential (ABX) discrimination task. Each point represents the percentage of time subjects correctly discriminated between two neighboring morph stimuli. The *x*-axis labels denote which pair of stimuli were being compared (e.g., 70/90 along the transition from sadness to disgust denotes a comparison of a 70% disgust/30% sadness morph with a 90% disgust/10% sadness morph). Discrimination was significantly better near the prototypes than near the category boundaries.

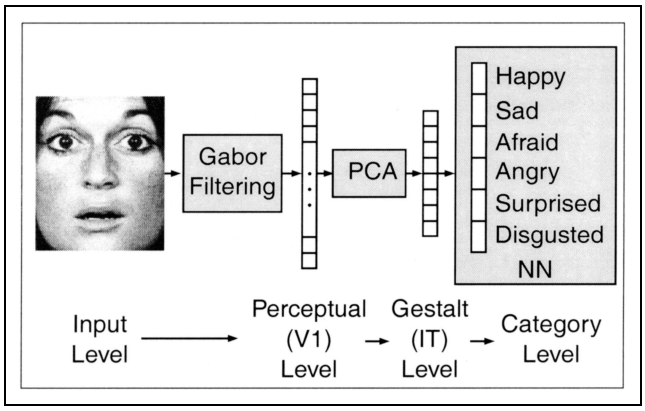

Figure 4. Model schematic.

using leave-one-image-out classification for the Ekman and Friesen database). The authors went on to compare their system with human data from psychological experiments. They found that their system behaved similarly to humans in a seven-way forced-choice task and that the principal component representation could be used to predict human ratings of pleasure and arousal (the two dimensions of Russell's circumplex).

Taken together, results with the above three computational models of facial expression recognition begin to hint that subjects' performance in psychological experiments can be explained as a simple consequence of category learning and the statistical properties of the stimuli themselves. Although Calder et al.'s system exhibits a surprising amount of similarity to human performance in a forced-choice experiment, it has not been brought to bear in the debate on multidimensional versus CP of facial expressions. In the present article, we show that a new, more biologically plausible computational model not only exhibits more similarity to human forced-choice performance, but also provides a detailed computational account of data supporting both categorical and multidimensional theories of facial expression recognition and perception. Our simulations consist of constructing and training a simple neural network model (Figure 4); the system uses the same stimuli employed in many psychological experiments, performs many of the same tasks, and can be measured in similar ways as human subjects. The model consists of three levels of processing: perceptual analysis, object representation, and categorization. The next section describes the model in some detail.

The Model

The system is a feed-forward network consisting of three layers common to most object recognition models (Riesenhuber & Poggio, 2000). Its input is a 240 by 292 manually aligned, grayscale face image from Ekman and Friesen's POFA (Ekman & Friesen, 1976). This data set contains photos of 14 actors portraying expressions that are reliably classified as happy, sad, afraid, angry, surprised, or disgusted by naive observers (70% agreement was the threshold for inclusion in the data set, and the overall agreement is 91.6%). The expressions made by one of those actors, ''J. J.,'' are shown in Figure 1a. The strong agreement across human subjects, together with the use of these photos in a wide variety of psychological experiments exploring emotional expression perception, makes POFA an ideal training set for our model.

The first layer of the model is a set of neurons whose response properties are similar to those of complex cells in the visual cortex. The so-called Gabor filter (Daugman, 1985) is a standard way to model complex cells in visual recognition systems (Lades et al., 1993). Figure 5 shows the spatial ''receptive fields'' of several filters. The units essentially perform nonlinear edge detection at five scales and eight orientations. As a feature detector, the Gabor filter has the advantage of moderate translation invariance over pixel-based representations. This means that features can ''move'' in the receptive field without dramatically affecting the detector's response, making the first layer of our model robust to small image changes. With a 29 by 35 grid and 40 filters at each grid location, we obtain 40,600 model neurons in this layer, which we term the ''perceptual'' layer.

In order to extract a small set of informative features from this high-dimensional data, we use the equivalent of an ''image compression'' network (Cottrell, Munro, & Zipser, 1989). This is a back propagation network that is trained to simply reproduce its input on its output through a narrow channel of hidden units. In order to solve this problem, the hidden units must extract regularities from the data. Since they are fully connected to the inputs, they usually extract global representations we have termed ''holons'' in previous work (Cottrell & Metcalfe, 1991). We note that this is a biologically plausible means of dimensionality reduction in the sense that it is unsupervised and can be learned by simple networks employing Hebbian learning rules (Sanger, 1989). As a shortcut, we compute this network's weights

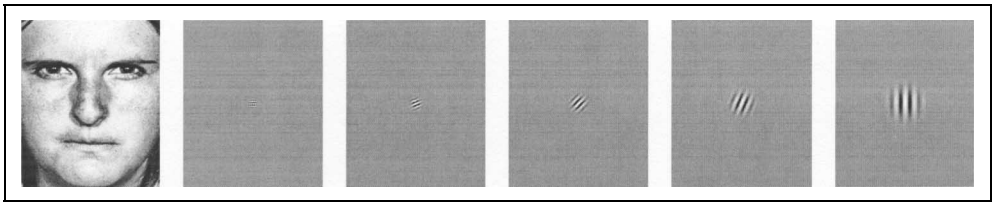

Figure 5. Example Gabor filters. The real (cosine shaped) component is shown relative to the size of the face at all five scales and five of the eight orientations used.

directly via the equivalent statistical technique of PCA. We chose 50 principal components (hidden units) for this layer based on previous experiments showing this leads to good generalization on POFA. It is important to note that the number of components (the only free parameter in our system) was not tuned to human data. The resulting low-dimensional object-level representation is specific to the facial expression and identity variations in its input, as is the population of so-called face cells in the inferior temporal cortex (Perrett, Hietanen, Oram, & Benson, 1992). We term this layer of processing the ''gestalt'' level.

Although PCA is one of the simplest and most efficient methods for coding a set of stimuli, other methods would probably also work. For the current model, it is only important that (1) the code be sensitive to the dimensions of variance in faces, to facilitate learning of expressions, and that (2) the code be low dimensional, to facilitate generalization to novel faces. In a more general object recognition system, a single PCA for all kinds of objects would probably not be appropriate, because the resulting code would not be sparse (a single image normally activates a large number of holistic ''units'' in a PCA representation, but object-sensitive cells in the visual system appear to respond much more selectively; Logothetis & Sheinberg, 1996). To obtain a good code for a large number of different objects, then, nonnegative matrix factorization (Lee & Seung, 1999) or an independent component mixture model (Lee, Lewicki, & Sejnowski, 2000) might be more appropriate. For the current problem, though, PCA suffices.

The outputs of the gestalt layer are finally categorized into the six ''basic'' emotions by a simple perceptron with six outputs, one for each emotion. The network is set up so that its outputs can be interpreted as probabilities (they are all positive and sum to 1). However, the system is trained with an ''all-or-none'' teaching signal. That is, even though only 92% (say) of the human subjects used to vet the POFA database may have responded ''happy'' to a particular face, the network's teaching signal is 1 for the ''happy'' unit, and 0 for the other five. Thus, the network does not have available to it the confusions that subjects make on the data. We term this layer the ''category'' level.

While this categorization layer is an extremely simplistic model of human category learning and decision-making processes, we argue that the particular form of classifier is unimportant; so long as it is sufficiently powerful and reliable to place the gestalt-level representations into emotion categories, we claim that similar results will obtain with any nonlinear classifier.

We should also point out that the system abstracts away many important aspects of visual processing in the brain, such as eye movements, facial expression dynamics, size, and viewpoint invariance. These complicating factors turn out to be irrelevant for our purposes; as we shall see, despite the simplifying assumptions of the model, it nevertheless accounts for a wide variety of data available from controlled behavioral experiments. This suggests that it is a good first-order approximation of processing at the computational level in the visual system.

The next section reports on the results of several experiments comparing the model's predictions to human data from identical tasks, with special emphasis on Young et al.'s landmark study of categorical effects in facial expression perception (Young et al., 1997). We find (1) that the model and humans find the same expressions difficult or easy to interpret; (2) that when presented with morphs between pairs of expressions, the model and humans place similar sharp category boundaries between the prototypes; (3) that pairwise similarity ratings derived from the model's gestalt-level representations predict human discrimination ability; (4) that the model and humans are similarly sensitive to mixed-in expressions in morph stimuli; and (5) that MDS analysis produces a similar emotional similarity structure from the model and human data.

RESULTS

Comparison of Model and Human Performance

The connections in the model's final layer were trained to classify images of facial expressions from Ekman and Friesen's POFA database (see Figure 1a) (Ekman & Friesen, 1976), the standard data set used in the vast majority of psychological research on facial expression (see Methods for details). The training signal contained no information aside from the expression most agreed upon by human observers—that is, even if 90% of human observers labeled a particular face ''happy'' and 10% labeled it ''surprised,'' the network was trained as if 100% had said ''happy.'' Again, there is no information in the training signal concerning the similarity of different expression categories. The first measurement we made was the model's ability to generalize in classifying the expressions of previously unseen subjects from the same database. The model's mean generalization performance was 90.0%, which is not significantly different from human performance on the same stimuli (91.6%; $t = .587$, $df = 190$, $p = .279$) (Table 1). Moreover, the rank-order correlation between the model's average accuracy on each category (happiness, surprise, fear, etc.), and the level of human agreement on the same categories was .667 (two-tailed Kendall's tau, $p = .044$; cf. Table 1). For example, both the model and humans found happy faces easy and fear faces the most difficult to categorize correctly.[1] Since the network had about as much experience with one category as another, and was not trained on the human response accuracies, this correlation between the relative difficulties of each category is an ''emergent'' property of the

Table 1. Error Rates for Networks and Level of Agreement for Humans

Expression	*Network Percentage Correct*	*Human Percentage Correct*
Happiness	100.0	98.7
Surprise	100.0	92.4
Disgust	100.0	92.3
Anger	89.1	88.9
Sadness	83.3	89.2
Fear	67.2	87.7
Average	90.0	91.6

Network generalization to unseen faces, compared with human agreement on the same faces (six-way forced choice). Human data are provided with the POFA database (Ekman & Friesen, 1976).

model. Studies of expression recognition consistently find that fear is one of the most difficult expressions to recognize (Katsikitis, 1997; Matsumoto, 1992; Ekman & Friesen, 1976). Our model suggests that this is simply because the fear expression is ''inherently'' difficult to distinguish from the other five expressions.

How does the network perform the expression recognition task? An examination of the trained network's representation provides some insight. We projected each unit's weight vector back into image space in order to visualize the ''ideal'' stimulus for each unit in the network (see Methods for details). The results are shown in Figure 6, with and without addition of the average face. In each of the images, each pixel value is the result of applying a regression formula predicting the value of the pixel at that location as a linear function of the 50-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. Each unit appears to combine evidence for an emotion based upon the presence or absence of a few local features. For instance, for fear, the salient criteria appear to be the eyebrow raise and the eyelid raise, with a smaller contribution of parted lips. The anger unit apparently requires a display in which the eyebrows are not raised. Clearly, the classifier has determined which feature configurations reliably distinguish each expression from the others.

Comparison on CP Data

Several studies have reported CP of facial expressions using morphs between portrayals of the six basic emotions by POFA actor ''J. J.,'' whose images have been chosen because his expressions are among the easiest to recognize in the database (Young et al., 1997; Calder et al., 1996). Since our model also finds J. J. ''easy'' (it achieves 100% generalization accuracy on J. J.'s prototypical expressions), we replicated these studies with 13 networks that had not been trained on J. J.[2]

We first compared the model's performance with human data from a six-way forced-choice experiment (Young et al., 1997), on 10%, 30%,50%, 70%, and 90% morphs we constructed (see Methods) between all pairs of J. J.'s prototypical expressions (see Figure 1b for two such morph sequences). We modeled the subjects' identification choices by letting the outputs of the networks represent the probability mass function for the human subjects' responses. This means that we averaged each of the six outputs of the 13 networks for each face and compared these numbers to the human subjects' response probabilities. Figure 2 compares the networks' forced-choice identification performance with that of humans on the same stimuli. Quantitatively, the model's responses were highly correlated with the human data ($r = .942$, $p < .001$), and qualitatively, the model maintained the essential features of the human data: abrupt shifts in classification at emotion category boundaries, and few intrusions (identifications of unrelated expressions) along the transitions. Using the same criteria for an intrusion that Young et al. (1997) used, the model predicted 4 intrusions in the 15 morph sequences, compared to 2 out of 15 in the human data. Every one of the intrusions predicted by the model involved fear, which

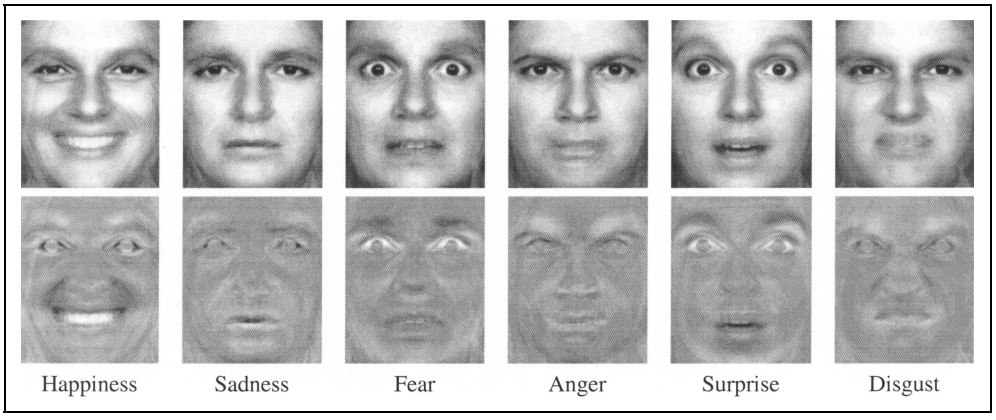

Figure 6. Network representation of each facial expression. (Top row) Approximation of the optimal input-level stimulus for each facial expression category. (Bottom row) The same approximations with the average face subtracted—dark and bright pixels indicate salient features.

is the least reliably classified expression in the POFA database, for humans and networks (Table 1).

We also compared subjects' RTs to those of the network, in which we model RT as being proportional to uncertainty at the category level. That is, if the network outputs a probability of .9 for happiness, we assume it is more certain and therefore faster than when it outputs a probability of .6 for happiness. We thus use the difference between 1 (the maximum possible output) and the largest actual output as our model of the reaction time. As explained earlier, subjects' RTs exhibit a characteristic scalloped shape, with slower RTs near category boundaries. The model also exhibits this pattern, as shown in Figure 2. The reason should be obvious: As a sequence of stimuli approaches a category boundary, the network's output for one category necessarily decreases as the other increases. This results in a longer model RT. The model showed good correlation with human RTs (see Methods) ($r = .677$, $p < .001$).

The model also provides insight into human discrimination behavior. How can a model that processes one face at a time discriminate faces? The idea is to imagine that the model is shown one face at a time and stores its representation of each face for comparison. We use the correlation (Pearson's r) between representations as a measure of similarity, and then use 1 minus this number as a measure of discriminability. An interesting aspect of our model is that it has several independent levels of processing (see Figure 4), allowing us to determine which level best accounts for a particular phenomenon. In this case, we compute the similarity between two faces at the pixel level (correlation between the raw images), the perceptual level (correlation between the Gabor filter responses), the gestalt level (correlation between the principal components), or the categorization level (correlation between the six outputs). We compared our measure of discriminability with human performance in Young et al.'s (1997) ABX experiment (see Methods for details). We found the following correlations at each processing level. Pixel: $r = .35$, $p = .06$; perceptual: $r = .52$, $p = .003$; gestalt: $r = .65$, $p < .001$ (shown in Figure 7b); category: $r = .41$, $p = .02$. Crucially, when the gestalt layer and the categorization layer were combined in a multiple regression, the categorization layer's contribution was insignificant ($p = .3$), showing that the explanatory power of the model rests with the gestalt layer. According to our model, then, human subjects' improved discrimination near category boundaries in this task is best explained as an effect at the level of gestalt-based representations, which were derived in an unsupervised way from Gabor representations via PCA. This is in sharp contrast to the standard explanation of this increased sensitivity, which is that categorization influences perception (Goldstone et al., 2001; Pevtzow & Harnad, 1997). This suggests that the facial expression morph sequences have natural boundaries, possibly because the endpoints are extremes of certain coordinated muscle group movements. In other domains, such as familiar face classification (Beale & Keil, 1995), the boundaries must arise through learning. In such domains, we expect that discrimination would be best explained in a learned

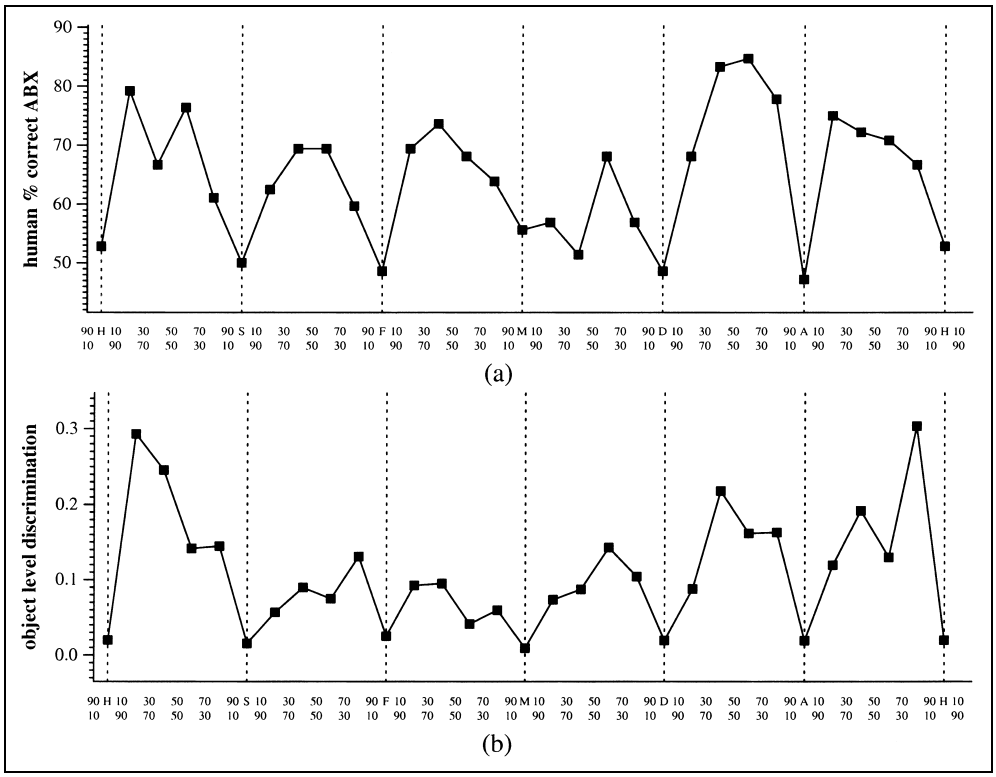

Figure 7. Discrimination of morph stimuli. (a) Percent correct discrimination of pairs of stimuli in an ABX task (Young et al., 1997). Each point represents the percentage of time subjects correctly discriminated between two neighboring morph stimuli. The *x*-axis labels to the left and right of each point show which two stimuli were being compared; (e.g., 70/90 along the transition from sadness to disgust denotes a comparison of a 70% disgust/30% sadness morph with a 90% disgust/10% sadness morph). Note that better performance occurs near category boundaries than near prototypes (highlighted by the vertical lines). (b) The model's discrimination performance at the gestalt representation level. The model discrimination scores are highly correlated with the human subjects' scores ($r = .65$, $p < .001$).

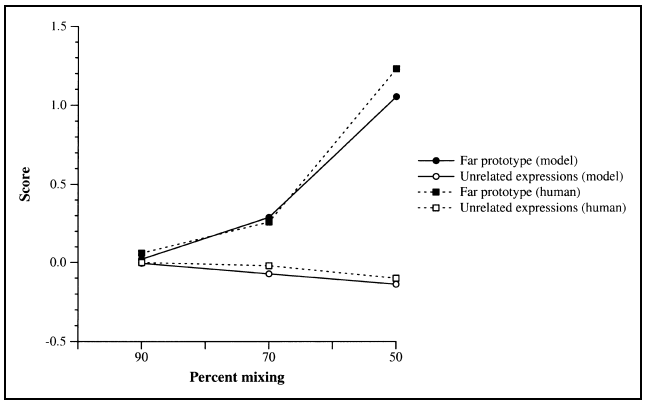

Figure 8. Ability to detect mixed-in expression in morph sequences, for humans and the model. In the human data (dashed lines), subjects chose the far prototype (the faint second expression in the morph) significantly more often than unrelated expressions when that expression's mix ratio was 30% or greater, even though they consistently identified the 70% expression in forced-choice experiments. The model is almost identical to the humans in its ability to detect the secondary expression in a morph image.

feature layer, as in other models (Goldstone, 2000; Tijsseling & Harnad, 1997).

The data above show that the model naturally accounts for categorical behavior of humans making forced-choice responses to facial expression morph stimuli. In another experiment, Young et al. (1997) found that subjects could reliably detect the expression mixed into a morph even at the 30% level. Can the model also account for this decidedly noncategorical behavior? In the experiment, subjects were asked to decide, given a morph or prototype stimulus, the most apparent emotion, the second-most apparent emotion, and the third-most apparent emotion (the scores on prototypes were used to normalize for inherent expression similarity). To compare the human responses with the model, we used the top three outputs of the 13 networks, and used the same analysis as Young et al. to determine the extent to which the network could detect the mixed-in expression in the morph images (see Methods for details). Figure 8 shows that the model's average sensitivity to the mixed-in expression is almost identical to that of human subjects, even though its behavior seems categorical in forced-choice experiments.

Comparison of Similarity Structures

We next investigated the similarity structure of the representations that the model produced. As before, we calculated the similarity between pairs of faces at each level of the network by computing the correlation between their representations. In order to evaluate the similarity structure qualitatively, we performed MDS both on the human forced-choice responses published by Ekman and Friesen (1976) and on the network's responses to the same stimuli, at each level of processing shown in Figure 4. At the pixel level, we found that the structure present in the MDS configuration was based mainly on identity, and as we moved toward the categorization level, the identity-based structure began to break down, and expression-related clusters began to emerge. Unsurprisingly, at the network's categorization layer, the configuration was clearly organized by emotion category (Figure 9). Of more interest, however, is that the "ordering" of facial expressions around the human and network MDS configurations is the same, a result unlikely to have arisen by chance

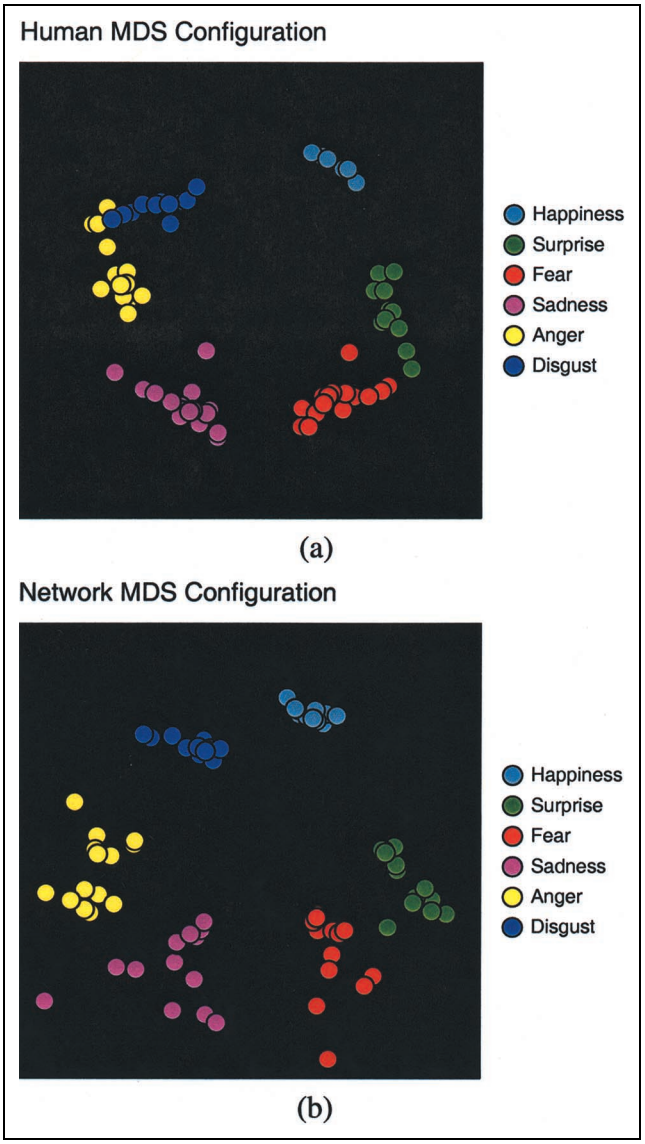

Figure 9. Multidimensional scaling of human and network responses reveals similar dimensions of emotion. Each point represents one of the 96 expressive faces in POFA. (a) 2-D similarity space induced by MDS from the average six-way forced-choice responses of human subjects (Ekman & Friesen, 1976) (stress = 0.218). (b) 2-D similarity space induced by MDS from the average training set responses of networks at their output layers (stress = 0.201). The arrangement of emotions around the circle is the same in both cases. Stress at the pixel level was 0.222, the perceptual level 0.245, and the gestalt level 0.286.

($p = 1/60 = .017$). Given that the network was never given any similarity information, this is a remarkable result. It suggests that the human similarity structure is a simple result of the inherent confusability of the categories, not necessarily the result of locality in some underlying psychological emotion space, as dimensional theories (e.g., Russell, 1980; Russell et al., 1989) might predict.

To measure the similarity between the human and network category structures quantitatively, we compared the human and model confusion matrices directly. For each pair of expressions, we computed the probability p_{ij} that humans or networks respond with emotion i when the intended emotion is j for all $i \neq j$, that is, the probability of confusion between two categories. The correlation between networks and humans on the networks' test sets (the stimuli the networks had never seen before) was .661 ($p < .001$).[3] Thus, there is a close correspondence between human confusion rates and response distributions at the model's categorization level. The system is never instructed to confuse the categories in a way similar to humans; nevertheless, this property emerges naturally from the data.

DISCUSSION

In this article, we have introduced a computational model that mimics some of the important functions of the visual system. The model is simply a pattern classifier incorporating a biologically plausible representation of visual stimuli. We first engineered the system to provide good classification performance on a small database of reliably recognizable facial expressions. The free parameters of the model, such as the number of principal components used in dimensionality reduction, were optimized to maximize the classifier's generalization performance. In contrast to most models in mathematical psychology, which seek to fit a low number of free parameters to maximize a model's agreement with human data, our model is compared to human data directly without any tuning.

The results of our comparison of the model's performance with human data were nevertheless remarkable. We first found that the relative levels of difficulty for six basic facial expressions of emotion were highly correlated with the levels at which humans agree on those same emotions. For example, humans are best at classifying happy expressions because the smile makes the task easy. The network model likewise finds it extremely easy to detect smiles because they are obvious visual features that discriminate happy expressions from the rest of the faces in its training set. On the other hand, humans usually find that fear expressions are very difficult to classify. Fear expressions are often classified as surprise, for instance. The network model likewise sometimes classifies fear expressions as surprise. Why are fear expressions so difficult for humans to classify? Could it be that true displays of fear are uncommon in our society? Or that portrayals of fear in the popular culture misguide us as to what fearful people really look like? Or simply that fear expressions are perceptually similar to other expressions? Our results suggest that the latter is the case—culture probably has very little to do with the difficulty of fear expressions. We have shown that perceptual similarity alone is sufficient to account for the relative difficulty of facial expressions.

The agreement between human and network similarity structure analyses (MDS) is also somewhat surprising. As pointed out earlier, Russell and colleagues have found that affective adjectives and affective facial expressions seem to share a common similarity structure. One might postulate that the underlying psychological space reflects the physiological similarity of emotional states, and that when we are asked to classify facial expressions, we are likely to classify a given face as any of the emotions close by in that psychophysiological space. However, we have shown that an emotionless machine, without any underlying physiology, exhibits a similarity structure very much like that of the humans. This is even more remarkable when one considers that the network was not given any indication of the similarity structure in the training signal, which was always all-or-none, rather than, for example, human subject responses, which would reflect subjects' confusions. Since we nevertheless match the human similarity structure, we have shown that the perceptual similarity of the categories corresponds to the psychological similarity structure of facial expressions. Why would this be? We suggest that evolution did not randomly associate facial expressions with emotional states, but that the expression-to-emotion mapping evolved in tandem with the need to communicate emotional states effectively.

The final question we set out to explore with this work is whether facial expressions are represented continuously or as discrete entities. As explained in the Introduction, the best evidence for discrete representations is that subjects appear to place sharp boundaries between facial expression categories and are better able to discriminate pairs of expressions near category boundaries, suggesting that our perception of the faces is influenced by the existence of the categories. As has been found in other modalities (cf. Ellison & Massaro, 1997), we find that it is unnecessary to posit discrete representations to account for the sharp boundaries and high discrimination scores. The network model places boundaries between categories, as it must to obtain good classification accuracy, but the categories are actually fuzzy and overlapping. The model can be thought of as a biologically plausible, working implementation of Russell and Bullock's (1986) theory of emotional facial expression categories as fuzzy concepts. Despite the network's fuzzy, overlapping category concepts, the categories appear

sharp when a rule such as ''respond with the category most strongly activated by the given facial expression'' is applied. A more difficult result to explain in terms of fuzzy categories placed in a continuous multidimensional space is the subjects' high discrimination scores near category boundaries. This result seems to call for an influence of the category boundaries on our perception of the expressions; indeed, that is Young et al.'s (1997) tentative interpretation. To the contrary, we have shown that in our model, discrimination scores best agree with human results at a purely perceptual level, where the category labels have no effect. With respect to image space, the low-dimensional PCA representation actually changes faster in boundary regions than in regions near the prototypes it derived from. In the context of our model, then, the seemingly contradictory results in different experiments conducted by Young et al. can be explained as simply tapping different computational levels of processing in a visual system organized much like our model.

We have found that our facial expression recognition model is ''sufficient'' to explain many aspects of human performance in behavioral tasks, but we have no proof of the ''necessity'' of any of our particular implementation decisions. In fact, we predict that many similar systems would obtain similar results. For instance, a system beginning with derivative-of-Gaussian edge filters (Marr, 1982) whose responses are combined to produce smooth responses to translations (like complex cells) should exhibit the same behavior. Replacing the PCA dimensionality reduction method with, say, factor analysis or the Infomax algorithm for independent components analysis (Bell & Sejnowski, 1995) should not dramatically affect the results. Finally, a category level using Support Vector Machines (Vapnik, 1995) should likewise produce similar behavior. The point of our simulations is that the category boundaries, discriminability, and similarity structures previously seen as being at odds are in a sense ''present in the data and tasks themselves,'' and are easily exposed by any reasonable computational model.

METHODS

Network Details

The first step of processing in the model is to filter the image with a rigid 29 by 35 grid of overlapping 2-D Gabor filters (Daugman, 1985) in quadrature pairs at five scales and eight orientations (some example filters are shown in Figure 5). The quadrature pairs are used to compute a phase insensitive energy response at each point in the grid. These linear energy responses, or ''Gabor magnitudes,'' are often used as a simplifying model of the spatial responses of complex cells in the early visual system (Lades et al., 1993). Though the model loses some information about precise feature localization when phase information is thrown away, the overlapping receptive fields compensate for this loss (Hinton, McClelland, & Rumelhart, 1986). Each Gabor magnitude is z-scored (linearly transformed to have mean 0 and variance 1 over the training data) so that each filter contributes equally in the next representation layer.

The second step of processing in the model is to perform linear dimensionality reduction on the 40,600-element Gabor representation via a PCA of the training set. The actual computation is facilitated by Turk and Pentland's (1991) algebraic trick for ''eigenfaces.'' This produces a 50-element representation typically accounting for approximately 80% of the variance in the training set's Gabor data.

The 50-element vector p output by PCA can then be classified by a simple statistical model. We locally z-score (scale to mean 0, variance 1) each input element then use a single-layer neural network (a generalized linear model) containing six outputs, one for each of the six ''basic'' emotions happiness, sadness, fear, anger, surprise, and disgust. Each of the six units in the network computes a weighted sum $o_i = \sum_j w_{ij} p_j$ of the 50-element input vector, then the ''softmax'' function $y_i = e^{o_i} / \sum_j e^{o_j}$ is applied to the units' linear activations to obtain a vector of positive values whose sum is 1.0. The network is trained with the relative entropy error function so that its outputs correspond to the posterior probabilities of the emotion categories given the inputs (Bishop, 1995).

Network Training

We train the expression recognition system using leave-one-out cross-validation and early stopping. A given network is trained to minimize output error on all the images of 12 of the 14 actors in POFA, using stochastic gradient descent, momentum, weight decay, and the relative entropy error function (Bishop, 1995). A thirteenth actor's images are used as a ''hold out'' set to determine when to stop training: Training is stopped when the error on the hold out set is minimized. After training is stopped, we evaluate the network's generalization performance on the remaining (fourteenth) actor. We performed training runs with every possible combination of generalization and hold out sets, for a total of 182 (14 by 13) individual networks.

Network Weight Visualization

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 well-defined task; though PCA is an easily inverted orthonormal transformation, the Gabor magnitude representation, besides being subsampled, throws away important phase information. As a workaround, we assume that each pixel's value is an approximately linear function of

the 50-component gestalt-level (Gabor + PCA) representation. We chose one of the 192 trained networks, and for each pixel location, we used regression to find the linear function of the 50-element gestalt-level representation best predicting the pixel's value over the network's training set. Then to visualize, say, the classifier's representation of happiness, we apply the regression function directly to happiness unit's weights. An image was constructed from each of the six units' weight vectors using the regression functions learned from that network's training set.

Generation of Morphs

We generated morphs from the original stimuli (Ekman & Friesen, 1976) using the Morph program, version 2.5, from Gryphon Software, as described elsewhere (Jansari, Tranel, & Adolphs, 2000). Briefly, for each of the 15 pairs of expression prototypes, corresponding features between the two images were manually specified. The images were then tessellated and linearly transformed both with respect to pixel position (a smooth warping) and pixel grayscale value (a smooth fade in luminance). The 10%, 30%, 50%, 70%, and 90% blends (see Figure 1b for examples) were retained for each transformation.

Multidimensional Scaling

MDS seeks to embed a set of stimuli in a low-dimensional space in such a way that the distances between points in the low-dimensional space are as faithful as possible to ratings of their similarity (Borg & Lingoes, 1987). We performed MDS on the human responses (Ekman & Friesen, 1976) and on the network's responses to the same stimuli, at each level of processing in the network. Each analysis requires a distance (dissimilarity) matrix enumerating how dissimilar each pair of stimuli is. For the human data, we formed a six-element vector containing the probability with which humans gave the labels happy, sad, afraid, angry, surprised, and disgusted, for each of the 96 nonneutral photographs in POFA. We obtained a 96 by 96 element similarity matrix from these data by computing the correlation r_{ij} between each pair of six-element vectors. Finally, we converted the resulting similarity matrix into a distance (dissimilarity) matrix with the transform $d_{ij} = (1 - r_{ij})/2$ to obtain values in the range 0–1. For the network, we measure the similarity between two stimuli i and j as the correlation r_{ij} between the representations of the two stimuli at each level of the network, corresponding to similarity at different levels of processing: the pixel level (70,080-element image pixel value vectors), the perceptual level (40,600-element Gabor response patterns), the gestalt level (50-element Gabor/PCA patterns), and the network's output (six-element category level). We ran a classical nonmetric MDS algorithm, SSA-1, due to Guttman and Lingoes (Borg & Lingoes, 1987), on each of the four resulting distance matrices described above and then plotted the stimuli according to their position in the resulting 2-D configuration.

Model RTs

We assume a network's RT is directly proportional to the ''uncertainty of its maximal output.'' That is, we define a network's model RT for stimulus i to be $t_i^{\text{model}} = 1 - \max_j y_{ij}$ (the time scale is arbitrary). Here y_{ij} is the network's output for emotion j on stimulus i. This is similar to the standard approach of equating RT with a network's output error (Seidenberg & McClelland, 1989), except that for morph stimuli, there is no predetermined ''correct'' response. For comparison, the human data available are t_{ij}^{human}, the mean RT of subjects responding with emotion j to stimulus i. To compare the model to the humans, given these data, we treated each network as an individual subject, and for each morph stimulus, recorded the network's response emotion j and model reaction time t_i^{model} then averaged t_i^{model} over all networks making the same response to the stimulus. The quantitative comparison is the linear fit between network predictions and human RTs for all stimuli and response pairs for which both human and network data were available. Young et al.'s criterion for reporting t_{ij}^{human} was that at least 25% of the human subjects responded with emotion j to stimulus i, and for some of these cases, none of the networks responded with emotion j to stimulus i, so the missing human and network data were disregarded.

Model Discrimination Scores

We assume discrimination is more difficult the more similar two stimuli are at some level of processing. We use the same measure of similarity as in the MDS procedure: The correlation r_{ij} between the network's representation of stimuli i and j (either at the pixel, perceptual, gestalt, or output level). To convert similarity scores to discrimination scores, we used the transform $d_{ij} = 1 - r_{ij}$. This was measured for each of the 30 pairs of J. J. images for which human data were available. The discrimination scores were then averaged over the 13 networks that had not been trained on J. J. and compared to the human data.

Mixed-in Expression Detection

To measure the ability of the model to detect the secondary expression mixed into a morph stimulus, we followed Young et al.'s (1997) methods. For each network's output on a given stimulus, we scored the first, second, and third highest outputs of the networks

as a 3, 2, and 1, respectively, and assigned the score 0 to the three remaining outputs. For each morph and prototype stimulus, we averaged these score vectors across all 13 networks. Now for each of the 30 possible combinations of near prototype (i) and far prototype (j), using the 90%, 70%, and 50% morphs moving from expression i to expression j, we subtracted the score vector for prototype i from the score for each of the three morphs. This eliminates the effect of the intrinsic similarity between J. J.'s prototype expressions. Now, averaging these score vectors across all 30 sequences, we obtain the data plotted in Figure 8.

Acknowledgments

We thank Stevan Harnad, Bill Kristan, Paul Munro, Alice O'Toole, Terry Sejnowski, and Gary's Unbelievable Research Unit (GURU) for helpful comments on previous versions of this manuscript. We also thank Andrew Young for permission to use the human data plotted in Figures 2, 3, 7a, and 8. We are also grateful to Paul Ekman for providing us with the Pictures of Facial Affect. This research was funded by NIMH grant MH57075 to GWC.

Reprint requests should be sent to Garrison W. Cottrell, UCSD Computer Science and Engineering, 9500 Gilman Drive, La Jolla, CA 92093-0114, USA, or via e-mail: gary@cs.ucsd.edu.

Notes

1. For comparison, the order of difficulty for the Calder et al. (2001) PCA model is happiness (98%), fear (97%), surprise (92%), anger (86%), sadness (72%), and disgust (72%).
2. We used 13 networks for technical reasons (see Methods for details).
3. The same comparison of our model with the confusion matrix from the Calder et al. (2001) forced-choice experiment is slightly better, .686, whereas their PCA model's correlation with their subjects' confusion data is somewhat lower, .496.

REFERENCES

Beale, J., & Keil, F. (1995). Categorical effects in the perception of faces. *Cognition, 57,* 217–239.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7,* 1129–1159.

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

Borg, I., & Lingoes, J. (1987). *Multidimensional similarity structure analysis.* New York: Springer.

Calder, A., Young, A., Perrett, D., Etcoff, N., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3,* 81–117.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research, 41,* 1179–1208.

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 526–551.

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

Cottrell, G. W., Dailey, M. N., Padgett, C., & Adolphs, R. (2000). Is all face processing holistic? The view from UCSD. In M. Wenger & J. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Context and challenges* (pp. 347–395). Mahwah, NJ: Erlbaum.

Cottrell, G. W., & Metcalfe, J. (1991). EMPATH: Face, gender and emotion recognition using holons. In R. P. Lippman, J. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems 3* (pp. 564–571). San Mateo: Morgan Kaufmann.

Cottrell, G. W., Munro, P., & Zipser, D. (1989). Image compression by back propagation: An example of extensional programming. In N. E. Sharkey (Ed.), *Models of cognition: A review of cognitive sciences.* New Jersey: Norwood.

Craw, I., & Cameron, P. (1991). Parameterising images for recognition and reconstruction. In *Proceedings of the British Machine Vision Conference* (pp. 367–370). Berlin: Springer.

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

Donato, G., Barlett, M. S., Hager, J. C. Ekman, P., & Sejnowsky, T. J. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21,* 974–989.

Ekman, P. (1999). Basic emotions. In T. Dagleish & M. Power (Eds.), *Handbook of cognition and emotion* (chap. 3, pp. 45–60). New York: Wiley.

Ekman, P., & Friesen, W. (1976). *Pictures of facial affect.* Palo Alto, CA: Consulting Psychologist Press.

Ellison, J. W., & Massaro, D. W. (1997). Featural evaluation, integration, and judgement of facial affect. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 213–226.

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

Goldstone, R. L. (2000). A neural network model of concept-influenced segmentation. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representation through category learning. *Cognition, 78,* 27–43.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (1998). A comparison of two computer-based face recognition systems with human perception of faces. *Vision Research, 38,* 2277–2288.

Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory and Cognition, 24,* 26–40.

Harnad, S. R. (1987). *Categorial perception: The groundwork of cognition.* Cambridge: Cambridge University Press.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. *Parallel distributed processing: Explorations in the microstructure of cognition* (vol. 1, chap. 3, pp. 77–109). Cambridge: MIT Press.

Jansari, A., Tranel, D., & Adolphs, R. (2000). A valence-specific lateral bias for discriminating emotional facial expressions in free field. *Cognition and Emotion, 14,* 341–353.

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

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., & Konen, W. (1993). Distortion

invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42,* 300–311.
Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401,* 788–791.
Lee, T.-W., Lewicki, M. S., & Sejnowski, T. J. (2000). ICA mixture models for unsupervised classification of non-Gaussian sources and automatic context switching in blind signal separation. *IEEE Transactions on Pattern Recognition and Machine Intelligence, 22,* 1–12.
Lien, J. J.-J., Kanade, T., Cohn, J. F., & Li, C.-C. (2000). Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems, 31,* 131–146.
Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19,* 577–621.
Lyons, M. J., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. *Proceedings of the third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 200–205). Los Alamitos, CA: IEEE Computer Society.
Lyons, M. J., Budynek, J., & Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21,* 1357–1362.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.
Matsumoto, D. (1992). American–Japanese cultural differences in the recognition of universal facial expressions. *Journal of Cross-Cultural Psychology, 23,* 72–84.
Padgett, C., Cottrell, G., & Adolphs, R. (1996). Categorical perception in facial emotion classification. *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.
Padgett, C., & Cottrell, G. W. (1998). A simple neural network models categorical perception of facial expressions. *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 806–807). Mahwah, NJ: Erlbaum.
Perrett, D., Hietanen, J., Oram, M., & Benson, P. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London, B: Biological Sciences, 335,* 23–30.
Pevtzow, R., & Harnad, S. R. (1997). Warping similarity space in category learning by human subjects: The role of task difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Proceedings of SimCat 1997: Interdisciplinary workshop on similarity and categorization* (pp. 189–195). Edinburgh, Scotland: Department of Artificial Intelligence, Edinburgh University.
Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3,* 1199–1204.
Rosenblum, M., Yacoob, Y., & Davis, L. S. (1996). Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks, 7,* 1121–1138.
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161–1178.
Russell, J. A., & Bullock, M. (1986). Fuzzy concepts and the perception of emotion in facial expressions. *Social Cognition, 4,* 309–341.
Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of circumplex model of affect. *Journal of Personality and Social Psychology, 57,* 848–856.
Sanger, T. (1989). An optimality principle for unsupervised learning. In D. Touretzky (Ed.), *Advances in neural information processing systems* (vol. 1, pp. 11–19). San Mateo: Morgan Kaufmann.
Schiano, D. J., Ehrlich, S. M., Sheridan, K. M., Beck, D. M. (2000). *Evidence for continuous rather than categorical perception of facial affect.* Abstract presented at The 40th Annual Meeting of the Psychonomic Society.
Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44,* 229–237.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.
Tijsseling, A., & Harnad, S. R. (1997). Warping similarity space in category learning by backprop nets. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Proceedings of SimCat 1997: Interdisciplinary workshop on similarity and categorization* (pp. 263–269). Edinburgh, Scotland: Department of Artificial Intelligence, Edinburgh University.
Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3,* 71–86.
Vapnik, V. N. (1995). *The nature of statistical learning theory.* New York: Springer.
Woodworth, R. S. (1938). *Experimental psychology.* New York: Holt.
Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition, 63,* 271–313.