

# Is *All* Face Processing Holistic?

## The View from UCSD

Garrison W. Cottrell ([gary@cs.ucsd.edu](mailto:gary@cs.ucsd.edu))

Matthew N. Dailey ([mdailey@cs.ucsd.edu](mailto:mdailey@cs.ucsd.edu))

Computer Science and Engineering Department

University of California, San Diego

9500 Gilman Dr., La Jolla CA 92093-0114 USA

Curtis Padgett ([cpadgett@cs.ucsd.edu](mailto:cpadgett@cs.ucsd.edu))

Jet Propulsion Laboratory

National Aeronautical and Space Administration

California Institute of Technology Pasadena CA 91109 USA

Ralph Adolphs ([ralph-adolphs@uiowa.edu](mailto:ralph-adolphs@uiowa.edu))

Department of Neurology

University of Iowa

Iowa City, IA 52242 USA

Running Head: Is *all* face processing holistic?

Preprint version of a chapter to appear in M. Wenger and J. Townsend (Eds.), *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*. Erlbaum.

## **Abstract**

There has been a great deal of progress, as well as controversy, in understanding how complex objects, in particular, human faces, are processed by the cortex. At the same time, sophisticated neural network models have been developed that do many of the same tasks required by these cortical areas. Such simplifying models allow us to explore hypotheses concerning relatively complex domains such as face processing.

In this chapter, we give a somewhat idiosyncratic history of the development of neural network models of face processing, concentrating on work at UCSD, and show how these models have led to a novel hypothesis concerning processing of facial expression. While our models have suggested a role for holistic representations of faces in identification, *general* local features appear to be important in recognition of expression.

## Introduction

What are the features underlying face processing? This is a question that appears to have been asked and answered multiple times. The most recent answer appears to be that the features used in recognizing faces are “holistic” in nature (Farah et al., 1998; Biederman and Kalocsai, 1998). Generally, holistic is used to refer to a system that is sensitive to *configural* properties of the stimulus, and which displays context-dependent interpretation of the parts (Farah et al., 1998). Both Farah and Biederman consider this to be characteristic of face recognition versus visual object recognition, where the evidence suggests that “parts” of the object are processed more independently, and changes in configurations of the parts have less impact on recognition (see also discussions in Uttal, 1988, and in chapters by Uttal and Wenger & Townsend, this volume).

One approach to getting insight into these questions is to use computational pattern recognition models to discover which features are best for those models. The reasoning is inductive: if certain features prove superior for these models, then perhaps they are useful for brains as well. Negative (inductive) inferences may be drawn also: if one finds that there is *no* pattern recognition model (in the current stable of such models) that can effectively use a certain kind of visual feature, then one might be confident in predicting that the brain does not use such a feature either.

On the other hand, if certain features prove effective for multiple pattern recognition models — perhaps some kind of “universally efficacious” feature — then one might be confident in predicting that the brain does use such a feature. Unfortunately, no such universal features have been found. However, the wide variety of features in use today have proven useful in many situations. While we cannot possibly provide a good review of all of these in this chapter (see the chapter by O’Toole, Wenger & Townsend, this volume), we devote the next section to a modest framework for discussing features and recent successful (engineering-oriented) computational systems for face recognition.

In contrast to engineering-oriented approaches, our approach involves using pattern recognition models that are also supposed to be cognitive models. These models are supposed to provide a basis for the inductive leaps from model to human processing described above.

The degree one believes such inductive leaps from a cognitive model to human information processing depends on several things, such as (a) the biological plausibility of the model, (b) the extent to which the model actually performs the same task as human subjects, (c) the correlation between measures taken on human subjects and corresponding measures taken from the model, (d) the extent to which the model provides novel insights into the nature of the processing required for the task and (e) novel predictions extracted from the model that turn out to be correct.

In our work, we tend to rely on neural network models to provide us with some degree of biological plausibility. While the representation of “neurons” in such models is cartoon-like at best, one hopes that at the level of *network computation* we are at least within range of reality. We also make the effort to use the same stimuli for our models as are used in human experiments, and to train our models to perform the same task. While the particular training procedure we use (backpropagation) is not particularly biologically plausible, we consider it to be an efficient search technique for finding networks that perform the task. More biologically plausible learning schemes exist, but they tend to be extremely slow to converge, while often achieving basically the same end result (Movellan and McClelland, 1993; Plaut, 1991).

In this chapter, we first consider what some of the possible dimensions of “feature space” are, and illustrate these with a discussion of the features used in some recent engineering-oriented face recognition models. In doing so, we hope to clarify to some extent what kinds of features and systems might result in “holistic” processing. On the other hand, we explicitly avoid definition of that term, as we believe that, like other terms such as “consciousness,” it may not be a coherent category. We then review a sequence of neural network models of face processing that were developed at UCSD (see also processing models developed by Wenger & Townsend, this volume). We begin with a relatively simple model of face recognition that provided a basis for understanding holistic processing. We then turn to models of expression recognition that are successful in some of the ways suggested above: First, they work reasonably well on stimuli that are very similar to those presented to human subjects. Second, they demonstrate considerable agreement with several response variables in facial expression recognition. Finally, they provide insight into the kinds of features that may

prove most efficacious for this task.

Farah and Biederman's positions raise the interesting question as to whether holistic face processing is *mandatory*, or whether different face processing tasks require different types of processing. The double dissociation between facial identification and facial expression identification in brain damaged patients provides some evidence that the two tasks rely on different representations and/or processing mechanisms. Our work showing that local (non-holistic) features are efficacious for expression identification provides a possible explanation for the double dissociation and, if holistic representations are used in other face processing tasks, indicates that there may be multiple representations of faces in cortex. In any case, it is important to be specific about the task when discussing representational issues.

## Features

In this section we consider some of the dimensions of feature space. The space of possible features is huge, and while the set useful for face processing is probably somewhat smaller, it must still be large. In this necessarily abbreviated review (again, also see O'Toole, Wenger & Townsend's chapter, this volume), we will consider just three dimensions of this space (see Figure 1) that have proved useful in practical face identification systems. We should also be clear here that when we use the word "features," most of the time we are referring to what may be termed *first order features*, that is, features that are computed relatively directly from the image. This is to be contrasted with *second order features*, that is, features that are computed from the first order features. For example, traditional features such as measurements of the distance between the eyes are necessarily second order features, as first order features such as the locations of the eyes (perhaps computed by template matching) would have to be computed first.

The first consideration is the *spatial extent* of the features relative to the object of interest. These may be divided into global features, spanning roughly the image of the whole object at one extreme, and local features, spanning only small subregions of the image subtended by the object. A simple example is a system that uses stored images of the whole face as a feature. This is clearly a global feature. Systems that store images of a person's nose, eyes,

and mouth would be using local features. Both of these are examples of *template matching* approaches.

The second dimension we consider is whether the features are rigidly applied to a particular region of the image (e.g. Cottrell and Fleming, 1990; Fleming and Cottrell, 1990; Turk and Pentland, 1991), or are allowed to deform or move with respect to one another and the image being processed (e.g. Yuille, 1991). As a toy example, consider the idea given above of a system that stored images of face parts, such as eyes, mouths and noses. They might be stored with the expected relative distances between them (a second order feature). Allowing the individual parts of this template to move to the best match on the face is an example of a deformable template approach. This can be more flexible in terms of matching faces slightly rotated out of perfectly frontal views, which, in the image, will make the eyes relatively closer together. We call this the “rigid/deformable” distinction. Deformable templates have proved the useful in recent applications of computer techniques to face recognition (Wiskott et al., 1997). A recent innovation has been in the use of a probabilistic characterization of *how* the templates or features deform to characterize within-subject versus between-subject distinctions (Moghaddam et al., 1996).

Finally, with the third dimension, we would like to draw attention to the issue of whether the features are learned from examples in the domain of interest, that is, derived from the data itself, usually in the service of the task, or are pre-determined. We call this the “problem-specific/untuned” distinction. Gabor filters (defined below), for example, are not tuned to the task, although researchers have (over the years) tuned the exact form of the filters by manipulating free parameters such as scale, orientation, and spatial frequency (e.g. Buhmann et al., 1990; Wiskott et al., 1997). On the other hand, eigenfeatures (e.g. Padgett and Cottrell, 1997; Pentland et al., 1994), independent components (e.g. Bartlett and Sejnowski, 1998) and intensity surfaces (e.g. Nastar and Pentland, 1995; Moghaddam et al., 1996) are learned from the data (again, most of these feature types are defined below).

These three dimensions ignore many possible distinctions, such as 2-D (or view-based) vs. 3-D (or volume-based), static or temporal, and active versus passive sensing. However, most of the most successful approaches to face recognition to date use static, view-based, passively sensed features (Moghaddam et al., 1996; Okada et al., 1998; Wiskott et al., 1997).

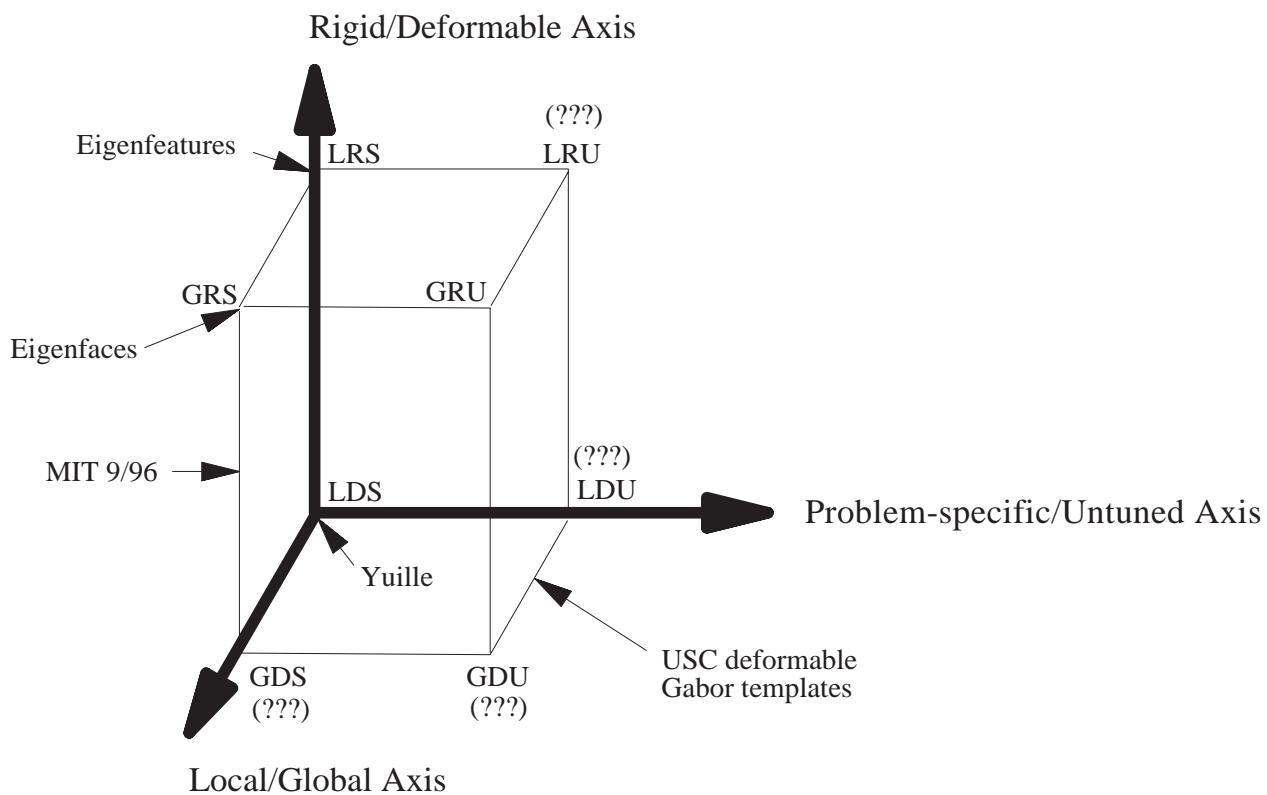


Figure 1: The Necker Cube of feature space: L/G: local/global; R/D: Rigid/Deformable; S/U: problem-specific/untuned. Examples of systems are placed near their regions in the space.

While these distinctions are important, and may provide the basis for even better approaches to face recognition than the current state of the art, for now, we will focus our attention upon the ones outlined.

Now, let us consider some of the features actually used in recent face recognition systems to illustrate some of the points in the feature space outlined above. The shortest step from using actual face images as templates (or features) is to use eigenfaces. These are features that are computed from the covariance matrix of the face images used as a training set. That is, face image  $i$  is treated as a vector, by simply concatenating the rows of the image into one long vector,  $\vec{F}_i$ . The mean face vector  $\vec{\mu}$  is computed, and then the covariance matrix is constructed:

$$\mathbf{C} = \sum_i (\vec{F}_i - \vec{\mu})(\vec{F}_i - \vec{\mu})^T \quad (1)$$

The eigenvectors of this matrix are called the *principal components* of the data.<sup>1</sup> These components amount to a rotated coordinate system for the data, centered on the mean of the data. They are typically ordered in terms of the amount of variance accounted for (which corresponds to the eigenvalues of each eigenvector). It turns out that when the data is projected onto the top  $k$  components (rather than all of them) and then reconstructed by mapping back to the original dimensions, the reconstruction is optimal in the least squared error sense for  $k$  components. This has uses in data compression, but it is also useful in throwing away variance that may be noise. When this technique is used with face images, the resulting eigenvectors are called “eigenfaces.” An example set of eigenfaces are shown in Figure 6. When the identical techniques are applied to images of eyes, mouths and noses separately, they are called “eigenfeatures,” or more specifically, “eigeneyes,” “eigenmouths,” etc. Examples of eigeneyes are shown in Figure 8.

Both of these kinds of features can be used for face identification systems.<sup>2</sup> Given a set of

---

<sup>1</sup>Some researchers call the eigenvectors of the covariance matrix the “principal component eigenvectors,” and reserve the word “principal components” for the values of the projections of the data onto these eigenvectors.

<sup>2</sup>In the literature, these are often called “face recognition systems,” but this should not be confused with the “old/new” task used in psychological studies of face recognition. See also the discussion in the O’Toole,

training images, the eigenfaces or eigenfeatures are computed from carefully aligned images. Then each face can be represented as the vector of numbers resulting from its projection onto the first  $k$  principal components ( $k$  can be chosen to maximize performance on a subset of the training data held out to use for this purpose). This gives a vector of  $k$  numbers representing the face. Then, given a new image, it is also first aligned by the same procedure as the original images, projected onto the same  $k$  components, and then a match score is computed via some distance or similarity measure between its  $k$ -dimensional representation and that of the stored images. The best match is the presumed “identity” of the new face.

Placing these features in our three-dimensional feature space, both of them are “learned” from the data, as they inherently depend on the statistics of the training set. Both of them are rigidly applied, as both the training and new images must be subject to careful alignment procedures. Finally, eigenfaces are clearly global features, while eigenfeatures are local with respect to the whole face. Pentland and his colleagues (Pentland et al., 1994) have shown that these two kinds of features perform about equally well for face identification systems. Systems that use eigenfaces are now used as a baseline for comparison with new approaches.

A second popular feature for face identification is the 2-D Gabor wavelet filter (Daugman, 1985). A Gabor filter is a two-dimensional sinusoid localized by a Gaussian envelope; it can be tuned to a particular orientation and spatial frequency. Examples are shown in Figure 2. One way of defining the kernel function is:

$$G(\vec{k}, \vec{x}) = \exp(i\vec{k} \cdot \vec{x}) \exp\left(-\frac{k^2 \vec{x} \cdot \vec{x}}{2\sigma^2}\right),$$

where

$$\vec{k} = [k \cos \phi, k \sin \phi]^T$$

and  $k \equiv |\vec{k}|$  controls the spatial frequency (scale) of the filter function  $G$ ,  $\vec{x}$  is a point in the plane relative the wavelet’s origin,  $\phi$  is the angular orientation of the filter, and  $\sigma$  is a constant. This filter is biologically motivated — it is a good model of observed receptive fields of simple cells in cat striate cortex (Jones and Palmer, 1987). von der Malsburg and colleagues form a “jet” by extracting the response of several filters with different orientation and spatial frequency tunings from a single point in the image. As an image feature detector,

the jet exhibits some invariance to background, translation, distortion, and size (Buhmann et al., 1990b), and provides a good basis for face identification systems (Lades et al., 1993; Wiskott et al., 1997).

As an example, the face identification systems developed at USC and Bochum by von der Malsburg and colleagues use Gabor jets. The original approach assumed a fixed rectangular grid of feature locations on the images (Lades et al., 1993). A Gabor jet was extracted at each location, using five scales and eight orientations, and both sine and cosine waves within the Gaussian envelope. This give rise to 80 numbers at every location. The magnitude of the filter response is recorded (that is, square root of sum of the squares of the sine and cosine responses, which reduces the set of numbers to 40). Now, given a new face, the same grid is placed over the face. However, the grid is allowed to *deform* by moving the feature points according to how well they match each stored face. The model assumes the grid points are connected by “springs,” but are “attracted” to the best fitting nearby point (for each face). This is computationally intensive (because it has to be done for each face in the database), but it results in a match that is much less sensitive to, for example, moderate rotations of the face left or right.

Gabor filters are an example of an untuned feature, since they are not learned from the data but predefined. They are also clearly deformable in the way they are commonly used. However, they span the local/global axis, to the extent that the larger scale components of the jet subtend a large fraction of the face.

To give an example of a recently developed feature that goes beyond our simple classification scheme given above, the September 1996 MIT system (Phillips et al., 1997)<sup>3</sup> is a hybrid system. It uses eigenfaces to extract the top matches, followed by a ranking of these via a more computationally intensive algorithm. The refinement algorithm uses the idea of intensity surfaces, which are face images treated as a 3-D surface, with the intensity values forming the third dimension. Examples of several intensity surfaces are shown in Figure 3.

---

<sup>3</sup>The designation of this system as the “September 1996 MIT system” is due to its participation in the Army’s Face Recognition Technology (FERET) program. This program is intended to quantify how well face identification technology has developed by having face identification “contests” on large galleries of faces. The MIT (usually eigenfaces) and USC systems (Gabor jets) have been consistently high-performing competitors.



Figure 2: A set of Gabor filters of five different scales and orientations.

While these first order features fit into our classification scheme (they are global, learned and rigid), the use of them is not. The system learns a probabilistic characterization of how the intensity surface deforms between two images of the same person (for example, between the left and right surfaces in Figure 3). It also learns how different subjects' intensity surfaces deform when mapped to one another (for example, between the top and bottom surfaces in Figure 3). Using an eigenspace decomposition of these two transformations, the system compares the transformations of the top matches to the probe face. The similarity of these *transformations* to the within-person deformation is used to rank the images (Moghaddam et al., 1996). This is a third order feature: The transformations between intensity surfaces are a second order feature (a relationship between the first order features, the intensity surfaces). The similarity of these transformations is then a third order feature.

The above examples have been taken from face identification systems developed in the computer science community. While these systems have not generally been put forth as cognitive models (but see Phillips et al. 1999), as mentioned above, we believe it is still important to consider what kinds of features are used in real systems, as these may be useful intuition pumps for considering how the human brain may perform the task. Again, if these systems did not perform well, then the insights to be gained would be dubious. Hence it is of interest to know just how well current systems are doing at this task. The face identification task (matching a new image with a face from a known gallery of faces) can be performed quite well when the novel images are taken in the same session, with minor variations in expression. Performance is over 95% on galleries of over 800 such images (Phillips et al., 1998), when measured in terms of the algorithm ranking the correct image within the top 10.<sup>4</sup> Deciding which algorithm is “best” often depends on which test one looks at. All of the algorithms appear to be constantly improving even though the gallery size is increasing, but some are better at some kinds of tests than others. The hardest tests administered as part of this program involve images taken over a year apart using a different camera and lighting (somewhat ironically termed “duplicates”), and images where the probe face is rotated away

---

<sup>4</sup>These results are based on a large database of images developed as part of the Department of Defense's Face Recognition Technology (FERET) program. The datasets have become more difficult over the years, as the FERET competitions uncover which variants make face identification difficult.

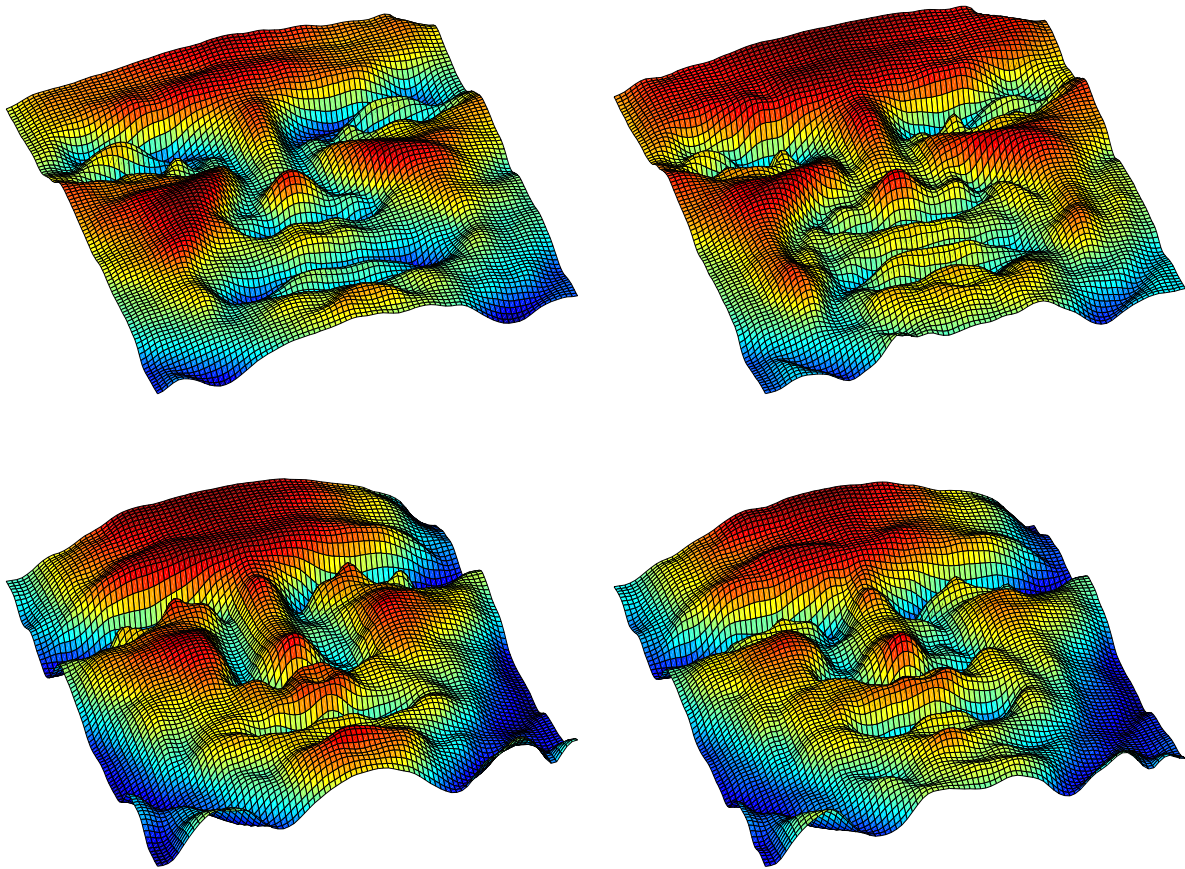


Figure 3: Examples of intensity surfaces. Rows are same face, columns are same expression.

from the camera. The top systems perform about equally well at the frontal view, same day task. However, the systems using deformable templates based on Gabor filters appear to do somewhat better with faces slightly turned away from the camera.

Finally, we would like to point out that our review of these systems suggests that it is important to consider the entire face identification system when thinking about possible measures that may resolve the holistic/part-based question. For example, one can easily build a face recognition system from Gabor jets that are restricted to the smaller scales. Thus, each feature would correspond to local features. However, these local features would be connected in a grid, and matching to a new face includes deforming that grid to find the best fit. Since the grid “resists” deformation via the spring analogy, there are positional relationships between the features that affect the matching process. This is clearly configural. Hence a system that starts with first order local features can end up looking “holistic,” by recent definitions of the term.

Finally, we should note that the particular face processing task being performed will also impact the kind of representations that are best for the task. In what follows, we show that there are differences in efficacy (given a particular kind of classifier) between different feature types for the task of facial expression identification.

## A sequence of models

Here, we describe a sequence of neural network models of face processing developed at UCSD. We shall see that while global, problem-specific, rigid features “work” on the small databases we tested them on (and on larger databases by extrapolating from the early FERET tests of the MIT eigenfaces system), and they provide an intuitive basis for understanding holistic processing, they are not ideal for all face processing tasks, even on small databases.

## Face recognition via holons

In a series of papers (Cottrell and Fleming, 1990; Fleming and Cottrell, 1990; Cottrell, 1990), we described a model of face recognition that used an *autoencoder* as a preprocessor (Ackley et al., 1985; Rumelhart et al., 1986a; Rumelhart et al., 1986b). The basic architecture of

the preprocessor is quite simple (see Figure 4). It consists of a input layer of units, one for each pixel in the image, a smaller hidden layer, and a layer of outputs equal in size to the input layer. The network is trained by backpropagation to replicate the input pattern on the output. This identity mapping task is interesting for several reasons. First, by replicating the input through a narrow channel of hidden units, the network must extract regularities in the input patterns at the hidden unit layer. Second, since the output is the same as the input, the network can be considered to be *self-supervised*. Thus, we have a system that extracts regularities from its environment in a self-organizing way. By simply identifying the output and input layers (that is, by “folding over” the network), it is easy to imagine this as a system that extracts *features* from its environment that preserve the most information, in the mean-squared error sense, as this is the error minimized by standard back propagation.

This error criterion is important, because it leads to a particular *kind* of feature being extracted by the network: the principal components of the data (Baldi and Hornik, 1989; Cottrell and Munro, 1988)<sup>5</sup>. In this case, the principal components of facial data resemble ghostly-looking “faces” we called *holons* (later called “eigenfaces” by Turk and Pentland – again, see Figure 6 for examples of eigenfaces), and we suggested that they may provide a computational basis for understanding single-cell recordings from so-called “face cells.”

In Cottrell and Fleming (1990), we specified the term “holon” to apply more generally to any representational element “if its receptive field subtends the whole object whose representation it is participating in. Further, we want[ed] to require that the information in a set of holons in the ideal case be maximally distributed: i.e., the entropy of any unit is maximized. The latter restriction eliminates grandmother cells, insures that the representation be noise resistant, and also distributes the processing load evenly... A weak point of our definition is the difficulty of defining precisely the notion of a ‘whole object’.”<sup>6</sup>

---

<sup>5</sup>We should make clear that the principal components are extracted *only* by autoencoders. Neural networks trained to perform classification tasks do not, in general, extract principal components. Also, the hidden unit weight vectors do not actually line up with the principal components; rather, they span the principal subspace.

<sup>6</sup>This idea has been better formalized by the recent work on Independent Component Analysis. There, the correct notion is the maximization of the joint entropy, which captures the idea of maximal distribution of information.

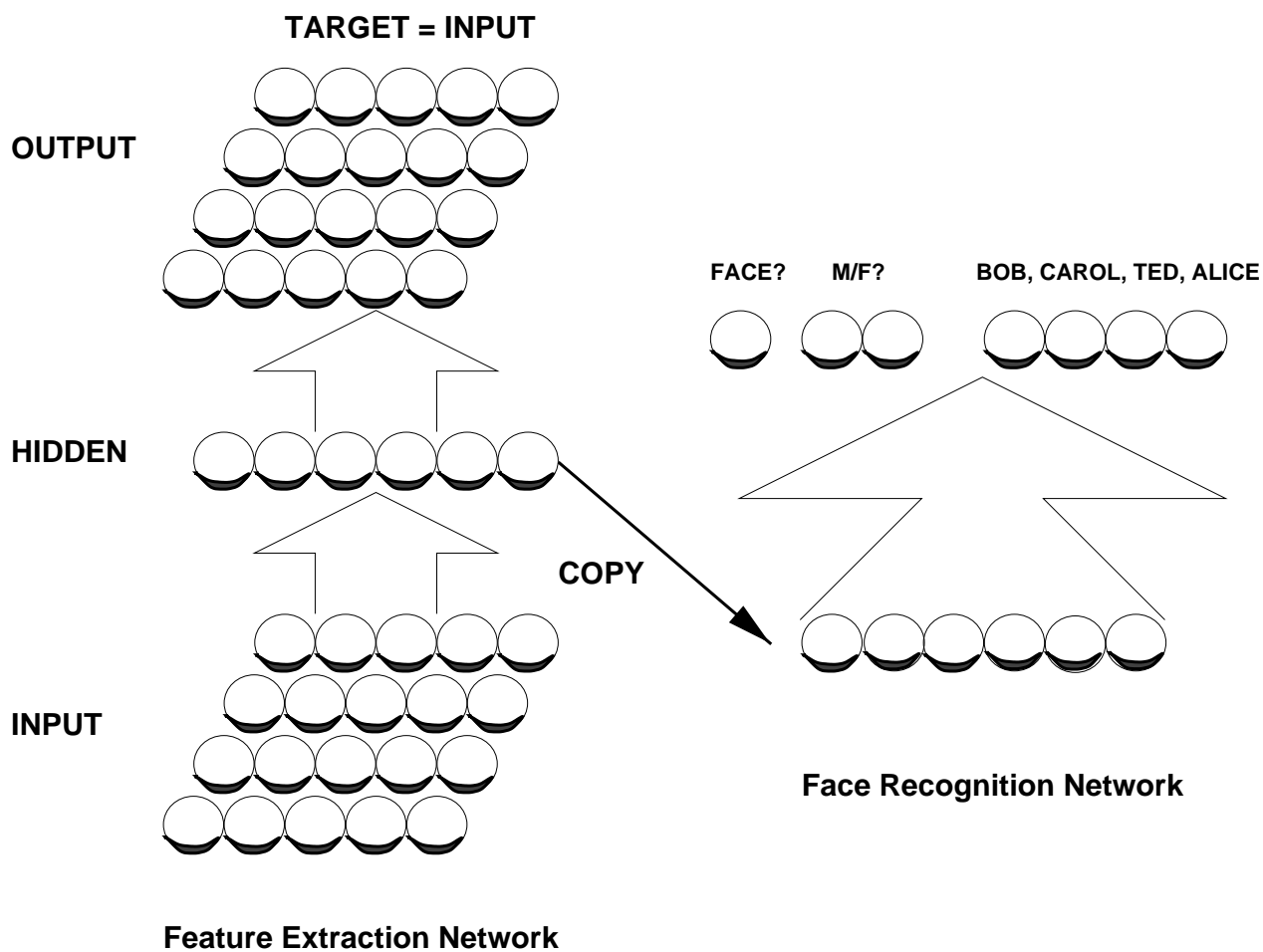


Figure 4: The face recognition system of Cottrell & Fleming. The left side shows the autoencoder network that extracts features from faces. Faces are represented as gray scale values scaled to the 0-1 range. The right side shows the face recognition network, which uses the hidden layer representations as input.

The intuition gained by analyzing this system was that the holistic templates developed by our system were *not* grandmother cells. All of the faces presented to the system activated multiple hidden units. As another check, we drove the recognition network by individual hidden units. All of the cells activated multiple name units at the output, and these usually included both males and females. We can think of these, then, as *distributed templates* that participate in representing all of the individuals in the training set. While the input to our system was not representative of the inputs to so-called “face cells” in inferior temporal cortex, the network provides an intuition pump for understanding how cells could be responding quite strongly to faces, yet not specifically to any one face. These units show how face cells can provide a distributed representation of individuals. At the time, this computational example of features that spanned the whole object (what we have called global features here) and yet did not represent any one object seemed as good an example of “holistic features” as any.

These systems demonstrated that holistic features are useful for face recognition. However, the issue is not settled. In other relatively recent work, Pentland and colleagues (Pentland et al., 1994) have shown that locally-derived features, eigeneyes and eigenmouths) did just about as well as eigenfaces for face recognition purposes on a large database. This kind of evidence suggests that either kind of feature is good enough, which, unfortunately, leaves us where we started.

It is worthwhile to repeat that one must be cautious, and not restrict the application of the term “holistic” to global features. As mentioned above, some of the most successful recent systems for face recognition have used a representation that is not based on global features, yet the system may still be termed holistic. In particular, the von der Malsburg group’s system used Gabor filters tied together by an elastic network, which was “stretched” when finding the best match of the filters to a novel face. Thus there is relational, second order information in the application of this filter-matching process to a face. In other words, there is *configural* information used in the matching process, which is one of the definitions of holistic as the term is used by Farah.

Finally, a hybrid system that includes both local and global features may be considered. Systems have been demonstrated (Bartlett et al., 1996) that usefully combine the results of

local and global classifiers to achieve better results than any of the component classifiers. This suggests that one must be careful in interpreting the results of experiments that show degradation in tasks that presumably use only part-based representations. This could be a degradation in a system that uses both kinds of features. Farah and colleagues are careful to state (parenthetically) that their hypothesis concerning holistic representations may simply mean that part-based features are relatively less important in face processing. This is a weak form of the hypothesis that, in the context of a hybrid system, would be suggesting that the weighting on a component of the system using part-based features is simply less than the weighting on a component of the system using more holistic features.

However, it is still instructive to know what features are most efficacious for which tasks. In particular, our results suggest that local, untuned features are best for expression recognition. This is the subject of the remainder of the chapter.

## EMPATH

The typical task used in face processing experiments that assess the holistic versus part-based question is that of face recognition and face identification. *Face recognition* may be defined by the typical old/new judgement. That is, was this face in the study set? This task is often used to suggest how we recognize faces as *familiar*. On the other hand, face identification is the task of saying *who* this face is (cf. the O’Toole et al. chapter, this volume).

The question naturally arises as to whether all face processing tasks are best served by holistic features. In order to test this question, Cottrell & Metcalfe embarked on a project to apply the above face identification system to facial emotion recognition. We gathered facial images of undergraduate psychology students at UCSD, whom we asked to “look happy,” “look sad,” “look bored,” etc. During this process, except for one student who also worked as a mime, we noticed that the subjects were much better at portraying positive emotions than negative ones, and that, in general, the variance in their expressions was not high.

The expression recognition system was identical to the face recognition system described earlier (see Figure 4), except that in addition to identity and gender outputs, the classifier network had emotion labels. The resulting network, dubbed EMPATH (for EMotion PATtern recognition using Holons), was a dismal failure (Cottrell and Metcalfe, 1991). The

system reliably identified the individuals and correctly learned their gender. However, when it came to emotion labels, the system could not even learn the training set. Consistent with the observations made during image capture, the network was able to perform reasonably well on the positive emotions. On the other hand, the confusion matrix for the system, when compared to human subjects trying to classify the same images, showed a pattern inconsistent with the human data, with the network confusing positive and negative emotion portrayals.

There are two possible conclusions one can make from this. First, we may have had bad data, as we observed that psychology sophomores from UCSD were poor at feigning emotions. Many studies of facial expression recognition suffer from feigned emotion portrayals. For example, when asked to “look happy,” many subjects display a so-called “cognitive smile,” which does not engage the muscles around the eyes. Second, we may have been using features that were inappropriate for the task. Emotion recognition is a quite different task than identification. Recognizing a person’s expression may require *ignoring* their identity. Perhaps features appropriate for identity are particularly bad for expression recognition. Additional evidence that the processes of face identification and emotion recognition may be independent may be gleaned from studies that show that (a) prosopagnosics are capable of recognizing facial expression in photographs, without knowing who it is (or even that successive photographs are of the same person) (Tranel et al., 1988); and (b) that brain-damaged subjects who show difficulty identifying certain facial expressions will nevertheless be able to discriminate identity (Adolphs et al., 1994; Adolphs et al., 1995).

## EMPATH-II

In more recent work, we have shown that both of the above conclusions were correct. We developed a relatively successful expression recognition system<sup>7</sup> using the Ekman and Friesen Pictures of Facial Affect (POFA) database (Ekman and Friesen, 1976). The initial experiment, described below, resolves to some extent the question of “which features are good for expression recognition, global or local?” It also examines the question of whether features

---

<sup>7</sup>Here, we dub this system “EMPATH-II,” although for reasons that will become obvious, the expansion of the acronym no longer applies.

tuned to the data are good for the task or whether nonspecific features are superior. The following work represents a recent replication of (Padgett and Cottrell, 1997; Padgett, 1998), with minor changes in methodology.

A crucial resource for this project was the POFA database. This database was developed using actors trained in the Facial Action Coding System (FACS). The FACS system was developed by Ekman and colleagues to provide an objective standard for both producing and coding pictures of facial affect. It involves a set of 46 facial actions, using muscles and muscle groups that can be moved independently. In the database, by combining particular facial actions, the actors portrayed one of six emotions: happiness, sadness, fear, anger, surprise, or disgust. There are also images in the database corresponding to no facial actions, otherwise known as “neutral” faces. The database has been verified via testing on undergraduates, and only those images that over 70% of subjects agree are the attested emotion are included in the dataset.

The database comes in the form of 35mm slides. We scanned the slides and normalized the digitized images. For each face, we manually located three points (the center of the left eye, the center of the right eye, and the bottom of the top row of teeth) and linearly warped these points to standard locations by rotation, scaling, and cropping. As a final step before any processing or feature extraction, we normalized each image’s pixel values to the same mean and standard deviation. Figure 5 shows several example faces.<sup>8</sup>

A goal of this work was to compare the efficacy of different kinds of features, including eigenfaces, for emotion recognition (Padgett & Cottrell, 1995, 1997). For our study, we used 14 individuals (6 male, 8 female). The database contains multiple examples of some emotions from the same individual, and for some individuals, not all emotions are in the database. Since the dataset size was small, and neural networks require large amounts of data to generalize well, we used all applicable examples (slightly biasing our training set). This resulted in 18 happy faces, 17 sad, 17 angry, 15 afraid, 15 disgusted and 14 surprised, for a total of 96 images.

---

<sup>8</sup>Dr. Ekman has graciously allowed us to share this digitized version with other researchers for a reduced fee compared to the original database. Please send email to [gary@cs.ucsd.edu](mailto:gary@cs.ucsd.edu) for further information.



Figure 5: Examples from the Pictures of Facial Affect database normalized and cropped. These are images of “JJ” portraying Anger, Disgust, Neutral, Surprise, Happiness (twice), Fear, and Sadness.

## Feature sets

We examined three types of features for expression recognition: eigenfaces, eigenfeatures (PCA of eye and mouth regions), and eigenvectors of random  $64 \times 64$  pixel patches from the database (this technique is sometimes known as local PCA). All three representations involve projection of all or part of the face image onto a set of basis vectors computed with principal components analysis (PCA). A vector of the resulting projection coefficients can then be fed into a classifier such as a neural network ensemble (described below).

- **Eigenfaces:** The first feature type, the eigenface, is an eigenvector of the covariance matrix of the face image set. The vectors are ordered by the amount of variance they account for, so the first eigenface is the one that accounts for the most variance in the face data set. The most significant 10 are shown in Figure 6.<sup>9</sup> These correspond to global, rigid, problem-specific features.
- **Eigenfeatures:** The second representation, eigenfeatures, uses the same process, except that the data used to find the principal components are restricted to the rectangular regions around the eyes and mouth (treated separately). Figure 7 shows the location of the eye and mouth eigenfeature samples (the larger outer boxes). Our data set’s top ten left “eigeneyes” are shown in Figure 8. These correspond to local, rigid, problem-specific features.
- **Local principal component eigenvectors:** For the third representation, local PCA, we computed the principal components of 1000  $64 \times 64$  pixel patches sampled uniformly from random positions in random faces in the database, with overlap allowed.<sup>10</sup> These eigenvectors yield a “basis image” representation that resembles the filtering performed

---

<sup>9</sup>Though the actual eigenvectors are relative to an origin defined by the “average” face, we visualize them by stretching each vector’s maximum and minimum values to 0 and 255 *individually*. Some researchers will do a *global* stretch, where the *same* transformation, based on the global maximum and minimum value, is applied to each eigenvector. The results can look quite different, so it is important for visualization purposes to know which technique is being used.

<sup>10</sup>We left out two of the 14 subjects, initials “A” and “C,” to ensure that classifier generalization does not depend on using the same face images for feature computation and classification.

by cells in primary visual cortex; Figure 9 shows 15 of the basis images. The features correspond to local, rigid features that are not tuned to the problem, nor are they explicitly part-based, though they are used to extract responses from regions around salient parts of the face.<sup>11</sup>

Now, to treat these basis vectors as features, for the eigenface network, a given face image is first normalized for luminance (to the same mean and variance used in obtaining the eigenfaces). We then subtract the “average” face (obtained earlier) and project the result onto the first  $k$  principal component eigenvectors, or eigenfaces. Thus, an input to the network is a “loading” on these  $k$  features. Since these loadings have widely varying variances, we normalize them by subtracting their mean and dividing by their standard deviation to produce “Z-scores” (Bishop, 1995).<sup>12</sup> Similarly, for the eyes and mouth, we again normalize luminance then project the subimages around the eyes and mouth onto the  $k$  most significant eigeneyes and eigenmouths. Finally, for the local PCA features, we project the eye and mouth portion of the face image onto the first  $k$  overlapping square  $64 \times 64$  local principal component eigenvectors. Again, we normalized the resulting projection coefficients for both local representations by their standard deviation to produce Z-scores.

The locations of these overlapping regions are also shown in Figure 7. There, one can see the three large outer blocks around the mouth and each eye that were used for eigenfeatures, and the seven smaller overlapping  $64 \times 64$  squares within them that were used for the local PCA projections.

### Classification with neural network ensembles

To compare these representations, we trained classifiers for each input representation. The classifier is a combination of several standard neural networks (we call this a “flock” of networks).

---

<sup>11</sup>One can argue whether we have placed these in the proper part of the space of features, since they are derived from the data in this task. However, if we take the principal components of any natural scene at a small enough scale, we will get features similar to these (Baddeley and Hancock, 1991).

<sup>12</sup>The astute reader will note that, if the projections onto the eigenvectors are Z-scored after projection, subtraction of the mean in advance makes no difference! Since all of our features are Z-scored, we ignore this detail in what follows.



Figure 6: The first 10 eigenfaces of the normalized Pictures of Facial Affect (POFA) database. Each is individually normalized to the [0,255] brightness range (see text).

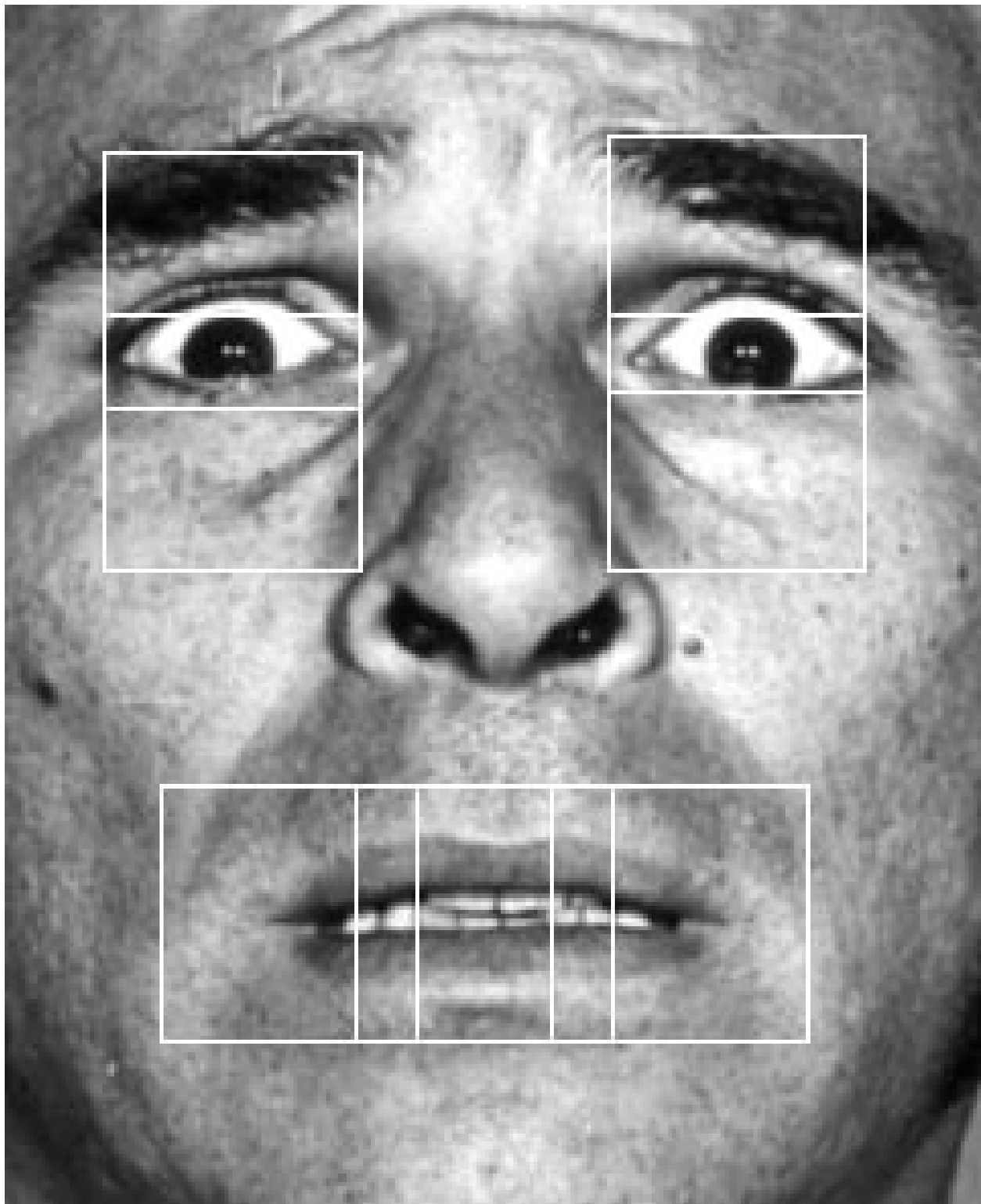


Figure 7: Regions used for eigenfeature and local PCA projection. Eigenfeatures use the three large regions around the eyes and mouth. Local PCA uses the seven overlapping square sub-blocks within the feature regions — two for each eye and three for the mouth.

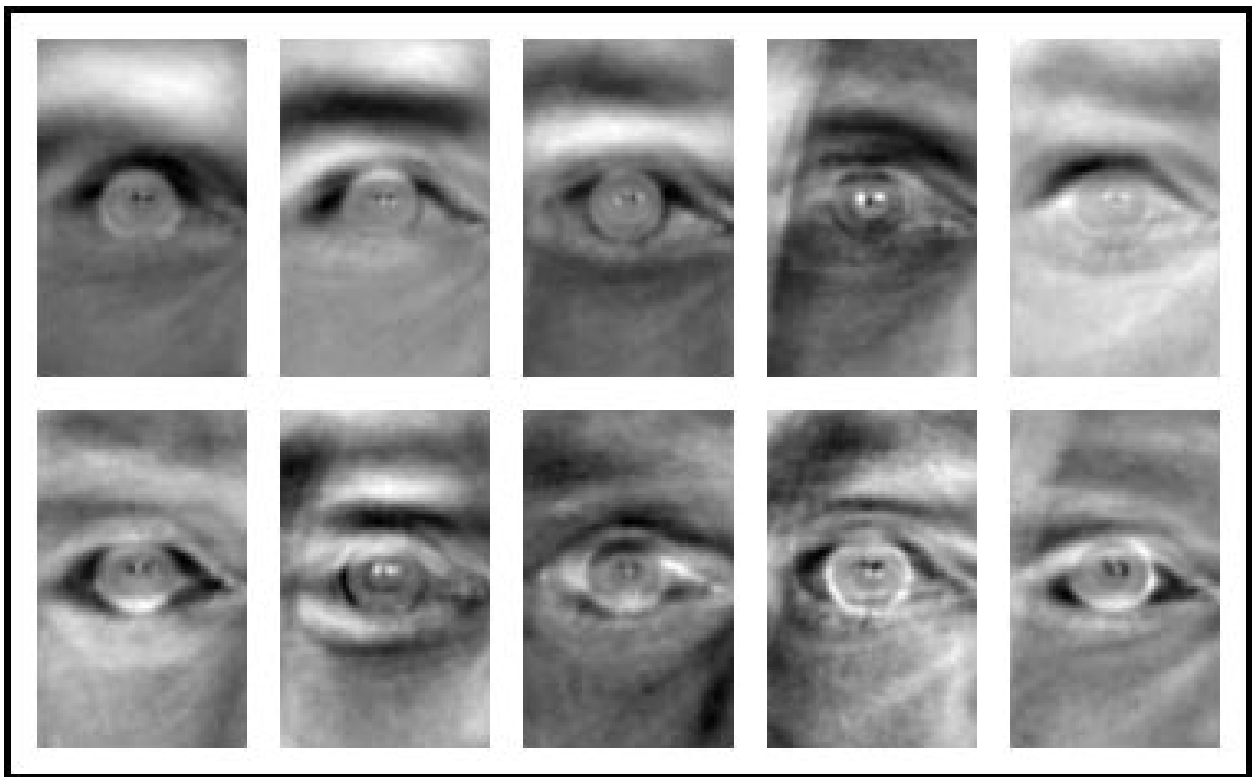


Figure 8: First 10 eigeneyes of the POFA database.

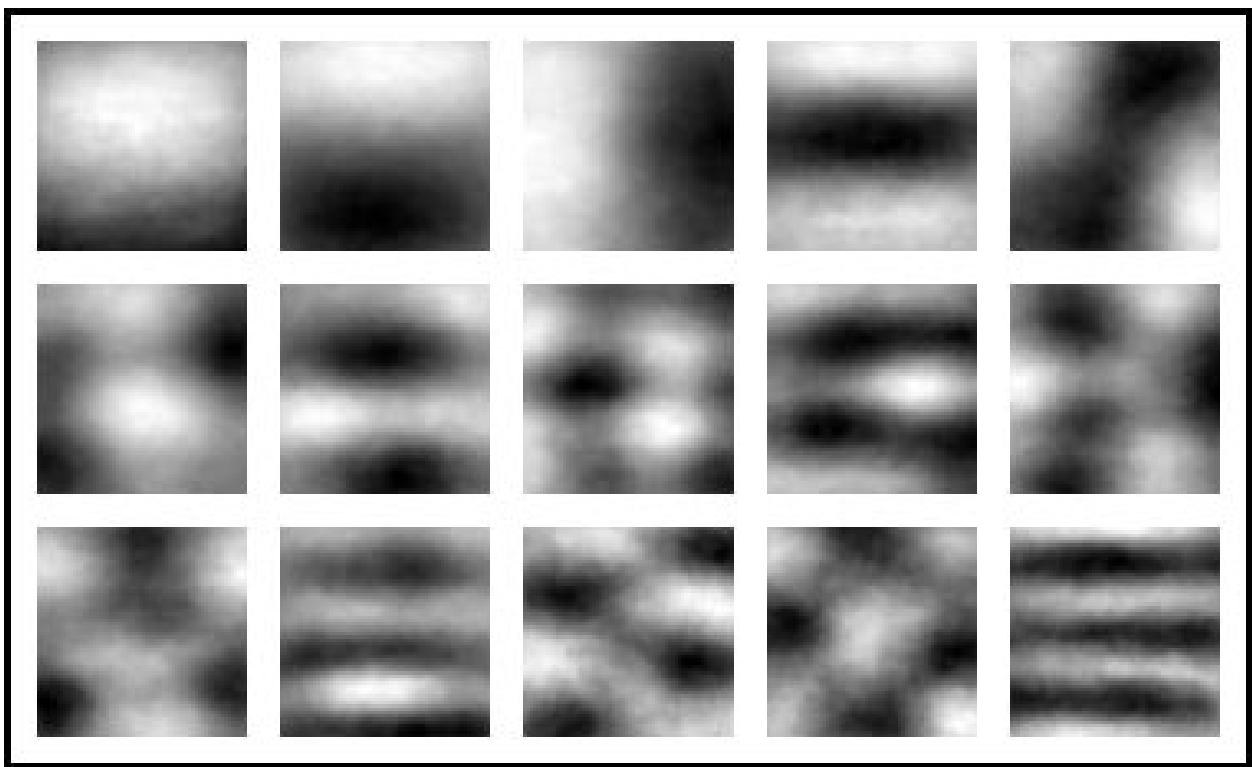


Figure 9: First 15 local principal component eigenvectors from the POFA database.

Each individual network in a flock has a hidden layer containing 10 units using the logistic activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

where

$$x = \sum_i w_i \times \text{input}_i \quad (3)$$

is the weighted sum of the inputs. The output layer contains 6 units, one for each emotion (also using Equation 2, and Equation 3 is applied to the hidden unit activations instead of the input). We trained the networks (that is, set the  $w_i$ 's) with backpropagation of mean squared error (Rumelhart et al., 1986a). For the training signal, we used the observed response vectors provided with the POFA slides by Ekman and Friesen. After training, the network outputs  $y_j$  for emotion  $j$  are converted to probabilities by using the softmax equation:

$$o_j = \frac{e^{y_j}}{\sum_k e^{y_k}} \quad (4)$$

The “flocks” of networks arose as follows. A standard approach to training a neural network, in order to avoid overtraining (which can result in poor generalization, if the training set is memorized), is to use a “hold-out set” of examples from the training set that are not used to change weights, but to estimate how the network will generalize to unseen data (Bishop, 1995). When generalization on this set of examples begins to decrease, training is halted. This is called “early stopping.” We performed early stopping, holding out a single individual’s images. Because of our limited training set, the technique can possibly perform poorly if we happen to pick a poor hold-out individual. If the hold-out individual’s expressions are too easy to decode, the training could stop too late, and if they are too hard, training could stop too early. In order to avoid this problem, we trained multiple networks, each one using a different individual as a hold-out. Thus, for a given training set, we ended up with a “flock” of networks. This is shown in Figure 10.

However, one also wants completely novel data after training to test generalization. So, we trained 14 different flocks, each flock using a different individual for testing generalization after training (this technique is called leave-one-out cross validation or the jackknife

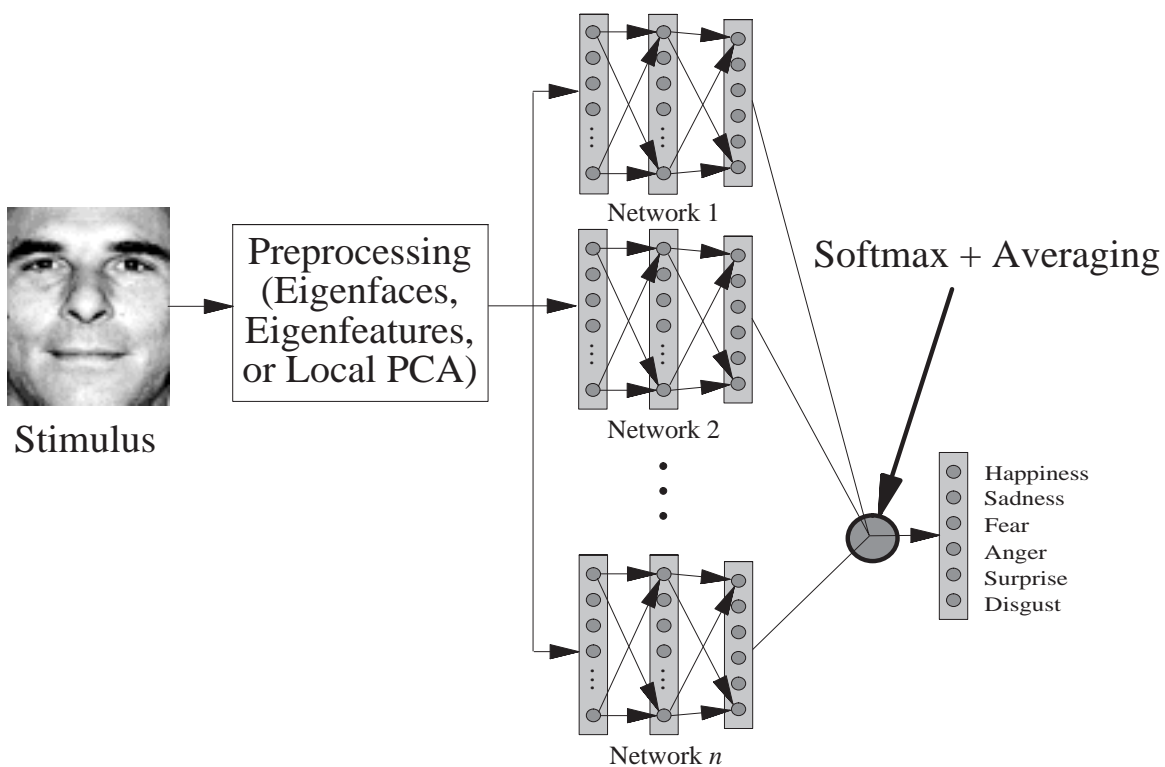


Figure 10: The classifier architecture. Each network is trained individually on a slightly different subset of the dataset, and their results combined using a softmax (see text).

procedure). Since each flock is trained on 13 individuals, then, we have 13 networks per flock. The final output of the flock,  $O_j$  for emotion  $j$ , is the average over the flock of each member's (softmax) output:

$$O_j = \frac{1}{13} \sum_k o_{jk} \quad (5)$$

where  $k$  indexes a flock member. This represents the probability that the facial emotion being portrayed in the input is the  $j$ th one.

## Results

Figure 11 displays the expected generalization achieved by ensembles initially seeded with small random weights for an increasing number of features in the respective representational spaces. That is, from left to right, we are adding one more principal component (in order of decreasing variance) to the corresponding representation. One should be aware that each point at a particular X-axis location represents quite different numbers of inputs to the networks, as adding one more eigenfeature adds 3 more inputs to the network, while doing the same for the eigenfaces and local PCA result in 1 and 7 more inputs, respectively. Hence this graph is biased in favor of the Local PCA representation, as it has many more inputs than the others. We consider the same data from the “number of inputs” point of view below. Each data point represents the average of 10 trials. The error bars reflect a 95% confidence interval on the mean. The curve (generalization rate vs. number of features) was evaluated at 19 pattern sizes for each representation.

Figure 11 shows that if Cottrell and Metcalfe had had a good database, they might have achieved generalization rates for novel individuals of on the order of 83% (eigenface curve), as this is essentially the EMPATH model with good data. This shows that eigenfaces are a good representation for expression recognition. However, the rest of the graph suggests that these are not the best features for expression recognition. Despite an early advantage at the smallest pattern size, eigenfeatures turn out to be relatively poor for this problem. On the other hand, these data demonstrate that better classification rates can be obtained using localized features that are not tuned to the data. An 87% generalization rate is achieved using local PCA features with 350 inputs (the top 50 local principal component eigenvectors

placed at seven locations on the face). This compares favorably with the results obtained from techniques that use dynamic facial action *sequences* (Cohn et al., 1999; Bartlett, 1998; Essa and Pentland, 1997; Yacoob and Davis, 1996). Such schemes make use of the dynamics of facial movement relative to the neutral face, which is not possible on novel, static face images, for humans or networks. The static image classification approach affords direct comparison of our model’s performance to that of humans in psychological studies using static images (Ekman and Friesen, 1977; Etcoff and Magee, 1992; Beale and Keil, 1995a; Young et al., 1997).

Another view of this same data is given in Figure 12. Here, we plot the generalization as a function of number of inputs to the network. This plot then corrects for the amount of information given to each network by the feature projections. The graph shows an early advantage for eigenfaces, with a late-arriving, but steadily increasing generalization level with the local PCA representation. Figure 13 carries the graph out further for the local PCA. There is an intrinsic limit on the number of eigenface features that can be obtained from this dataset (no more than the number of individual images minus one). The eigenfaces are able to essentially give complete information by the last eigenface, but there is still a generalization advantage to be gained from the local PCA representation.

What all of these graphs show is that there is a “sweet spot” for all of these representations, beyond which the generalization is hurt. There are two possible reasons for this. One is that more and more information about identity is given in the low-variance principal components, which is noise with respect to this task. Another is that the networks are unable to ignore this noise as it becomes a larger part of the input. Recall that z-scoring the projections makes them all equally strong in the input. Apparently, back-propagation is unable to filter out these uninformative inputs. Another observation to make here is that variance does not always equal information, especially with respect to a particular task.

## Discussion

This work shows that, at least for the face database and kind of classifier we used, localized, untuned features are best for extracting expression from faces. This is in contrast to previous successes using eigenfaces for facial identification. This reflects the difference between

these two tasks — facial expression identification requires finding something *common* across individuals, while face identification requires finding something *different*. Thus one might hypothesize that different features underlie these tasks in humans. Under this hypothesis, expression identification taps a lower-level, more spatially localized representation than does face identification.

One must be cautious, of course, in deriving such a conclusion, for several reasons. First, the differences we found between local and global features were relatively small. There is an optimal number of inputs for eigenfaces that leads to 83% correct, while local untuned features leads to 87% correct. Second, it has been shown in previous work that combining votes from local and global based classifiers can improve performance. Uttal and colleagues (Uttal et al., 1995a; Uttal et al., 1995b) and especially (Uttal, this volume) have proposed that face processing may use either or both kinds of features, depending on the available information in each. He has proposed a weighted AND/OR circuit, where input from either kind of feature each may have sufficient information for a decision, and when each is relatively weak, the combination of the two will do. The contribution from each kind of feature would give diminishing returns in the case where there is a lot of information on the other channel. The nonlinear sigmoidal function used in neural networks is, of course, an excellent example of such a circuit.

Finally, what is perhaps most interesting is that with appropriate feature extraction, both expression recognition and identity recognition can be learned quite well by a general-purpose pattern classifier such as the backpropagation network. There is nothing inherently “special” about such networks.

## Modeling human performance

As discussed in the introduction, aside from showing efficacy for a particular task, we also will tend to believe a model to the extent that it mimics human data on the task. In order to evaluate this aspect of our model, we applied the best version of EMPATH-II to the issue of categorical perception of emotions. In the following, we detail two experiments. The first compares human responses to model responses on the same images. The test data

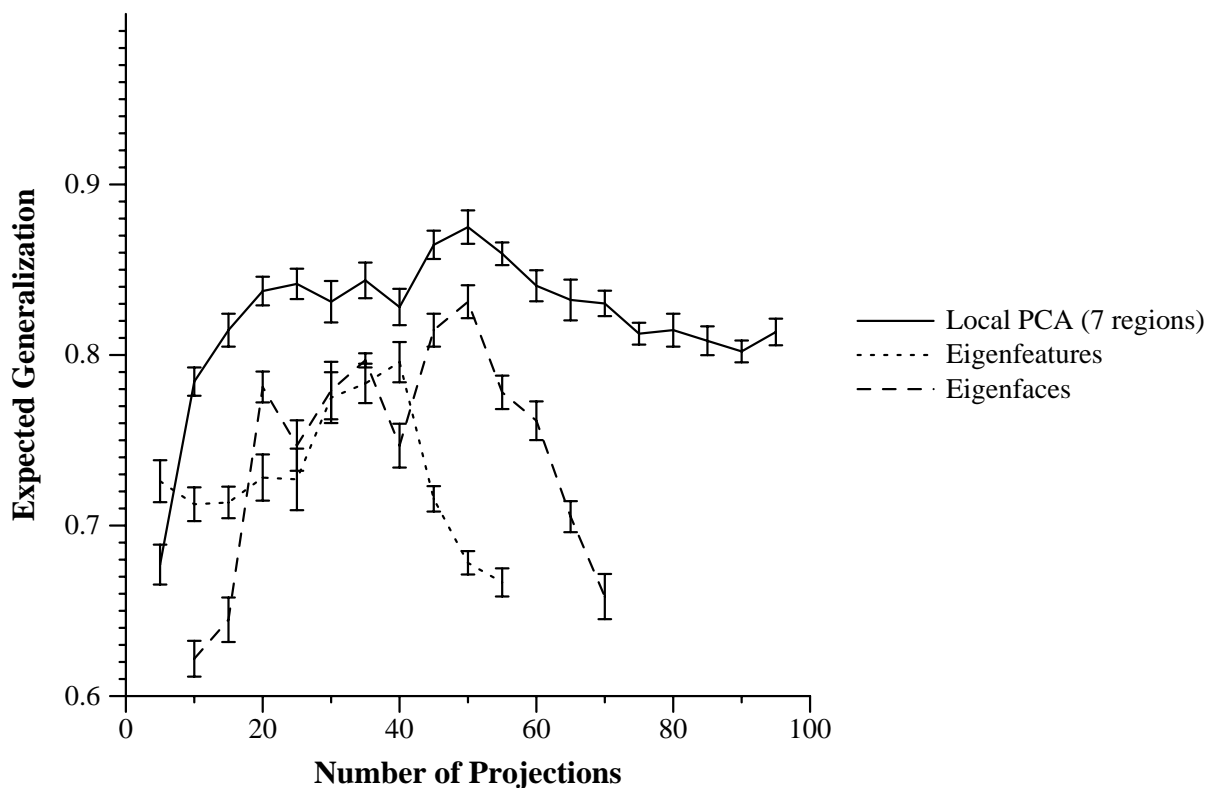


Figure 11: Generalization curves for feature-based representations and full-face representation. Error bars denote a 95% confidence interval on the mean over 10 runs with different initial random weights. X-axis is number of principal component eigenvectors used for each.

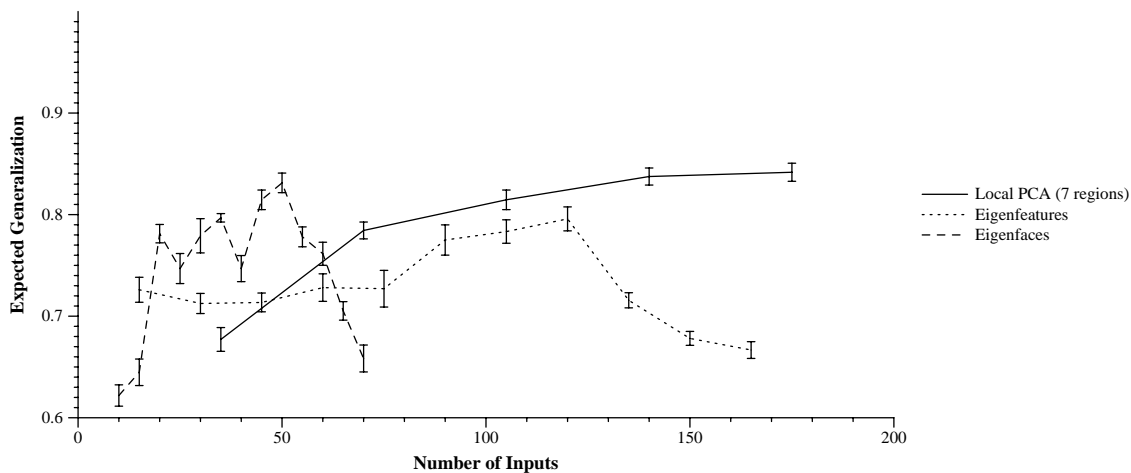


Figure 12: Generalization curves for feature-based representations and full-face representation. Error bars denote a 95% confidence interval on the mean over 10 runs with different initial random weights. X-axis is the total number of principal component projections used for each.

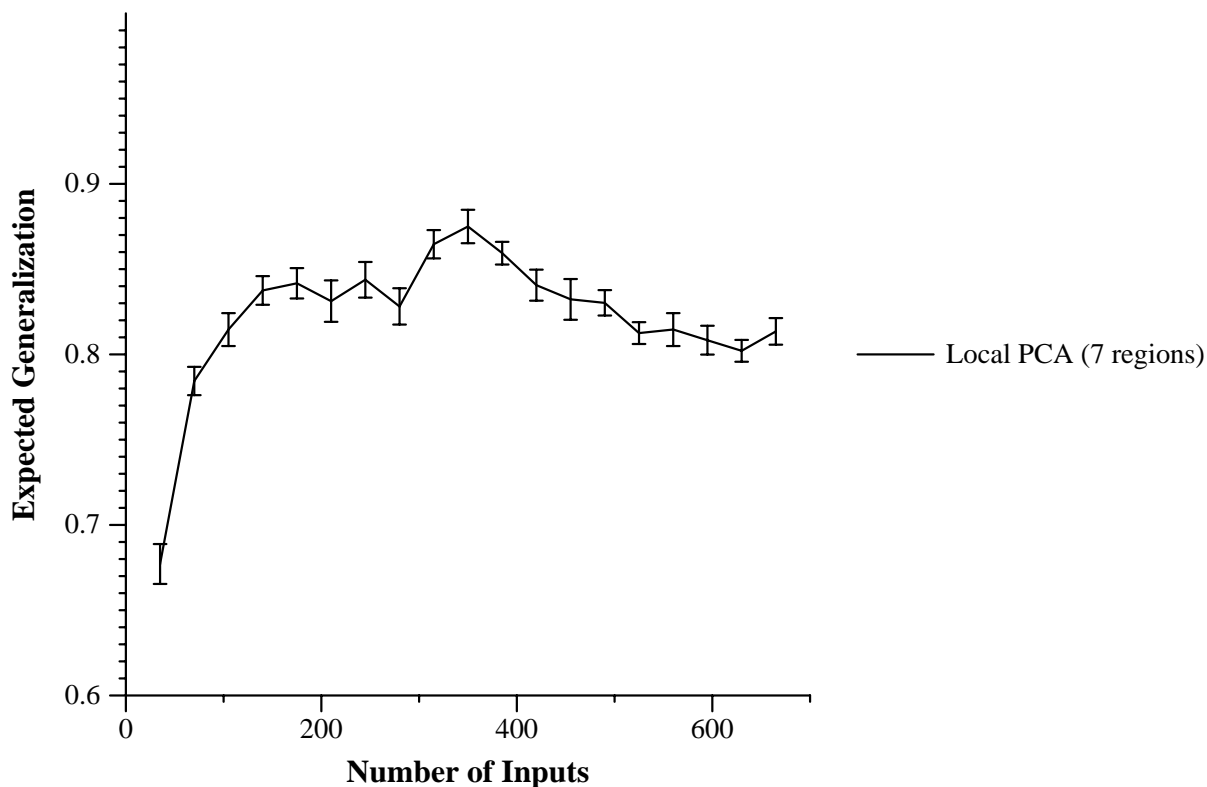


Figure 13: Generalization curves for the Local PCA-based representation. Everything else is as in previous Figure.

are dissolves (a kind of “poor man’s morphs”) from one expression to another. The second experiment compares the model to data from a published study that investigates human responses to morphs from all possible transitions for a particular subject in the database. The results are mixed: While some of the data is well-described by our model, further work is needed to account for all of it. To the extent that the model’s performance matches the data, we believe that this work shows that local untuned features are a good candidate for what is used in expression recognition by humans.

## Experiment 1

Inspired by Etcoff and Magee’s finding of categorical perception in transitions between emotional expressions, represented as line drawings (Etcoff and Magee, 1992), we wanted to approach the issue of categorical expression using images of actual facial expressions of emotion. This work was carried out apparently concurrently with the human subjects work of Calder et al. (1996), which used image-quality morphs of expression, and parallels it in some respects.

## Methods

To determine the type of transitions that the neural network model exhibits, we linearly transformed a face image of an individual expressing one emotion to the same individual expressing another at fixed intervals using a simple averaging technique in pixel space. That is, we simply formed the weighted average of images from two different emotions, with the weights moving progressively from one end of the sequence to the other. We used 9 steps: .9 of the first face with .1 of the other, .8 of the first with .2 of the other, etc. In computer graphics, this technique is called a *dissolve*. It worked reasonably well in this domain because we used registered images (the “normalization” we performed earlier), and the subject at either end was the same person. However, some artifacts do appear occasionally in the images, such as two mouths. Images in the resulting dissolve sequence can then be transformed into the input representation (the local PCA representation) and presented to the classifier as described in the previous section. Figure 14 shows a typical image sequence generated by

this process.

It is possible that the human response data that the network was trained with in the previous section could inform the network concerning the similarities seen by humans between the emotion images. If, for example, when subjects are presented with a portrayal of “disgust,” and respond “disgust” 90% of the time and “angry” 10% of the time, then the network training signal for that image would be .9 on the Disgust output and .1 on Angry. This potentially biases the network’s category boundaries to be more like the human ones, as this suggests that Angry is slightly similar to Disgust. Back propagation would then derive features such that this would be the case. To remove this possible source of information to the network, we trained a new set of networks on purely binary outputs, where the target for the emotion label most frequently used by Ekman’s subjects for a particular image is 1 and all others are 0. Otherwise, as before, this training was on the original, prototype images.

For this experiment, the “raw” ensemble output for emotion  $j$  is:

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

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

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

where  $\bar{a}_j$  and  $\sigma_j$  are the average and standard deviation of the raw ensemble output for emotion  $j$  over *all* training patterns.<sup>13</sup> The “final” ensemble output,  $O_j$  for emotion  $j$ , is the softmax of the Z scores:

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

The same training and stopping methodology for the flock is employed. However, for this study, we allowed ourselves a single fitting parameter,  $A$ , which was 1.0 during training, but then was adapted after training to human subject data (but not to the data we were trying to account for!):

$$O_j = \frac{e^{A*(a_j - ave_j)}}{\sum_k e^{A*(a_k - ave_k)}}$$

---

<sup>13</sup>Note that this work used a slightly different combination methodology for historical reasons; other simulations we have carried out suggest that this does not make much difference.

where  $ave_j$  is the average output over the training set on the  $j$ th emotion,  $a_j$  is the activity of the  $j$ th emotion (the average of the 11 networks outputs for emotion  $j$  for a particular image), and  $A$  is the scale factor. After training for best generalization by early stopping and cross validation as before, the value for  $A$  is determined by finding the minimum least square error between the outputs of the trained networks and the original response vectors that come with the Pictures of Facial Affect database (*not* the human data we gathered below). A single value of 7.0 for  $A$  was determined for all of the network ensembles. By varying initial random weights, we generated five different network flocks (“subjects”), all trained on the same data.

To compare the network responses to human responses, we tested human subjects on a sample of the same images. We tested Anger/Disgust, Anger/Fear, Disgust/Happiness, Happiness/Fear, Sadness/Disgust, Surprise/Anger, and Surprise/Happiness. Because we anticipated that category boundaries would be located near the middle of a transition between two emotions, we sampled this region more densely. The following numbers of stimuli were used: 6 images of each of the original two images (12 images total), 6 images of each of the two interpolated images closest to each of the two original images (12 images total), and 12 images of each of the remaining 7 interpolated images (84 images total), for a grand total of 108 stimuli. We showed human subjects these 108 stimuli in randomized order. Subjects were tested individually with no time limit on a two-alternative forced-choice labeling task between the endpoint emotion labels. Due to the time-intensive nature of the task, not all subjects were tested on all emotion transitions. All subjects had corrected-to-normal visual acuity and had given informed consent to participate in these studies.

## Results

In order to compare the behavior of human subjects and networks on the dissolve sequences, we decided to measure the width of the transition between the endpoint emotion categories. To do this, we counted the number of images in the dissolve sequence for which the rating on the endpoint emotion (averaged over subjects) was between 20% (.2 for the networks) and 80% (.8 for the networks). This is a simple measure of how wide the transition is from one emotion to the other. A narrow transition (low number) corresponds to behavior that has

been typically labeled as categorical perception, whereas a wide one (high number) does not. We compared the networks' numbers to those of the humans (see Table 1), and found using a two-tailed t-test adjusted for unequal sample sizes ( $df=15$ ) that the network data were not significantly different from the human subject data ( $p > 0.1$ ) for five of the transitions treated separately. The Surprise/Angry transition, however, was significantly different from the human data ( $t = 1.93, p < 0.1$ ). On the seventh transition, we inadvertently used a sequence to test for which one of the endpoint images (a Disgust image) was not correctly classified by any of the networks, and thus we had to throw this data out. Unfortunately, all this statistical analysis shows is that we cannot say that the performance of the two systems are different—it does not mean that we can now say they are the same!<sup>14</sup> Another shortcoming of this experiment is that we did not test the discriminability of the images for the human subjects at the transition points versus near the prototypes, a necessary condition in the standard definition of categorical perception (Harnad, 1987). However, in the next experiment, we do test the networks' discrimination of these stimuli and compare it to humans on similar stimuli.

We can also compare the human and network responses on particular transitions directly. A (particularly good!) sample graph of this sort is shown for one dissolve sequence in Figure 14. One obvious difference between the networks and the humans is the amount of variance at the transition (the larger standard deviation bars are the human data). It is a curious fact that variance does not seem to be reported often in the categorical perception of expressions literature. This high variance indicates the possibility that different subjects place the boundary between categories at different locations along the dissolve sequence.

## Discussion of Experiment 1

These results show that a simple neural network classifier can simulate fairly well the human responses to emotional expressions. While these results are suggestive, they do not address the complete picture required to assess categorical perception. In particular, we did not

---

<sup>14</sup>One might also compute the correlation between the network responses and the human responses. However, this kind of analysis is mainly useful for model comparison. It is hard to say what one isolated correlation means.

TABLE 1. Comparison of networks and humans on dissolve sequences.

Dissolve Sequence	Human $\mu$ ( $\sigma$ )	Network $\mu$ ( $\sigma$ )	Difference
Anger $\leftrightarrow$ Disgust	3.8 (1.4)	3.4 (0.6)	$t = 0.64, df = 15, p > 0.1$
Anger $\leftrightarrow$ Fear	3.1 (2.1)	2.4 (0.6)	$t = 0.72, df = 15, p > 0.1$
Happiness $\leftrightarrow$ Fear	1.5 (0.8)	2.0 (0.0)	$t = 1.44, df = 15, p > 0.1$
Sadness $\leftrightarrow$ Disgust	2.3 (0.5)	2.0 (0.7)	$t = 0.98, df = 15, p > 0.1$
Surprise $\leftrightarrow$ Anger	1.5 (0.5)	2.8 (0.7)	$t = 1.93, df = 15, p = 0.073$
Surprise $\leftrightarrow$ Happiness	1.7 (1.2)	2.4 (0.6)	$t = 1.27, df = 15, p > 0.1$

Table 1: Comparison of networks and humans on dissolve sequences.

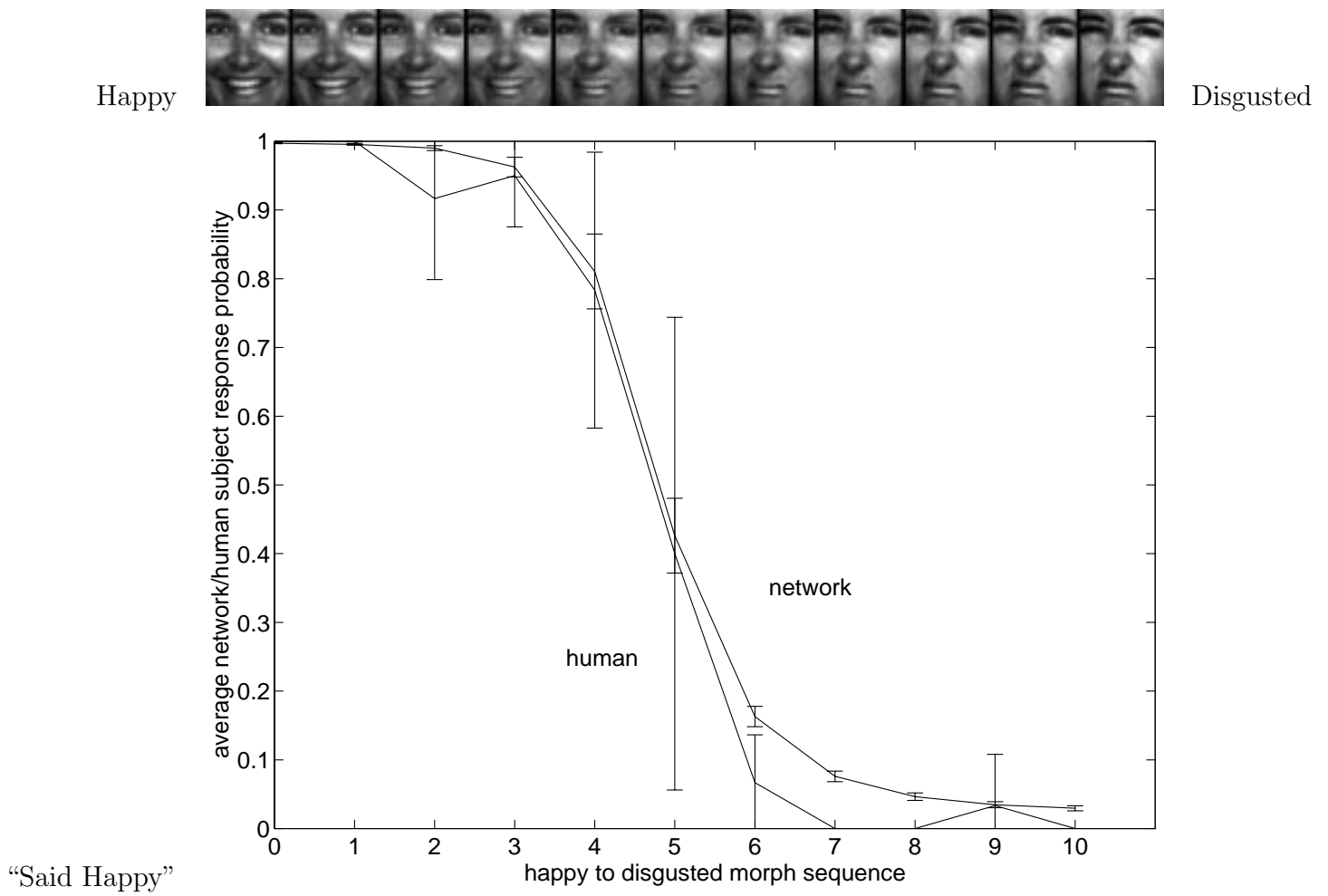


Figure 14: Comparison of network and human response on the same data.

measure the discriminability of the stimuli to our network, nor did we model reaction times. The first is important because categorization results are usually interpreted as “categorical” when subjects’ discrimination abilities are highest for images straddling the 50% response point, compared to images next to each other at other points in the transition, for equally sized steps in physical space. Reaction times also tend to be longer the closer the stimulus is to the decision boundary (Ashby et al., 1994). In the following experiment we remedy these omissions.

## Experiment 2: Modeling “Megamix”

In one of the most extensive studies with human subjects, Young et al. (1997, henceforth “Megamix”) show that image-quality morph sequences between six emotional expressions from the POFA database (Happy, Sad, Afraid, Angry, Surprised, and Disgusted) and “neutral” exhibit categorical behavior. In contrast to Etcoff and Magee’s work, they used photo quality images instead of line drawings. In contrast to Calder et al., *all* possible transitions between expression pairs for a single subject (“JJ”) from the POFA database (including neutral) were tested. This comprehensive study of human responses to facial expressions inspired us to assess our model against their data. In the following sections we review the results from the Megamix study in some detail, describe the application of EMPATH-II to modeling this data, and present preliminary results.

### Description of “Megamix”

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

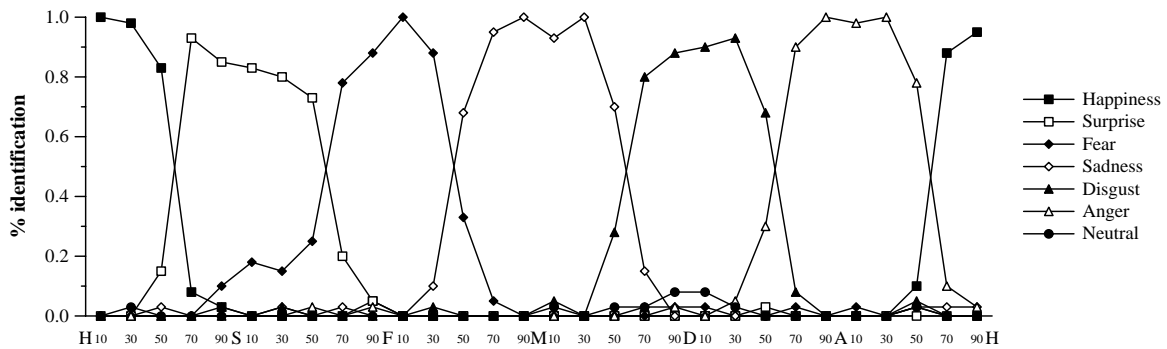
The focus of the Megamix study was in demonstrating that two dimensional accounts

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

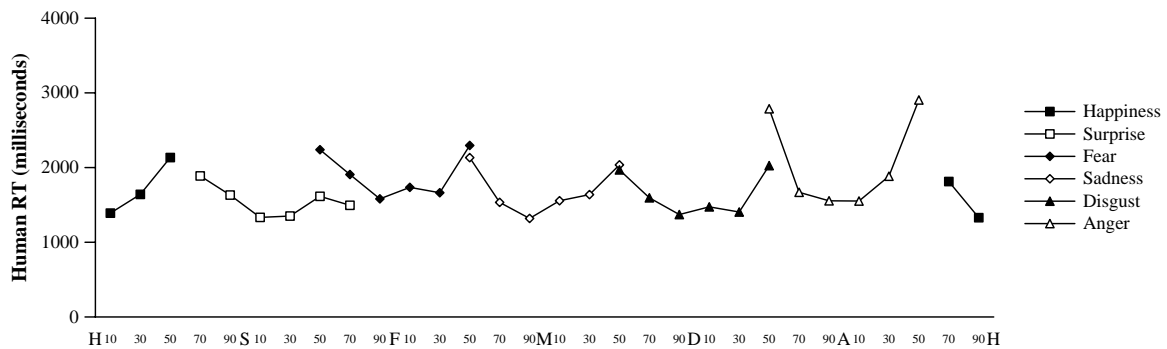
Thus, the issue is similar to what was discussed in the O’Toole et al. chapter (this volume)—what are the representational spaces involved for each kind of task? Are the same spaces used by humans in making similarity judgements also used for category judgements? While we don’t answer this question here, recent work in our lab with a system very similar to the one described has shown that the hidden units, when scaled using MDS, give the same circumplex as the human judgements that accompany the Pictures of Facial Affect database give when they are scaled using Euclidean distance. Thus, we are on our way to resolving this issue.

In the current study, we are interested in comparing the reported results of the human subject experiments in Megamix to the neural network model used in our previous experiments. In Megamix Experiment 1, Young et al. constructed image-quality morphs between all six of JJ’s emotional expressions. Megamix Experiment 2 added morphs from each expression prototype to JJ’s neutral image. Step sizes of 90%, 70%, 50%, 30%, and 10% were used between each pair of endpoints (105 unique images). These were presented in random order to subjects, who made a 7-way forced choice between the six emotion labels and neutral. Young et al. also recorded response times (RTs). They found the resulting RT curves were “scaloped,” with the slowest RTs near the decision boundary and faster RTs with increasing distance from the decision boundary (Ashby et al., 1994). Examples of the human subject expression identification response curves and response time curves are given in Figure 15.

In Megamix Experiment 3, subjects were required to discriminate (same/different judgments) simultaneously presented images that were one step away from each other along the transitions. The subjects showed better discrimination near category boundaries than near



(a)



(b)

Figure 15: Example response curves for subjects in Megamix Experiment 2. The expression sequence is Happiness – Surprise – Fear – Sadness – Disgust – Anger – Happiness. (a) Expression identification. (b) Response times. Data provided by Andrew Young.

prototypes, the standard requirement for categorical perception.

We wish to preview here one of the important aspects of our results. Given the above “laundry list” of requirements for categorical perception, the conclusion is often drawn that the underlying categories must be based upon discrete representations. We will show that our network is quite capable of accounting for the results of the Megamix experiment. However, there is no doubt that the representations used by our neural network are continuous nonlinear functions of the input features. The “categorical” part of our results are due to the decision process imposed on top of this continuous representation. We count the “answer” provided by the network to be the label on the output unit with the maximal activation after the S-shaped curve of the softmax function is applied. This gives rise, near the boundaries, to sharp transitions between the categories. Our other response variables, such as reaction time and discrimination scores, also arise naturally in this setting. Thus one should be careful in interpreting experiments that purport to show categorical perception as showing that the underlying representations are discrete. Massaro has been making this same point for years, most recently in (Ellison and Massaro, 1997).

Finally, in Megamix Experiment 4, Young et al. tested the extent to which their subjects could tell what two emotions were present in the morph images. This is important because, if the images are perceived categorically, then subjects should be poor at judging what other emotion is mixed into the image. They asked the subjects to give three responses to an image: which emotion it was closest to, then the next closest emotion, then the next, scored as 3, 2, and 1, respectively (the rest of the emotions were given a score of 0). One problem with this kind of measurement is that intrinsic similarity between certain emotional expressions, e.g. surprise and fear, could bias the subjects’ response. To control for this effect, the authors measured these biases by including the prototypes in the experiment. They then subtracted the subjects’ scores for the “near prototype” from the subjects’ scores for the morphs as the mix varied from 90% to 50%. Finally, they then averaged these difference scores across all emotions to obtain the average response to a prototype “being moved towards.” We plot their results with our model’s results in Figure 20; the figure shows that when the secondary expression (the “far prototype”) in a morph is 30% or 50% present, subjects are sensitive to its presence, giving it a significantly higher score than the unrelated expressions.

## Methods

We were unable to obtain the actual morph sequences used in the Megamix experiments, so in this preliminary study we use previously developed dissolve sequences for testing the transition behavior between expression pairs. The disadvantage of dissolves is that in some transitions, artifacts (multiple features) appear. On the other hand, dissolves have at least two advantages: (a) they are simple to create, and (b) they are truly linear in image space, unlike actual morphs. Hence we could make a claim that we are actually varying a physical quantity literally (albeit in a high dimensional space), as in studies of categorical perception of colors or sounds (indeed, Busey, 1998, has shown that perceptually, facial morph space is curved). Figure 16 shows dissolves for the human response curves displayed in Figure 15. The endpoints of the transitions are the same prototypical JJ faces used in Megamix.

Our model is the best-performing EMPATH-II classifier, which used projection of seven  $64 \times 64$ -pixel eye/mouth regions onto the top 50 local principal component eigenvectors (see Figures 7 and 9) as its face image representation. The training patterns consisted of 102 images of thirteen subjects (five male, eight female) from the POFA database (not including male subject JJ, the one used in Megamix for testing human subjects, who we reserve as a “novel” subject for the network)<sup>15</sup> These included images of faces portraying all six expressions plus thirteen neutral images.

Our network model of a “subject” consists of an ensemble of five individual neural networks, each trained on a different subset of the entire training set. We use a smaller ensemble size than in previous experiments to introduce more variability into the subject pool. As before, at test time, the outputs of an ensemble’s members are combined by softmaxing their output vectors (Equation 4) then averaging.

Within an ensemble, the individual networks have 350 inputs, 10 logistic hidden units, and 6 logistic output units, one for each emotion. Each individual network is actually trained on 10 randomly-chosen examples of each expression plus neutral (70 patterns) leaving the rest of the training set (32 patterns) as a hold-out set to stop training. In contrast to the

---

<sup>15</sup>The POFA database is unbalanced, in that not all subjects have all expressions represented in the database, and some have multiple occurrences of some expressions. Hence the number of images is not 13 times 7, or 91.

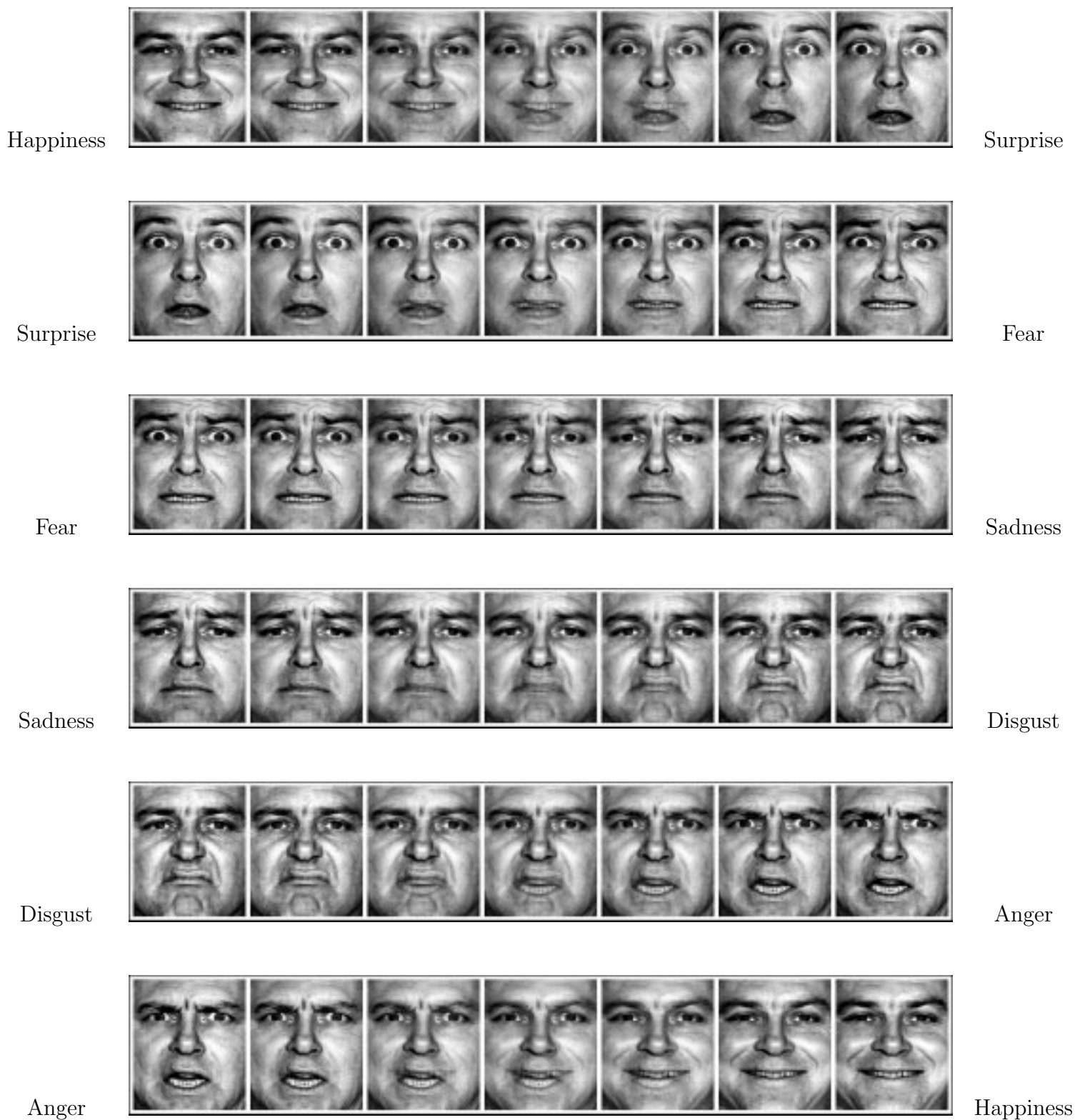


Figure 16: Example dissolve sequences of subject JJ from the POFA database. The six transitions reported in Megamix Experiment 2 are shown. The image sequences are linearly interpolated between the two database images at each extreme. Each sequence shows the prototypes at the extremes and the 10%, 30%, 50%, 70%, and 90% mixes between them.

previous EMPATH-II experiments, in which the teaching signal was the human response vector for each image, in this study the networks are trained to produce binary outputs — a “1” for the putative expression being portrayed and a “0” for the other five outputs. We have found that teaching networks to recognize neutral as a class does not generalize well. Instead, we teach the networks that the neutral faces are a *lack of* an emotion class, or the vector  $[0, 0, 0, 0, 0, 0]^T$ .

We trained 40 such network ensemble “subjects” and threw out those ensembles that did not generalize with 100% accuracy to the six expression prototypes, leaving 26 ensemble subjects.<sup>16</sup> The output values from the ensemble networks are used to generate *expression identification responses* to a given stimulus input, corresponding to a button press in the Megamix study. The highest output value for a particular input image is considered to be the emotion label of the button pressed. To model a “neutral” response given that the network has only six outputs, we compare the variance of the six outputs to a threshold and respond “neutral” when the variance is below threshold—since the sum of the ensemble’s output is always 1, a low-variance output vector best signals uncertainty. In this preliminary work, we set an ensemble’s variance threshold to a value slightly larger than its variance on the “neutral” JJ prototype. This is the *only* parameter we currently fit to the JJ data.

In addition to identification responses, we can also extract response times from our model. A standard measure of reaction time for a feed-forward neural network is to assume that RT is proportional to the output error (Seidenberg and McClelland, 1989). In the model’s case, since there is no predetermined correct response to the dissolve images, we simply use the difference between the maximum output (corresponding to the network’s *response*, see above), and the maximum *possible* output (1.0). Thus, the more uncertain the maximum response is (the farther from 1.0), the slower the RT. Since neutral is currently a special case, we do not model reaction time for a neutral response.

One point here that should be obvious is that this measure of reaction time will *naturally* lead to response times that are slower near the boundary between expression categories (as in the human data). This is because, by definition, the boundary is where the network is shifting its classification from one category to another. This then means that the network

---

<sup>16</sup>Not all subjects generalize perfectly because the number of networks per ensemble is so small.

output for the category on one side of the boundary is decreasing, while the output for the category on the other side of the boundary is increasing. Thus the maximal response will be lower than it is for a good example of the category. Thus, it should be no surprise when we show that the networks' responses show the same "scalloping" as the human subjects. However, the use of uncertainty in the response is independently motivated by previous work (Seidenberg and McClelland, 1989), and the observation that if a decision were being made by the subject, outputs that are weaker should naturally lead to longer reaction times. Indeed, we see this as a virtue of our approach that this result is simple to obtain.

To model the discriminability of a pair of stimuli as measured in Megamix Experiment 3, we suppose discriminability is based upon similarity. The more similar two stimuli are, the less discriminable they will be. We can think of similarity as deriving from the *gestalt* of the output of the network in this experiment.<sup>17</sup> Our similarity measure is the correlation between the output vector for stimulus  $i$  and stimulus  $j$ ,  $r_{ij}$ . Our dissimilarity measure is then  $d_{ij} = 1 - r_{ij}$ . Note that  $d_{ij}$  only makes sense as a measure of perceptual discriminability when the two stimuli are different images in the same expression transition sequence, because the exact same expression on two totally different faces could admit identical ensemble network output vectors. However, in this context, it is sufficient.

Again, it should be apparent that using this measure of discriminability will naturally lead to higher discriminability at the category boundaries, since this is where the outputs are changing most rapidly. Hence the correlation between the outputs will be lower, and discriminability will be higher. This is consistent with the basic tenets of categorical perception: since the stimuli are being "perceived" as different categories, they should be more discriminable. However, our measure does not assume that the actual categorization step (choosing the maximum response) has occurred! Hence this gives rise to a somewhat different explanation of the results.

Finally, to model the ranking of "closest emotions" given by the subjects in Experiment 4, we simply rank the corresponding emotion outputs in an ensemble's output vector.

---

<sup>17</sup>One could also use the hidden layer representation. We believe the results would be similar.

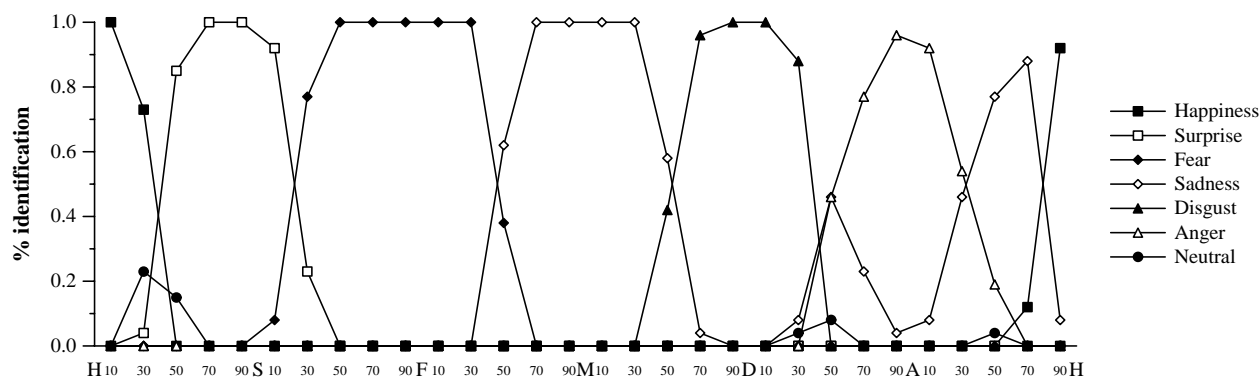


Figure 17: Average response of 26 ensemble networks to the same transitions displayed in Figure 15.

## Results

**Expression identification** The first analysis examines the average response curves (percentage of subjects giving a particular labeling to a stimulus) as the mixture of two emotion prototypes varies. The stimuli presented to both the neural network model and the human subjects were novel transitional faces. An example of the average response of the 26 ensemble networks to the Happiness – Surprise – Fear – Sadness – Disgust – Anger – Happiness transition series is presented in Figure 17, which can be compared directly with the human data in Figure 15a.

The most striking feature found in both the ensemble model and the subjects’ responses is very sharp transition regions from expression to expression across the sequence. This is true for all human transitions, including those not shown. For the model, the transition behavior was also sharp. However, at some points in some transitions, the most prominent emotion is not either of the prototypes involved in the transition. For instance, Figure 17 shows that Sadness intrudes upon the Anger–Happiness transition. Defining an intrusion as a point where the model identifies some unrelated expression more strongly than either of the mixed expressions for a given transition, we found that (a) Sadness intrudes on the Anger–Happiness transition (shown in Figure 17), (b) neutral intrudes on the Happiness–Fear transition, (c) Fear intrudes on the Sadness–Surprise transition, and (d) neutral intrudes on the Surprise–Disgust transition. Thus four of the 15 transitions between pairs of expressions show an intrusion by an unrelated expression.

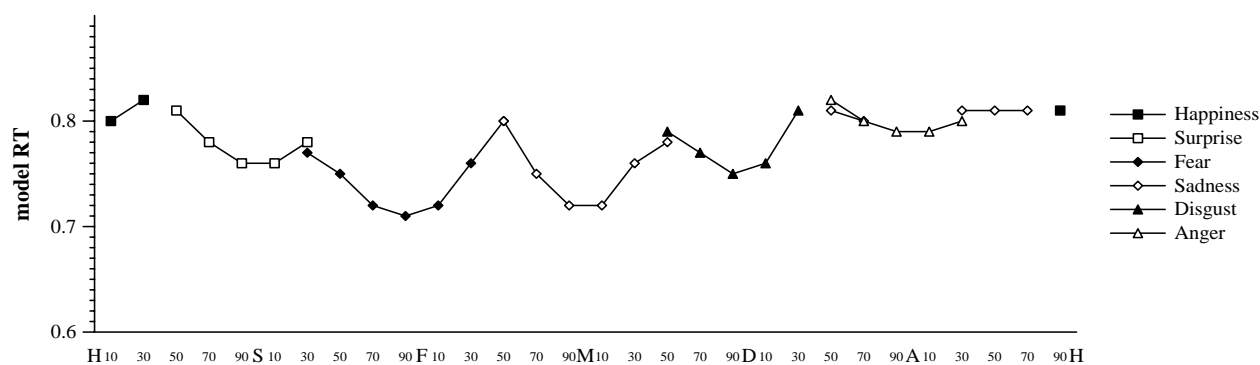


Figure 18: Model response times. See text for details on extracting response times from a network ensemble’s output. The model RTs show the same “scalped” pattern as in Megamix Experiment 2, with faster response near the prototypes and slower response further from the prototypes. The spurious Sadness RTs near Anger reflect the intrusions of Sadness around Anger in Figure 17.

We attribute intrusions like these to our use of dissolve sequences rather than image-quality morphs. Since dissolves are pure weighted averages, and morphs are inherently nonlinear, it makes sense that some mixes may actually resemble prototypes not actually involved in the mix. We plan on rerunning the experiment with image-quality morphs (currently under construction) to eliminate this possible confound.

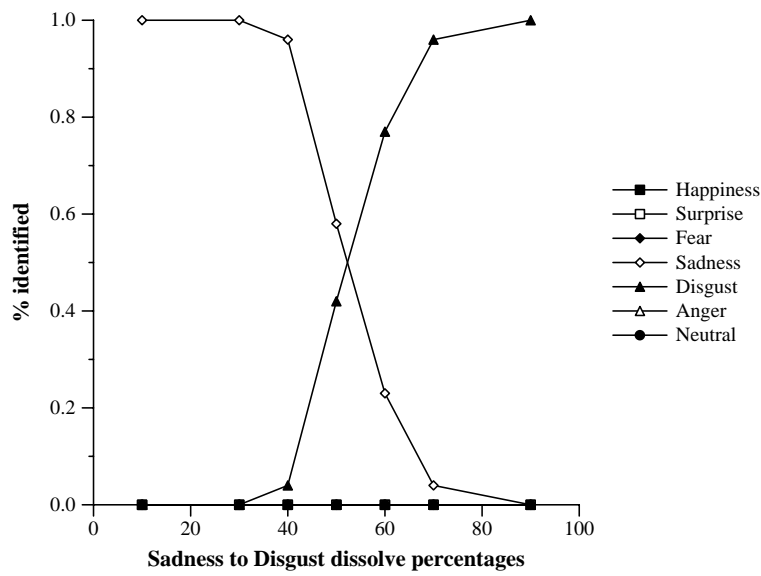
**Response times** Young et al. found that subjects’ response times were *scalped*, with faster RTs near prototype expressions and slower RTs further from prototypes. The graph in Figure 18 shows our model’s simulated reaction times for the same expression transition sequence previously shown in Figure 17. Given a stimulus  $s$ , for each emotion  $e$  that 23% or more of the network “subjects” identified in the stimulus, we plot the average response time for those subjects identifying emotion  $e$  in stimulus  $s$  (23% was the cutoff used in Megamix). The model’s RT curves show the same scalped pattern as the human RT curves in Megamix Experiment 2 (Figure 15b).

**Discriminability** In Megamix Experiment 3, Young et al. found that subjects were significantly better at discriminating pairs of stimuli near expression category boundaries than they were at discriminating pairs of stimuli closer to expression prototypes. Figure 19 shows

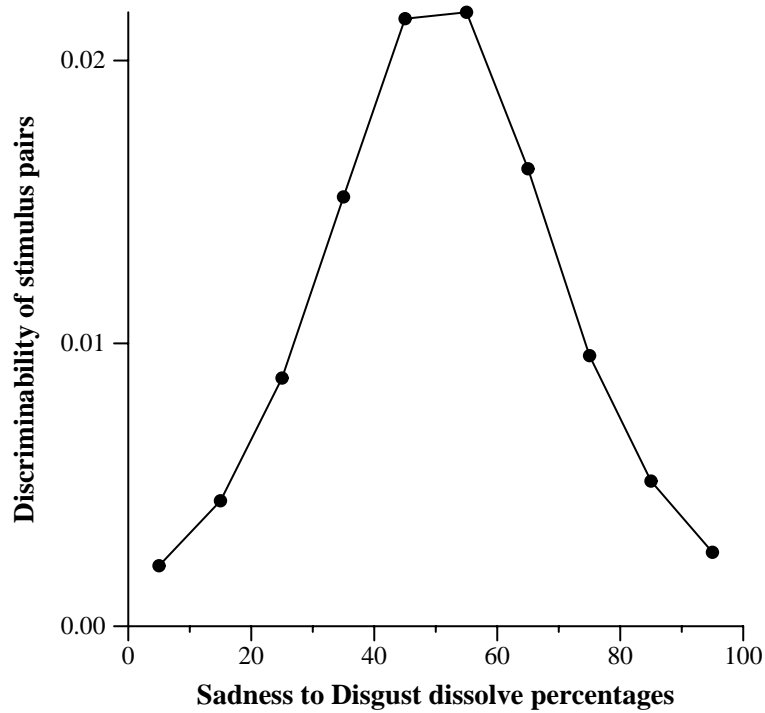
an example of the discriminability of pairs of stimuli close to and far away from expression prototypes in our model. The curves demonstrate that for the Sadness – Disgust transition, as in the human data, the model is most sensitive to stimulus changes near the boundary. To test this more formally, over all expression transitions, we found that the mean discriminability between stimulus pairs near prototypes (10% vs. 30% and 70% vs. 90%) is 0.04056 and that the mean discriminability between stimulus pairs closer to the transitions (30% vs. 50% and 50% vs. 70%) is 0.06237. The difference between the means is significant ( $z = 10.9, p < 0.001$ ). The same contrast was significant in Megamix Experiment 4, for both a sequential discrimination (ABX) and a simultaneous discrimination (same-different) task.

**Mixed-in expression identification** Megamix Experiments 1–3 strongly support the notion that emotional expressions are perceived categorically. In Experiment 4, Young et al. considered to what extent subjects were nevertheless sensitive to the second category mixed into a morph image. For example, can the subjects perceive the anger in a 90% happy/10% angry morph, even though they respond “happy” to the image? As described earlier, the authors asked their subjects to give three responses to an image: which expression it is closest to, then the next closest expression, then the next. They scored the three responses as 3, 2, and 1, and subtracted off the average score for the dominant prototype as described earlier. These difference scores were then averaged across subjects for the “prototype being moved towards” (they call this the “far prototype”). The scores for the other four unrelated emotion categories (those not represented in the morph) were averaged together as well. These two scores were then averaged across *all* transitions, and plotted. The human data is shown as the dashed lines in Figure 20.

We used the same methodology for our networks, using the rank order of the network outputs to extract scores. The results are shown in Figure 20 as the solid lines. The unfilled circles and squares show the difference scores for the four emotions *not* represented in the dissolves or morphs. The network’s behavior is very similar to that of the humans. The unrelated expressions consistently score near 0, indicating that neither the humans nor the networks false alarm on the morphs/dissolves more than one would expect given the intrinsic similarity of JJ’s prototypical expressions. The “far prototype,” on the other hand,



(a)



(b)

Figure 19: Discriminability of stimuli near a transition. (a) A closeup of the identification of emotions in the Sadness to Disgust transition. (b) Model discriminability of pairs of stimuli at 10% increments along the Sadness to Disgust transition, averaged across network ensembles. Discriminability is modeled as  $d_{ij} = 1 - r_{ij}$  where  $r_{ij}$  is correlation between the ensemble outputs for stimulus  $i$  and stimulus  $j$ .

is detected increasingly more reliably as its presence in the mix increases. The model is somewhat more sensitive to the secondary expression than the humans are; this may be attributable to the difference in stimuli.

## **Discussion of Experiment 2**

We have shown that a feed forward neural network model using a feature based representation of the face (projections of feature regions on a fixed filter set) accounts for the observations found in the human study. Specifically, the models exhibit categorical responses: sharp transitions in the response curves and higher discrimination across category boundaries. The scallop shape in the human RTs was also modeled by the same network. In addition, the models show a very good match to the human subjects' sensitivity to the non-dominant prototype being mixed into the images. Unlike the classical account of categorical perception, humans were able to make intra-categorical distinctions, and these results were reflected in the model as well.

As we discussed earlier, this result is consistent with Massaro's view that so-called "categorical perception" can be simply explained as the result of a decision process imposed upon an underlying continuous representational system (Massaro, 1987). In our model, the underlying continuous representation can be accessed at several levels in the model – the input, hidden, and output layers. At the output level, the sharp changes at the boundaries between categories are a natural consequence of the "soft competition" aspect of the softmax function. As the evidence for one category wanes, and the evidence for another waxes, this is reflected in a smooth, but steep, shift in the outputs for those two categories. Likewise, using the difference from the maximal output as our reaction time variable naturally leads to slower reaction times near the boundaries, where the evidence for each category is weakest. A similar story may be told for the discrimination scores. Thus, we have all of the "hallmarks" of categorical perception, but nowhere in the model is there a discrete set of categories. Indeed, our model is also consistent with the non-categorical aspect of the results, as is demonstrated forcefully in Figure 20.

The main point of departure between our model and the human data is intrusions of unrelated expressions into four of the 15 transitions between expression pairs. As we stated

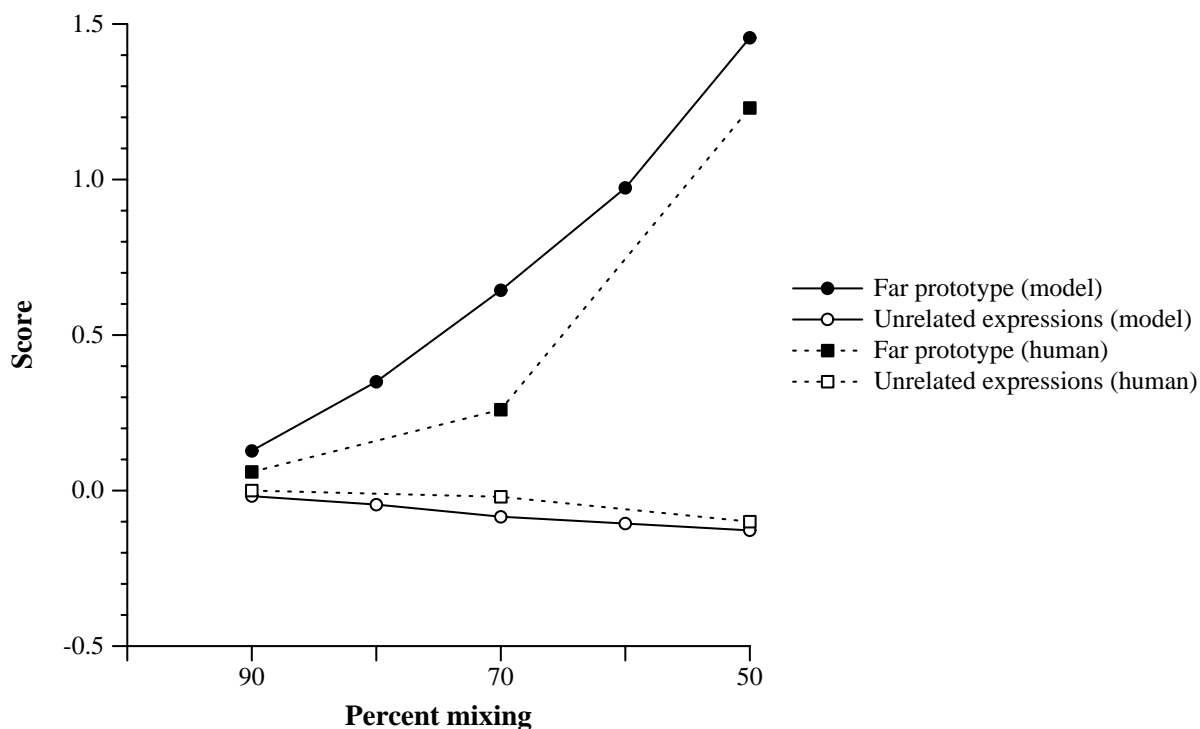


Figure 20: This graph compares the neural network model and human scores from the Megamix study, computed by the same method. The plots represent the average rating subjects give to the emotion in a face as it falls further from a given anchor prototype (see text for details). The emotions are lumped into two classes, the related emotion (the one being mixed with the prototype) and the unrelated emotions. Both the neural network model and the human subjects exhibit a step rise in prominence for the related emotion with no detectable increase for unrelated emotions. Human data estimated from graph in Young et al. (1997).

earlier, one likely reason for this difference is the way in which we constructed our transitions; simple pixel averages may be more easily confused with unrelated expressions than true morphs, which certainly do not fall on straight lines in pixel space (since morphing is a highly nonlinear operation) and also may not fall on straight lines in perceptual space (Busey, 1998). We plan to test this conjecture by applying our model to image-quality morphs in future work.

Unfortunately, we were unable to obtain the Megamix data for individual subjects, or even the average data in some cases. This prevented us from doing detailed statistical comparisons of our model with the Megamix data. Future work will include replications of the Megamix experiments, allowing more detailed comparisons between the model and data. We also have planned experiments suggested by the model. For example, we expect that training on certain categories of emotion more frequently than others will lead to shifts in transition points and steepening of category boundaries.

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

The reason that the neural network shows categorical perception is simple. Early in training, the network has small weights, which will result in shallow boundaries between the classes. As learning progresses, the weight sizes will increase, which sharpens the boundaries between the categories. Thus the region of ambiguity is shortened. However, different exemplars give different results. Easily identified emotions, as in the JJ images, give rise to steeper response changes than morphs between other subjects whose portrayals are not as pronounced. This is in agreement with other studies (Beale and Keil, 1995b) that show familiarity with the endpoints determines the steepness of the transition in human subjects.

## Conclusions

The main suggestion of this chapter is that the first order features underlying facial expression recognition are local, rigidly placed, untuned features. This falls in the “LRU” corner of the cube in Figure 1, one that (apparently) has been unexplored in face identification systems. Again, this means that the features are small in extent relative to the image, they are placed in the same location (not slid into a “best fit” location), and are not developed in the service of the task. Specifically for our model, this means that the features were small relative to the faces, that is, they were of the order of an eye in size, they were placed in the same position in every image (with the facial images normalized with respect to eye and mouth position), and they were obtained by principal components analysis of random patches from the image.

These first order features are not holistic, as they are not affected by features that would usually be considered contextual. As an example, the left eye features would not be affected by a distortion of the position of the right eye (for example, sliding it up) in the image.<sup>18</sup> Obviously, such a distortion would have an impact on our expression recognition system, with its rigidly-placed features, but presumably there would still be enough information from the one eye in alignment with the feature detectors to solve the problem.

The other features we tested, which did not generalize as well, were eigenfaces and eigeneyes. These fall in the LRS and GRS corners of the cube. Eigenfaces could legitimately be considered “holistic”, while eigenfeatures are not. We believe the reasons these features did not perform as well is because eigeneyes and eigenmouths code for both identity and expression, and identity information is noise with respect to expression recognition. The local PCA features, on the other hand, may also transmit identity information, but do it such a way that this noise is more easily filtered out by the rest of the network. As further evidence for this claim, consider another kind of untuned feature, Gabor filters. We have recently shown that these perform as well as local PCA in the expression recognition task (Dailey and Cottrell, 1999). This generalizes the results reported in this chapter to

---

<sup>18</sup>A one-eye position distortion makes an image look bizarre to normal subjects, while a two-eye movement makes a familiar face unrecognizable. How these distortions would affect facial expression recognition is unknown.

another example (perhaps more biologically realistic) of an untuned feature. Finally, we have not explored the final corner of the top of the cube, the global, rigidly-placed, untuned features (GRU). An example of such a feature would be Fourier components of the whole image.

We also have not explored the lower face of the cube, where features are adaptively placed on the face. One could view our normalization of the images in advance, to keep eye location and mouth location in the same place, as adapting the image to the feature location. The other way around is possibly more realistic, perhaps implemented by eye movements. We do not know of any research in expression recognition that does not allow the subject adequate time to make eye movements. Hence it is unclear what the performance of human subjects on the expression recognition task would be without eye movements.

The case we make here, that local untuned features are the best for expression recognition, suggests that this process taps a lower level set of features than face recognition does. If face recognition is using holistic features, then these would presumably be computed from local features. Disrupting the configuration of these would disrupt the face recognition process. However, if expression recognition “skips” a level, going straight to a categorization process, then configural changes may have less of an effect on expression recognition (but see the discussion below).

In the beginning of this chapter, we also described recent conjectures that face processing is “holistic” in nature. It is clear from this discussion that what we conceive of as holistic may need expansion. Generally, the term holistic is used to refer to a system that is sensitive to configural properties of the stimulus and displays context-dependent interpretation of the parts (Farah et al., 1998). Two of the prominent engineering-oriented face recognition systems we discussed, the USC system and the MIT system, display such characteristics, and suggest different ways in which this kind of phenomenon might occur. In the USC system, moving one of the components of the matching graph towards a better fit pulls on the other components. Thus, at least the matching process over one part of the face changes the response to other parts nearby. For the MIT system that uses the characteristics of mappings between intensity surfaces, changing a nose will change the intensity surface, and this may change the transformation from a reference intensity surface in a non-local manner.

These systems need to be tested on the stimuli used in psychological experiments, as others have suggested (Biederman and Kalocsai, 1998). This is the approach we have taken here in the context of expression recognition.

Given that we have stated that, for expression recognition, local features are the best, should we then conclude that our model does not use “holistic” features? Again, although we have resisted precise definition of this term, the phrase *configurational properties* keeps cropping up. Recall that we restricted our discussion of features to first order features – ones computed directly from the input. First order features such as eigenfaces are ones that could be called holistic at that level. Another way a system can be sensitive to configurational information is if it uses second order features, computed from first order features, that respond differently to different configurations of the first order features. This can certainly happen even when the first order features are local. Indeed, the visual cortex is characterized by larger receptive fields the farther upstream one goes from V1. The casual observer might note that at the hidden unit level, we have second order features whose “receptive fields” span the entire input. These may then respond to configurational information at the input level. This is a hypothesis that requires testing. For example, it may turn out that through training, there are hidden units that differentially respond to information around the eyes independently from variation around the mouth (Ellison and Massaro, 1997; Movellan and McClelland, 2000). This would represent a relatively non-holistic representation, consistent with work showing that human subjects process this information relatively independently. Or, it may turn out that all hidden units are sensitive to variation anywhere on the face. This would mean holistic processing, by the “context sensitive configural” definition, but would apparently deviate from the available data (Ellison and Massaro, 1997).

In any case, given a working computational model of the task, it is clear that we can investigate this model in ways that we cannot currently investigate human subjects. Pattern recognition models allow for the kind of analysis suggested in the previous paragraph. They play a role as both an intuition pump, and as a system in which our ideas concerning the term “holistic,” and other theoretical issues, such as the nature of so-called “categorical perception,” may be sharpened and tested. Indeed, our ability to obtain results taken to be consistent with categorical perception using a model with continuous representations

demonstrates that we need different criteria for assessing whether perception is indeed “categorical.”

Finally, models such as ours make predictions that can be tested in new experiments. We hope to carry out some of these experiments ourselves in order to validate or disprove our model. The idea that so-called “categorical perception” of higher level categories could be a learned phenomenon has appeared previously in the literature (Beale and Keil, 1995b). Our model shows that it can be learned. We plan to use the model to make specific predictions concerning how standard variables such as frequency and age of acquisition will affect the shape of the final categories.

## Acknowledgements

We would like to thank Andy Young for providing us with some of the data from the Megamix study, which is reproduced in Figure 15. We also thank the members of Gary’s Unbelievable Research Unit (GURU), one anonymous reviewer, Hervé Abdi, and the Editors for helpful comments. This research was supported in part by NIMH grant MH57075 to GWC.

## References

- Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9. Reprinted in (Anderson and Rosenfeld, 1988).
- Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. R. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, 372(6507):669–672.
- Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. R. (1995). Fear and the human amygdala. *Journal of Neuroscience*, 15:5879–5892.
- Anderson, J. A. and Rosenfeld, E., editors (1988). *Neurocomputing: Foundations of Research*. MIT Press, Cambridge, MA.

- Ashby, F. G., Boynton, G., and Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, 55:11–27.
- Baddeley, R. J. and Hancock, P. J. B. (1991). A statistical-analysis of natural images matches psychophysically derived orientation tuning curves. *PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON SERIES B BIOLOGICAL SCIENCES*, 246(1317):219–223.
- Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58.
- Bartlett, M. and Sejnowski, T. (1998). Learning viewpoint invariant face representations from visual experience in an attractor network. *Network: Computation in neural systems*, 9(3):399–417.
- Bartlett, M., Viola, P., Sejnowski, T., Larsen, J., Hager, J., and Ekman, P. (1996). Classifying facial action. In *Advances in Neural Information Processing Systems 8*, Cambridge, MA. MIT Press.
- Bartlett, M. S. (1998). *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California, San Diego.
- Beale, J. and Keil, F. (1995a). Categorical effects in the perception of faces. *Cognition*, 57:217–239.
- Beale, J. and Keil, F. (1995b). Categorical perception as an acquired phenomenon: What are the implications? In Smith, L. and Hancock, P., editors, *Neural Computation and Psychology: Workshops in Computing Series*, pages 176–187, London. Springer-Verlag.
- Biederman, I. and Kalocsai, P. (1998). Neural and psychophysical analysis of object and face recognition. In Wechsler, H., Phillips, P. J., Bruce, V., Soulie, F. F., and Huang, T., editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press, Oxford.

- Buhmann, J., Lades, M., and von der Malsburg, C. (1990a). Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the IJCNN International Joint Conference on Neural Networks*, pages 411–416.
- Buhmann, J., Lades, M., and von der Malsburg, C. (1990b). Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the IJCNN International Joint Conference on Neural Networks*, volume II, pages 411–416.
- Busey, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, 9(6):476–483.
- Calder, A., Young, A., Perrett, D., Ectoff, N., and Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, 3:81–117.
- Cohn, J. F., Zlochower, A. J., Lien, J., and Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology*, 36:35–43.
- Cottrell, G. W. (1990). Extracting features from faces using compression networks. In Touretzky, D. S., Elman, J. L., Sejnowski, T. J., and Hinton, G. E., editors, *Connectionist Models Proceedings of the 1990 Summer School*. Morgan Kaufmann Publishers, Inc: San Mateo, CA.
- Cottrell, G. W. and Fleming, M. (1990). Face recognition using unsupervised feature extraction. In *Proc. of the Int. Neural Network Conf.*, pages 322–325, Paris, France. Kluwer.
- Cottrell, G. W. and Metcalfe, J. (1991). Empath: Face, gender and emotion recognition using holons. In Lippman, R. P., Moody, J., and Touretzky, D. S., editors, *Advances in Neural Information Processing Systems 3*, pages 564–571, San Mateo. Morgan Kaufmann.
- Cottrell, G. W. and Munro, P. (1988). Principal components analysis of images via back propagation. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, Cambridge, MA. SPIE.

- Dailey, M. N. and Cottrell, G. W. (1999). PCA = Gabor for expression recognition. Technical Report CS-629, University of California, San Diego.
- 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.
- Ekman, P. and Friesen, W. (1976). *Pictures of Facial Affect*. Consulting Psychologists, Palo Alto, CA.
- Ekman, P. and Friesen, W. (1977). *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. *Journal of Experimental Psychology: Human Perception and Performance*, 23:in press.
- Essa, I. and Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763.
- Etcoff, N. L. and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44:227–240.
- Farah, M. J., Wilson, K. D., Drain, M., and Tanaka, J. N. (1998). What is “special” about face perception? *Psychological Review*, 105(3):482–498.
- Fleming, M. and Cottrell, G. W. (1990). Categorization of faces using unsupervised feature extraction. In *Proc. of the Int. Joint Conf. on Neural Networks*, volume 2, pages 65–70, San Diego, CA.
- Harnad, S. R. (1987). *Categorical perception: The groundwork of cognition*. Cambridge University Press, Cambridge, UK.
- Jones, J. and Palmer, L. (1987). An evaluation of the two-dimensional Gabor filter model of receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.

- Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311.
- Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Moghaddam, B., Nastar, C., and Pentland, A. (1996). Bayesian face recognition using deformable intensity surfaces. In *IEEE Conf. on Computer Vision & Pattern Recognition*, San Francisco, CA.
- Movellan, J. R. and McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, 17(4):463–496.
- Movellan, J. R. and McClelland, J. L. (2000). Connectionist models of perception and the morton-massaro law. In Solla, S. A., Leen, T. K., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, Cambridge, MA. MIT Press.
- Nastar, C. and Pentland, A. (1995). Matching and recognition using deformable intensity surfaces. In *IEEE International Symposium on Computer Vision*, Coral Gables, FL.
- Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., and von der Malsburg, C. (1998). The Bochum/USC face recognition system and how it fared in the FERET phase III test. In Wechsler, H., Phillips, P. J., Bruce, V., Soulie, F. F., and Huang, T., editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F. Springer-Verlag.
- Padgett, C. (1998). *A Neural Network Model for Facial Affect Classification*. PhD thesis, University of California, San Diego.
- Padgett, C. and Cottrell, G. (1995). Identifying emotion in static face images. In *Proceedings of the 2nd Joint Symposium on Neural Computation*, volume 5, pages 91–101, La Jolla, CA. University of California, San Diego.

- Padgett, C. and Cottrell, G. W. (1997). Representing face images for emotion classification. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, volume 9, pages 894–900, Cambridge, MA. MIT Press.
- Pentland, A. P., Moghaddam, B., and Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*.
- Phillips, P., Moon, H., Rauss, P., and Rizvi, S. (1997). The FERET evaluation methodology for face recognition algorithms. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 137–143, San Jose, Puerto Rico. IEEE.
- Phillips, P. J., O’Toole, A. J., Cheng, Y., Ross, B., and Wild, H. (1999). Assessing algorithms as computational models for human face recognition. Technical Report NISTIR 6348, National Institute of Standards and Technology. <http://www.nist.gov/itl/div894/894.03/pubs.html#face>.
- Phillips, P. J., Wechsler, H., Huang, J., and Rauss, P. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306.
- Plaut, D. (1991). *Connectionist Neuropsychology: The Breakdown and Recovery of Behavior in Lesioned Attractor Networks*. PhD thesis, Carnegie Mellon University. CMU-CS-91-185.
- Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning representations by back-propagating errors. *Nature*, 323:533–536. Reprinted in (Anderson and Rosenfeld, 1988).
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

- Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4):523–568.
- Tranel, D., Damasio, A. R., and Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, 38:690–696.
- Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *The Journal of Cognitive Neuroscience*, 3:71–86.
- Uttal, W. R. (1988). *On seeing forms*. Hillsdale, NJ: Erlbaum.
- Uttal, W. R., Baruch, T., and Allen, L. (1995a). Combining image degradations in a recognition task. *Perception & Psychophysics*, 57:682–691.
- Uttal, W. R., Baruch, T., and Allen, L. (1995b). The effect of combinations of image degradations in a discrimination task. *Perception & Psychophysics*, 57:668–681.
- Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779.
- Yacoob, Y. and Davis, L. S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642.
- Young, A. W., Rowland, D., Calder, A. J., Etcoff, N., Seth, A., and avid I. Perrett, D. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, 63:271–313.
- Yuille, A. L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70.