Human and computer recognition of facial expressions of emotion

J.M. Susskind[1], G. Littlewort[2], M.S. Bartlett[2], J. Movellan[2] and A.K. Anderson[1]

[1]Department of Psychology, University of Toronto

[2]Machine Perception Laboratory, Institute of Neural Computation,

University of California, San Diego

Shortened title: Computer and human expression recognition

Address correspondence to:
Adam K. Anderson
Department of Psychology
University of Toronto
100 St. George Street
Toronto, ON MS5 3G3
Canada
E-mail: anderson@psych.utoronto.ca
Phone: (416) 946-0207
Fax: (416) 978 4811

Abstract

Neuropsychological and neuroimaging evidence suggests that the human brain contains facial expression recognition detectors specialized for specific discrete emotions. However, some human behavioral data suggest that humans recognize expressions as similar and not discrete entities. This latter observation has been taken to indicate that internal representations of facial expressions may be best characterized as varying along continuous underlying dimensions. To examine the potential compatibility of these two views, the present study compared human and support vector machine (SVM) facial expression recognition performance. Separate SVM's were trained to develop fully automatic optimal recognition of 1 of 6 basic emotional expressions in real-time with no explicit training on expression similarity. Performance revealed high recognition accuracy for expression prototypes. Without explicit training of similarity detection, magnitude of activation across each emotion-specific SVM captured human judgments of expression similarity. This evidence suggests that combinations of expert classifiers from separate internal neural representations result in similarity judgments between expressions supporting the appearance of a continuous underlying dimensionality. Further, these data suggest similarity in expression meaning is supported by superficial similarities in expression appearance.

Introduction

The premise that emotions are discrete entities with distinct physiological signatures dates back to Charles Darwin's observations of continuity in prototypical displays of emotion across animal species (Darwin, 1872). Darwin speculated that displays across species mapped onto such emotion states as pain, anger, astonishment, and terror. In revisiting Darwin's observations, the universality of emotions was examined in cross-cultural human studies in which participants were asked to identify (Ekman & Friesen, 1971) and pose (Ekman, 1972) facial expressions associated with emotion-specific described contexts. A primary set of basic emotions was identified with characteristic facial signatures with substantial cross-cultural expression and recognition (Ekman & Friesen, 1971). Thus emotional experience and expression has been characterized as a set of discrete dimensions coding activation of specific states, such as fear, anger, sadness, or happiness (Ekman, 1992). More complex emotions, like love, may occur from secondary mixtures of these proposed basic prototypes. Basic emotions would then provide the palette from which more complex emotions are mixed (Plutchik, 1980).

Behavioral evidence from forced choice recognition of morphs between prototypical expressions demonstrates nonlinearities consistent with categorical perception, implying the existence of discrete expression categories (Calder et al, 1996; Etcoff & Magee, 1992; Young et al, 1997). Neuropsychological and neuroimaging evidence likewise provide evidence consistent with neurally localized discrete representations of facial expressions. Damage to the amygdala differentially impairs fear recognition whilst leaving other discrete emotions such as disgust recognition largely intact, while damage to anterior insula differentially impairs disgust recognition leaving fear recognition intact (Adolphs et al, 1999; Phillips et al, 1998).

Convergent evidence from functional neuroimaging demonstrates that fear expressions maximally activate the amygdala while disgust expressions maximally activate the anterior insula (Anderson et al, 2003; Philips et al, 1998). Similarly, discrete neural representations have recently been proposed for recognition of anger in the ventral striatum (Calder et al, 2004). To the extent that such dissociations in recognition can be found for a variety of basic prototypes would provide further evidence for their status as the primaries on which emotional experience and communication depend.

The alternative view of emotional space is characterized by lower order dimensions that suggest that emotions are fuzzy categories clustered on axes such as valence, arousal, or dominance (Russell, 1980; Russell & Bullock, 1986; Schlosberg, 1952). As such, emotions can be understood according to their relatively continuous ordering around a circumplex characterized by a few underlying dimensions. In these models, recognizing facial expressions relies on an ability to find the nearest cluster to the current exemplar in this continuous low-dimensional space rather than by matching to basic emotion prototypes. Behavioral evidence is consistent with some form of lower-order dimensional representation of emotions, whereby emotion types are closer (e.g., anger and disgust) than others (sadness and happiness) in emotion space. As such, expression judgments tend to overlap, indicating that emotion categories are not entirely discrete and independent. Proximity of particular expression exemplars (e.g. anger) to other expression exemplars (e.g. disgust) is tightly clustered across individuals, reflecting the possibility that categorization tasks force boundaries to be drawn in the lower dimensional expression space. In contrast with these lower order dimension theories, basic prototype accounts do not make explicit the similarity relationships between the basic emotions, as they do not explain the tight or distant clustering between expression types.

Although integrating behavioral accounts with neuropsychological and neuroimaging studies provides important data towards explaining emotional space, progress in the field of machine perception and machine learning offers an opportunity to test the computational consequences of different representational theories. Such an approach also affords examining to what extent recognition of emotional expressions directly reflects the statistical structure of the images to which humans are exposed. Parallel interest in facial expression recognition has been evolving in computer science as researchers focus on building socially interactive systems that attempt to infer the emotional state of users (Fasel, 2003). Progress in computer facial expression analysis has just begun to contribute to understanding the information representations and brain mechanisms involved in facial emotion perception because approaches from the various disciplines have not been integrated and closely compared with human recognition data.

Machine learning approaches to facial expression recognition provide a unique opportunity to explore the compatibility or incompatibility of different theories of emotion representation. To the degree that human data on facial expression recognition are consistent with basic prototype accounts, it is unclear if such representations can support the similarity relationships between the basic emotions, as do models that describe emotions in terms a small number of underlying dimensions. To address this issue, in the present study, we compared a computer model trained to make a 7-way forced choice between basic expressions plus neutral with human behavioral data. The system was developed by machine learning methods with the only goal of providing strong expression discrimination performance by developing distinct expert classifiers for different basic emotions. No attempt was made to fit human data. In the model, support vector machine (SVM) classifiers were trained to maximally discriminate a particular emotion. In contrast to traditional back-propagating neural networks that minimize the training error between network output and target for each training example (e.g., Dailey et al, 2002), SVM's learn an optimal decision boundary between two labeled classes by focusing on difficult training examples. This method finds features that maximally separate decision boundaries resulting in a high level of discrimination performance between expression types, minimizing false alarms to non-target expressions. Each expert is trained independently from all the other experts and then their opinions are integrated. The extent to which such a computer model of expression recognition correlates with human judgments of expression similarity will be a strong test of whether separate internal representations can support similarity judgments attributed to continuous underlying dimensions. Such a comparison can provide important computational constraints on how emotional expression recognition may take place in the human brain.

Methods

Computer Model Details

The system we tested was developed at UC San Diego's Machine Perception Laboratory (Littlewort et al, 2004). The software is currently distributed as part of the MPT/MPTX library (available online at http://mplab.ucsd.edu). This system was developed with the explicit purpose of performing robustly and in real-time in a fully automatic manner. The system can operate over video images at 30 frames per second, automatically extracting frontal faces, and categorizing the expression of the detected faces. During development of the model no attempt was made for it to fit human perceptual data.

The computer model (see Figure 1a) automatically finds and registers faces in images, extracts visual features, makes binary decisions about the presence of each of seven expressions (happiness, sadness, fear, disgust, anger, surprise, neutral), and then makes a multiple class decision. Face detection was performed by a system developed by Fasel et al (2005). The face detector uses a cascading decision tree based on the thresholded outputs of local oriented intensity difference detectors selected by a training process designed to detect frontal faces, and returns a rectangular face box with the candidate face region. It has an approximate hit rate of 90% for a false alarm rate of 1/million. The detector can process 320x240 pixel images in 1/30 of a second on a Pentium 4 personal computer. For the present study, faces were correctly detected in each expression exemplar used in the human behavioral experiment. After detecting a face, the system automatically extracted the face region from the image, converted the pixels to grayscale values, and rescaled the region to a common 96x96 window to standardize all training and test images. No further registration was performed. The computer system employed machine learning for subsequent feature selection as well as class decisions. No assumptions about facial expression appearance were programmed into the model.

Face images at the pixel level were then converted to Gabor magnitude representations using a bank of Gabor filters at 8 orientations and 5 spatial frequencies (4:16 pixels per cycle at octave steps). Gabor filters are Gaussian modulated sinusoidal gratings that approximate response properties of simple cells in primary visual cortex, essentially performing edge detection over locations, orientations, and scales (Lades et al, 1993). Figure 1b shows a single Gabor filter overlaid on a face. Gabor magnitude filters add the squared output of two filters with the same spatial frequency and orientation but out of phase by 90 degrees (Movellan, 2002). Converting face images to Gabor magnitudes results in image representations that are relatively resistant to slight translations in image registration compared to pixel representations. Moreover, using Gabor filters allows for representations that include overlapping features, a property seen in receptive fields of higher-level visual areas such as area IT (see Figure 2).

The resulting matrix of Gabor magnitudes contains on the order of 10^6 elements for a single image at 96x96 pixel resolution. Feature selection by AdaBoost was performed to reduce computational complexity and to encode images with a minimal set of highly useful features (Friedman et al, 2000). By reducing complexity, systematic feature selection by AdaBoost eliminates redundancy in representation and decreases the propensity for making false alarms. AdaBoost selects features iteratively, resulting in a reduced feature set in which each successive feature is contingent on previously selected features. The process can be interpreted as a maximum likelihood sequential optimization process for the generalized linear model (Freeman, 1979). In contrast to principal component analysis (PCA), which is an unsupervised technique, AdaBoost is supervised. PCA selects features that maintain as much information as possible about the input images, whereas AdaBoost selects features that maintain as much information as possible about the categories of interest. In practice, AdaBoost was a much more effective feature selection technique than PCA for expression classification using the computer model (Bartlett et al, 2004).

Gabor features are combined into a single feature vector after AdaBoost selects those features that are deemed most useful in discriminating each one-versus-rest combination of expression types. Linear SVM classifiers were used to make local expert decisions on one expression versus the rest, using the Gabor feature representations selected by AdaBoost. In a former study, discrimination performance on facial expressions with SVM’s exceeded that of alternative methods such as traditional neural networks (multilayer perceptrons) and linear discriminant analysis (Littlewort et al, 2004). Support vector classification (Vapnik, 1998) is particularly useful in situations where feature data are high dimensional and not necessarily linearly related to the input space. SVM classifiers are a regression technique that provides a generic framework for finding a hyperplane decision boundary that achieves the largest separating margin between positive (target) and negative (non-target) training exemplars. The decision boundary is generic in the sense that any non-linear decision function can be used (e.g. polynomial, Gaussian). However, in the current study, a linear hyperplane was chosen based on comparable performance to more complicated functions. Those training examples that fall closest to the boundary between positive and negative classes are called support vectors. The separating margin is defined as the distance between the support vectors on each side of the boundary.

The model has 7 different SVM classifiers, one for each of the 6 basic expressions plus neutral. Each SVM was trained separately to discriminate one expression from the other 5 plus neutral. A particular exemplar feeds to each of the classifiers, which produce a weighted “Yes” or “No” answer for whether the emotion specific to each SVM is detected. Positive output values indicate one side of the decision boundary while negative output values indicate the other. The output magnitude for each classifier corresponds to the distance from the decision boundary that an exemplar falls. Maximizing the decision margin between two classes of data within an SVM optimizes the trade-off between model complexity and goodness of fit to the data. Thus the model is set up to drive expression types apart (within a SVM) while attempting to maintain the ability to generalize to new exemplars from the same expression type. As such, the model is designed to minimize similarity in response between expression types. Figure 3 provides a pictorial representation of the weights learned by an SVM for discriminating two expressions.

AdaBoost feature selection and SVM model parameters were trained using 625 posed expression images from Cohn and Kanade's DFAT-504 (Kanade et al, 2000) and 110 exemplars from Ekman and Friesen's Pictures of Facial Affect (POFA; Ekman & Friesen, 1976) datasets, totaling more than 50 independent face exemplars for each of the six emotion types plus comparable numbers of neutral. DFAT-504 consists of 100 university students ranging in age from 18 to 30 years. 65% were female, 15% were African-American, and 3% were Asian or Latino. State of the art performance was achieved using leave-one-out benchmarking (93% generalization on a 7-alternative forced choice test).

Experiment Participants

Twenty-three undergraduate participants from the University of Toronto Psychology Department volunteered for this study for optional course credit. Informed consent was obtained from each participant prior to his or her involvement in this study in accordance with the ethics guidelines at the university.

Stimuli

To test the expression recognition and generalization performance for both the computer model and human subjects, we first compared computer performance on the POFA set with human norms (Ekman & Friesen, 1976). Generalization performance was tested on a distinct set of eight exemplars (2 Male/2 Female Caucasians and 2 Male/2 Female Asians) from each category including Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral obtained from Ekman and Matsumoto's Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and Neutral Faces (JACNeuF) datasets (Biehl et al, 1997). Each stimulus was converted to grayscale using Adobe Photoshop 7.0 and was normalized for contrast differences between stimuli. Stimuli were displayed at 418x463 image resolution to both the human subjects and the computer model.

Experimental Design and Procedure

Participants were asked to rate images of facial expressions with respect to labels corresponding to 6 basic emotions (anger, happiness, surprise, sadness, fear, and disgust) on a scale from 1 to 7 (1 – not at all; 7 – very much) (Adolphs et al, 1994). Stimuli and rating scales were presented in random order, continuing until each exemplar was rated on each of the six scales. All stimuli were presented and responses were recorded via computer. The experiment required 30-45 minutes to complete.

The same stimuli were presented to the computer model. For the purposes of evaluating discrimination performance and comparing to human ratings similarity, all ratings were converted into standard score (z) units. Outputs from each SVM classifier were converted from a signed distance to the decision boundary to a z-score, using the mean and standard deviation computed across all exemplars.

Results

Discrimination Performance

Computer model outputs were measured on generalization to untrained exemplars on which human subjects made their judgments. As illustrated in Figure 4, standardized ratings for each of the target emotions (ratings for humans and SVM activations) demonstrate that the model performed comparably to human ratings for all expressions (falling within 1 standard deviation). For both human and model judgments, consistent with accurate discrimination, the target expression received the highest average ratings for each expression type (i.e., surprise ratings were highest for surprise, sadness ratings highest for sad, etc.).

As a different index of discrimination performance, the continuous data were converted into a force choice format by defining correct responses as the proportion of exemplars on which the maximal response was for the target prototypical label. Humans correctly classified the target expressions with differing degrees of accuracy (mean=89.2%, SD=0.17). Happiness (mean=98.4%, SD=0.04), followed by sadness (91.8%, 0.10) and surprise (92.9%, 0.10) were discriminated accurately followed by anger (88.0%, 0.16), fear (84.8%, 0.17) and disgust (79.3%, 0.30). An ANOVA demonstrated statistically reliable differences in accuracy across expression types, $F(5,132)=3.71$, $p<0.005$, consistent with expression recognition success differing across expression type. The computer model showed good generalization performance on the untrained exemplars (mean=79.2%, SD=.292). Similar to human performance, accuracy was highest for happiness=100%, sadness=100% and surprise=100%, with less accurate performance on anger=75%, disgust=75% and relatively poor performance on fear (25%). Despite the model's high average ratings of fear for the fear prototypes, the low forced choice performance for fear expressions reflected an overlap with surprise and sadness ratings. Fear expressions were classified as surprise 62.5% of the time, and as sadness 12.5% of the time. The low accuracy for fear relative to the other expressions is consistent with evidence that fear recognition is particularly difficult (Rapzcak et al., 2000).

Inspection of Figure 4 revealed rather than all human perceivers demonstrating identical expression recognition performance there was substantial variability across subjects, in particular, for anger, disgust and fear. We used principal component analysis (PCA) to explore the patterns of accuracy variability across subjects and the model. The PCA two-factor solution indicated two clear clusters each containing 7 of the 23 subjects. As shown in Figure 5a, the cluster that contained the computer model, was characterized by subjects who had difficulty with fear. The other cluster, not containing the model, was defined by difficulty in classifying disgust (Figure 5b). Thus, to the degree the model differs from idealized mean group performance, it also behaved similarly to a major subgroup of human participants.

Similarity performance

We next examined whether the model's appreciation of expression similarity was comparable to human observers. Similarity of exemplars in terms of average subject ratings across expression types was computed and visualized using multidimensional scaling (MDS) analyses of the human data (average rating for each exemplar on the six emotion scales). The same analysis was performed for the computer model. Human and computer MDS plots were then compared for similarity of the relative positions of exemplars on a circumplex across the six basic expressions.

Trained exemplars. Human rating norms from the original POFA rating study (Ekman & Friesen, 1976) were compared to SVM ratings using MDS. Focusing on training exemplars allowed examination of similarity on data where discrimination between expressions classes was most accurate. The MDS circumplex for the human ratings, shown in Figure 6a, demonstrates that each expression class is clustered tightly together with no overlap between adjacent classes. In addition, exemplars were clustered in a characteristic order, replicating MDS analyses in previous studies (Adolphs et al, 2000; Dailey et al, 2000; Dailey et al, 2002). Although more diffuse than mean human performance, highly distinct clusters were also formed in the computer model (Figure 6b). Critically, the ordering of the clusters and their relative positions was identical to human observers. For example, anger exemplars were rated between sadness and disgust, surprise was between happiness and fear, with sadness rated maximally distant from happiness. MDS for human and computer model data resulted in similar levels of stress (Stress-I) in two-dimensional solutions (0.256 vs. 0.257). Thus, where supervised training achieved maximal discrimination of expressions types, a secondary unsupervised aspect of performance was the model's capturing of the similarity between expression types.

Untrained exemplars. We next assessed similarity performance on exemplars not in the model's training set. Human and computer ratings were again converted to standard scores for comparison. MDS on the human ratings verified that the circumplex ordering matched the above reported human norms for POFA (see Figure 7a). Adjacent clusters were no longer equidistant; angry exemplars fell in close proximity to disgust exemplars while fear exemplars fell close to surprise exemplars, suggesting greater perceived similarity in these expression pairings in comparison to the POFA images.

Our above finding of individual differences in discrimination of fear and disgust expressions may be due in part to the perceived similarity with adjacent clusters on the circumplex. To address this further, we examined MDS solutions on subjects who formed the two major sub-clusters in discrimination performance reported above (see Figure 5). As illustrated in Figure 7b, individuals reveal different clustering from idealized mean performance, with much less separation of expression types, such as fear and surprise, or disgust and anger. This demonstrates that ordering and clustering on the circumplex is somewhat variable and that averaging over subjects reveals a stronger tendency towards clustering than may be present in individual subjects.

When MDS was applied to the computer model, despite overall similarity in the circumplex solution, generalization to new exemplars revealed looser clustering of exemplars and more overlap between expression types than the mean human ratings, as depicted in Figure 7c. However, the model's performance appears more similar to individual subjects; in particular those with less pronounced discrimination of fear (Figure 7b). Critically, the circumplex for the computer ratings followed the same order as the human circumplex, demonstrated in both group and individual subject data. In particular, where the computer model fails to define distinct clusters it largely captures the similarity amongst exemplar types in humans. MDS solutions for human and computer model data resulted in similar levels of stress with a two-dimensional projection (0.157 vs. 0.221).

Despite the more sparse clustering found in the computer model relative to average human data, the correlation coefficient between human and computer judgments across expression types was very high ($r = 0.80$, $p < .001$), suggesting a great deal of similarity in the rating patterns. Examining how well the activation of distinct expert SVM's (anger, fear, disgust, etc) corresponded to humans, we found that specific correlations for each expression type were consistently high (anger, $r = 0.96$; sadness, $r = 0.94$; happiness, $r = 0.89$; fear $r = 0.85$; surprise, $r=0.83$; disgust, $r = .60$). For example, as illustrated in Figure 8 for fear expressions, humans and SVM experts agreed upon fear as the target expression, and also rated surprise as the most similar relative to the other expression types. The model's capturing of the similarity between fear and surprise underlies its poor discrimination of fear, often providing false alarms to surprise. Similarly, with anger expressions, humans and SVM's agreed upon angry faces as the target expression, and rated disgust as the most similar relative to the other expression types.

Discussion

In the present study, we show that a computer model of facial expression recognition performed comparably to human observers in two critical capacities: 1) the discrimination of distinct basic emotion classes and 2) judgments of the similarity between distinct basic emotions. With respect to the latter, the similarity space in the model was driven entirely from visual input, without any inferences about the meaning of an expression or the similarity of one emotion to another. Without direct training or implementation in the model, expression similarity was found to be a secondary consequence of training to discriminate between basic emotions. Thus, the judgment of similarity in affective experience across different facial expressions requires no explicit understanding of emotions or their relations. For instance, the observation that individuals expressing disgust may portray feelings of anger but little happiness can be computed from their similarity in high dimensional visual feature space. This emotional comparison does not necessarily require an appreciation of their similarity in somatic space (Adolphs et al 2000; Anderson & Phelps, 2000; Damasio, 1994; 1996), nor does it entail accessing linguistic or conceptual descriptions of the relation between different expression types (e.g., Russell, 1991; Shaver, Schwartz, Kirson, & O'Connor, 1987).

Recognition of facial expressions and the phenomenology of internal affective states have been characterized in two seemingly incompatible ways. The notion of a primary set of distinct basic emotion types (Ekman, 1992) has been contrasted with lower dimensional accounts, whereby emotions fall in specific locations in lower order affective space, and thus are better characterized as varying along underlying continua, such as valence, arousal, or dominance (e.g., Russell & Mehrabian, 1977; Schlosberg, 1952). A few recent studies have systematically compared human and computer performance on recognition of facial expressions using holistic neural network models (Dailey et al, 2000; Dailey et al, 2002). These studies show how both continuous and discrete-like representational performance can coexist in the same holistic network model. In particular, it is shown that categorical perception of facial expressions can occur in a distributed architecture. In addition, these models not explicitly trained to exhibit continuous dimension-like performance can nevertheless capture aggregate human similarity judgments. Thus, distributed models can capture both the continuous and categorical nature of expression recognition. However, such models that have impressively captured human similarity data have employed an all-or-none teaching signal dependent on the output activity across the single network rather than on independently trained expert classifiers for each category. As such, these models may not necessarily reflect the organization of the human brain where there is good evidence for the existence of distinct neural substrates specialized for recognition of expressions of specific types, such as fear, disgust and anger (Adolphs et al 1994; Calder et al 2004; Phillips et al 1997,1998). To the degree to which human brain data are consistent with the existence of distinct specialized representations, the present study examined whether similarity in judgments can be captured by specialized representations alone. A strong test of whether judgments of similarity can be supported by specificity coding alone is to examine a model based on specialized experts each focused on discriminating a particular facial expression from all others.

Our results demonstrate that judgments of similarity can arise from the patterns of activity across outputs of local expert classifiers, which were trained to optimally discriminate specific target emotions. As such, the present model provides strong evidence that activations across specialized emotional classifiers can support the judgments of expression similarity. We first tested similarity on training set examples, which the computer model was trained to specifically discriminate with very high accuracy. Although never intended to be a model of human affective judgments, the computer model's ratings matched the human data both in terms of ordering on the circumplex and equidistance between neighboring clusters. These results strongly suggest that human-like judgments of similarity naturally arise out of the problem of sculpting categorical boundaries between expression types in order to maximize accuracy, rather than developing internal representations for how emotions relate to each other. Although the computer model presented here makes a case for functional specialization for expression discrimination, it is not intended to address how such functional structure arises. Indeed, functional specialization can emerge in a fully distributed system from the statistical structure of the data. This is supported by a number of computational models (e.g. Linkser, 1988), as well as by neurophysiological studies of plasticity (e.g. Neville & Bavelier, 2000).

The evidence from computational modeling suggests that underlying expression similarity can be achieved by superficial visual feature analysis. That facial displays of basic emotions are not entirely independent, but portray related states, may then simply depend upon shared component features (Scherer,1984; 1988). Visual analysis of feature overlap would then be sufficient to capture the relations between emotions. This can be interpreted as evidence against facial expression recognition depending on computations of the similarity in underlying emotional/somatic activity across facial expression types (Adolphs et al, 2000). Facial feedback theories of emotional experience suggest configurations of the face play an important role in emotional experience (see Adelman and Zajonc, 1989). Similar feeling states between two emotions would be mirrored in facial efference (Adelman and Zajonc, 1989). This correspondence in the activation of specific facial muscles would result in visual similarity (Dailey et al, 2002). As such, the present study does not argue against human observers extracting underlying emotional-somatic representations from facial displays; rather, these results are consistent with subjectively similar emotions depending on objectively (i.e., structurally) similar facial displays produced by underlying facial musculature.

Computational and neural representations of emotion recognition

In contrast to studies of emotion experience, where there is neural evidence supporting the existence of emotional dimensions such as approach-withdrawal, valence, and emotional intensity (e.g., Anderson & Sobel, 2003; Davidson, 1995), support for dimensional correlates in facial expression recognition is limited (but see Anderson et al, 2000). Neuroimaging and neuropsychological data demonstrating neural representations selective for distinct expression classes including fear, anger, and disgust are on the surface most consistent with the existence of a set of primary or basic emotions supported by discrete neural substrates. The amygdala is implicated in numerous studies as a crucial component of fear recognition relative to other expressions (e.g., Russell, et al, this issue; Ashwin et al, this issue; Adolphs et al, 1999; Anderson et al, 2003; Philips et al, 1998). In contrast, disgust expressions maximally activate the anterior insula (Anderson et al, 2003; Philips et al, 1998), and patient studies have implicated a basal ganglia-insula system in disgust recognition dysfunction in Parkinson’s and Huntington’s diseases (Suzuki et al, 2006; Hennenlotter et al, 2004). Anger recognition may involve the ventral striatum (Calder et al, 2004) and deficits in anger recognition have been linked to Parkinson’s disease (Lawrence et al, this issue). These data provide strong evidence consistent with local accounts of brain information processing, suggesting that facial expression recognition is supported by distinct expert systems that process specialized information and result in selective deficits when damaged (e.g., Downing, Jiang, Shuman, & Kanwisher, 2001; Kanwisher, McDermott, & Chun, 1997).

On the other hand, distributed accounts of brain function argue that representations are patterns shared across brain areas (e.g., Haxby et al, 2001). The degree to which a particular individual perceives anger and disgust in the same expression, or detects similarity between sadness and fear, may reflect the combinatorial response across distinct expert neural classifiers. Consistent with this view, studies have also shown that regions “specialized” for a specific facial expression also demonstrate reliable responses to other expressions. For instance, regions of the anterior insula responsive to disgust are also responsive to fear in faces (Anderson et al, 2003; Morris et al, 1998), and conversely, the amygdala can reveal robust responses to expressions of disgust (Anderson et al, 2003), anger (Wright, Martis, Shin, Fischer, & Rauch, 2002) and sadness (Blair et al., 1999). Although a brain region may be maximally responsive to a specific emotion, these non-maximal responses to other expressions may have important functional significance for expression recognition. The present computational model suggests that specialized representations of basic emotions classes can support dimension-like gradients of similarity when magnitude of activation across neural local experts is considered. Thus, specialized representations are not antithetical to dimensional-like performance, but represent two compatible modes of information representation. Such combinatorial coding across neural classifiers allows the simultaneous maintenance of discrimination attributed to basic emotions theories and similarity/generalization attributed to dimensional theories. Consistent with this combinatorial coding hypothesis, patients with selective impairments in facial expression recognition following amygdala damage maintain a largely intact capacity to judge similarity between expression classes (Anderson et al, 2000; Hamann & Adolphs, 1999), which may result from the profile of activation across the remaining spared neural classifiers. These profiles, whether facial, auditory, or somato-visceral, may be integrated in a convergence zone, such as the right somatosensory cortices (Adolphs et al, 2000). Contrary to the emotion specific impairments described above, lesions of the right somatosensory cortices result in more global facial expression recognition impairments. According to this hypothesis, in contrast to lesions of expert classifiers, we would predict lesions of this region to be particularly harmful to judgments of expression similarity.

Another view is that the expert classifiers do not represent entire facial expressions configurations, but important subcomponents of expressions. A crucial aspect of the computer model in this study is the common visual Gabor feature layer shared by all expert SVM classifiers. The computational evidence for feature overlap between expressions implies that the dimensions on which facial expressions are related are literally sets of visual features that are themselves important for discrimination. The Component Process Model of Emotion (CPM) (Scherer,1984; 1988; 2001) emphasizes that expression configurations are composed of subunits, with component appraisals such as novelty detection being associated with specific physical features of the face (e.g., eye opening) that may be common across basic expressions (e.g., fear and surprise). Recent studies supporting the importance of feature components to facial expression recognition suggest that the amygdala is not critical for the entire expression configuration but serves primarily as a detector of eye opening (Whalen et al, 2004; Adolphs et al, 2005). Vuilleumier and colleagues (2003) showed in an fMRI experiment that the amygdala response to fearful expressions is greatest for low spatial frequency components, which may preferentially encode the presence of eye whites. In combination with neural and behavioral evidence, emergent similarity in the computer model as a consequence of overlapping features indicates that expression recognition in the brain may depend on detecting important component features, such as eye opening (e.g., in surprise and fear), and not basic emotion prototypes or dimensions such as valence/arousal.

Deconstructing idealized discrimination performance

One question that has not been addressed well by either basic emotion or circumplex accounts of subject ratings is whether aggregate subject discrimination performance is characteristic of individual subject performance. Comparing human and computer accuracy scores for the six emotion ratings revealed that the computer model generally matches the mean performance trend of humans. However, for untrained exemplars, the computer model demonstrated lower forced choice discrimination accuracy for fear expressions relative to the other expression types. This is consistent with fear recognition being least accurate in human observers. The MDS analysis performed on standardized subject ratings indicates that subjects are not a homogenous group in terms of discrimination errors. There are substantial individual differences in facial expression recognition (Elfenbein et al, 2002) as well as differences in neural response across individuals (Canli et al, 2002). Our analyses of individual differences revealed that some well-defined clusters arise, suggesting that various groups of subjects may share rating patterns (e.g., some subjects perform lower recognizing fear due to similarity with surprise while others perform lower on disgust due to similarity with anger). These individual differences are consistent with a CPM account of facial emotion recognition, as an individual can attend to certain features and ignore others within a facial configuration, resulting in predictable patterns of facial expression confusion.

The computer model was found to make similar discrimination errors to the cluster of subjects characterized by a relatively selective difficulty with fear. Thus, while comparing accuracy for the computer model to aggregate human accuracy suggests that the computer model performs atypically on fear, this comparison is flawed by benchmarking the computer model against an idealized human observer based on average subject performance. A more detailed analysis reveals that the model performs similarly to specific subgroups of human observers. There may be important individual differences in how humans recognize facial expression that are often glossed over in treatment of group mean data alone. To the degree that computer simulations capture human performance, it may be argued that an appropriate index is their mirroring of how specific individual human observers categorize rather than their capacity for modeling aggregate behavior.

The current experiment relied on facial expression datasets coded and tested within a basic emotions theoretical framework (Ekman & Friesen, 1976; Kanade et al, 2000). It remains possible that assumptions made in this framework bias exploration of actual emotion space. Exploring emotion space by training computer models on exemplars of spontaneous expression data that have not been coded into basic emotions may provide a different picture of facial expression behavior and the representations underlying their recognition. Further, the present model is context independent, relying solely on facial features for recognition. The present study suggests that significant statistical regularity of image features across expressions types allows for recognition of expression similarity and distinctiveness. Recent work nevertheless emphasizes the role of temporal and spatial situational context in interpreting facial expressions.

For example, a cropped image of a face of an Olympian gold medal winner at the podium may portray extreme grief but will resemble extreme happiness with the context of the scene incorporated (Fernandez-Dols & Carroll, 1997). Such context dependence suggests an understanding of the full range of human competence in emotional communication cannot be characterized by statistical regularities in image structure alone.

Acknowledgments

We thank Maha Adamo for ideas leading to individual subject analyses of expression ratings and Iris Gordon for providing experiment programming. This work was supported by the National Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chairs Program. Partial funding was also provided by National Science Foundation grants NSF IIS-032987, NSF IIS-0220141, and CNS-0340851, in addition to a University of California Discovery Grant.

References

Adelman, P., & Zajonc, R. (1989). Facial efference and the experience of emotion. Annual Review of Psychology, 40, 249–280.

Adolphs R., Damasio H., Tranel D., Cooper G., & Damasio A. R. (2000). A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *J. Neurosci. 20*(7):2683-90.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*(7021), 68-72.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. R. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature, 372*, 669-672.

Adolphs, R., Tranel, D., Hamann, S., Young, A. W., Calder, A. J., & Phelps, E. A., et al. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia, 37*(10), 1111-1117.

Anderson, A. K., Christoff, K., Panitz, D., De Rosa, E., & Gabrieli, J. D. (2003). Neural correlates of the automatic processing of threat facial signals. *The Journal of Neuroscience, 23*(13), 5627-5633.

Anderson, A. K., & Sobel, N. (2003). Dissociating intensity from valence as sensory inputs to emotion. *Neuron, 39*(4), 581-583.

Anderson, A. K., Spencer, D. D., Fulbright, R. K., & Phelps, E. A. (2000). Contribution of the anteromedial temporal lobes to the evaluation of facial emotion. *Neuropsychology, 14*(4), 526-536.

Anderson AK, Phelps EA. (2000). Perceiving emotion: There's more than meets the eye. Current Biology, 10(15):R551-4.

Ashwin, C., Baron-Cohen, S., Wheelwright, S., O'Riordan, M., & Bullmore, E.T. (This Issue). Differential activation of the amygdala and the 'social brain' during fearful face processing in Asperger Syndrome. *Neuropsychologia*

Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., & Movellan, J. (2004). Machine learning methods for fully automatic recognition of facial expressions and facial actions. *IEEE international conference on systems, man & cybernetics*, The Hague, Netherlands, 592-597.

Beaupré, M. G., & Hess, U. (in press). Cross-cultural emotion recognition among canadian ethnic groups. *Journal of Cross-Cultural Psychology.*

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., & Kudoh, T. et al. (1997). Matsumoto and Ekman's japanese and caucasian facial expressions of emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*, 2-21.

Blair, R. J., Morris, J. S., Frith, C. D., Perrett, D. I., & Dolan, R. J. (1999). Dissociable neural responses to facial expressions of sadness and anger. *Brain; a Journal of Neurology, 122 (Pt 5)*(Pt 5), 883-893.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121-167.

Calder, A. J., Keane, J., Lawrence, A. D., & Manes, F. (2004). Impaired recognition of anger following damage to the ventral striatum. *Brain, 127*(9):1958-69.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3*(2), 81-117.

Canli T, Sivers H, Whitfield SL, Gotlib IH, Gabrieli JD. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296(5576):*2191.

Dailey, M. N., Cottrell, G. W., & Adolphs, R. (2000). A six-unit network is all you need to discover happiness. *Proceedings of the 22nd annual cognitive science conference,* Philadelphia, PA, *22* 101-106.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience, 14*(8), 1158-1173.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351*(1346), 1413-1420.

Damasio, A. R. (1994). *Descartes' error*. New York: Putnam.

Darwin, C. (1872). *The expression of the emotions in man and animals*. New York: D. Appleton and Company.

Davidson, R. J. (1994). Asymmetric brain function, affective style and psychopathology: The role of early experience and plasticity. *Development and Psychopathology, 6*, 741-758.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470-2473.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3-4), May-Jul 1992, pp. 169-200.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation, 1971* (pp. 207-283). Lincoln: University of Nebraska Press.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, Calif.: Consulting Psychologists Press.

Ekman, P., & Friesen, W. (1976). In Ekman P. (Ed.), *Pictures of Facial Affect*. San Francisco, CA 94143: Human Interaction Laboratory, UCSF, HIL-0984.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124-129.

Elfenbein, H. A., Marsh, A., & Ambady, N. (2002). Emotional Intelligence and the recognition of emotion from the face. In L. F. Barrett & P. Salovey (Eds.), The wisdom of feelings: Processes underlying emotional intelligence (pp. 37-59). New York: Guilford Press.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227-240.

Fasel, I., Dahl, R., Hershey, J., Fortenberry, B., Susskind, J. & Movellan, J. (2004). *The machine perception toolbox.* Retrieved 10/07, 2004 from http://mplab.ucsd.edu/projects-home/project1/free-software/MPTWebSite/API/index.html

Fasel, I., Fortenberry, B., & Movellan, J. (2005). A generative framework for real time object detection and classification. *Computer Vision and Image Understanding, 98*(1), 182-210.

Fernandez-Dols, J. M., & Carroll, J. M. (1997). Is the meaning perceived in facial expression independent of its context? In J. A. Russell, & J. M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 275-294)Cambridge University Press.

Freeman, P. R. (1970). Optimal bayesian sequential estimation of the median effective dose. *Biometrika*, *57*, 79-89.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28*(2), 337-407.

Hamann SB, Adolphs R. (1999). Normal recognition of emotional similarity between facial expressions following bilateral amygdala damage. Neuropsychologia, 37(10):1135-41.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Hennenlotter, A., Schroeder, U., Erhard, P., Haslinger, B., Stahl, R., & Weindl, A. et al. (2004). Neural correlates associated with impaired disgust processing in pre-symptomatic huntington's disease. *Brain; a Journal of Neurology, 127*(Pt 6), 1446-1453.

Kanade, T., Cohn, J. F., & Tian, Y. L. (2000). Comprehensive database for facial expression analysis. *Proceedings of the 4th IEEE international conference on automatic face and gesture recognition (FG'00),* 46-53.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 17*(11), 4302-4311.

Keltner, D., & Shiota, M. N. (2003). New displays and new emotions: A commentary on rozin and cohen (2003). *Emotion, 3*(1), 86-91; discussion 92-6.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., Malsburg, C. v. d., & Wurtz, R. P. et al. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42*(3), 300-311.

Lawrence, A. D., Goerendt, I. K., & Brooks, D. J. (This Issue). Impaired Recognition of Facial Expressions of Anger in Parkinson's Disease Patients Acutely Withdrawn from Dopamine Replacement Therapy. *Neuropsychologia*

Linkser, R. (1988). Self-organization in a percentual network. *IEEE Computer, 21*, 105-117.

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2004). Dynamics of facial expression extracted automatically from video. *IEEE conference on computer vision and pattern recognition*, (Workshop on Face Processing in Video)

Morris, J. S., Friston, K. J., Buchel, C., Frith, C. D., Young, A. W., & Calder, A. J. et al. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain; a Journal of Neurology, 121 (Pt 1)*(Pt 1), 47-57.

Movellan, J. (2002). *Tutorial on gabor filters* (http://mplab.ucsd.edu/tutorials/pdfs/Gabor.pdf).

Neville, H. J., & Bavelier, D. (2000). Specificity and plasticity in neurocognitive development in humans In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 83-98). Cambridge, MA: MIT Press.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., & Giampietro, V. et al. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 265*(1408), 1809-1817.

Phillips, M.L., Young, A.W., Senior, C., Brammer, M., Andrew, C., Calder, A.J., Bullmore, E.T., Perrett, D.I., Rowland, D., Williams, S.C., Gray, J.A., & David, A.S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*(6650):495-8.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion* (pp. 3-33). New York: Academic.

Rapcsak SZ, Galper SR, Comer JF, Reminger SL, Nielsen L, Kaszniak AW, Verfaellie M, Laguna JF, Labiner DM, Cohen RA. (2000). Fear recognition deficits after focal brain damage: a cautionary note. *Neurology, 54(3):*575-81.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin, 110*(3), 426-450.

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

Russell, J. A., & Bullock, M. (1985). Multidimensional scaling of emotional facial expressions: Similarity from preschoolers to adults. *Journal of Personality and Social Psychology*, *48*, 1290-1298.

Russell, J. A., & Bullock, M. (1986). Fuzzy concepts and the perception of emotion in facial expressions. *Social Cognition*, *4*(3), 309-341.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*, 273-294.

Russell, T.A., Reynaud, E., Pietura, K., Ecker, C., Benson, P.J., Zelaya, F.,Giampietro, V., Brammer, M., David, A. & Phillips, M.L. (This Issue). Neural responses to dynamic expressions of fear in schizophrenia. *Neuropsychologia*

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92-120). New York and Oxford: Oxford University Press.

Scherer, K. R. (1988). Criteria for emotion-antecedent appraisal: A review. In V. Hamilton, G. H. Bower & N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 89-126). Dordrecht: Kluwer.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer, & P. Ekman (Eds.), *Approaches to emotion* (pp. 293-317). Hillsdale, NJ: Erlbaum.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44*, 229-237.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*(6), 1061-1086.

Suzuki, A., Hoshino, T., Shigemasu, K., & Kawamura, M. (2006). Disgust-specific impairment of facial expression recognition in parkinson's disease. *Brain; a Journal of Neurology, 129*(Pt 3), 707-717.

Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley and Sons, Inc.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *Int.J.Comput.Vision, 57*(2), 137-154.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience, 6*(6), 624-631.

Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., & Polis, S. et al. (2004). Human amygdala responsivity to masked fearful eye whites. *Science, 306*(5704), 2061.

Wright, C. I., Martis, B., Shin, L. M., Fischer, H., & Rauch, S. L. (2002). Enhanced amygdala responses to emotional versus neutral schematic facial expressions. *Neuroreport, 13*(6), 785-790.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*(3), 271-313.

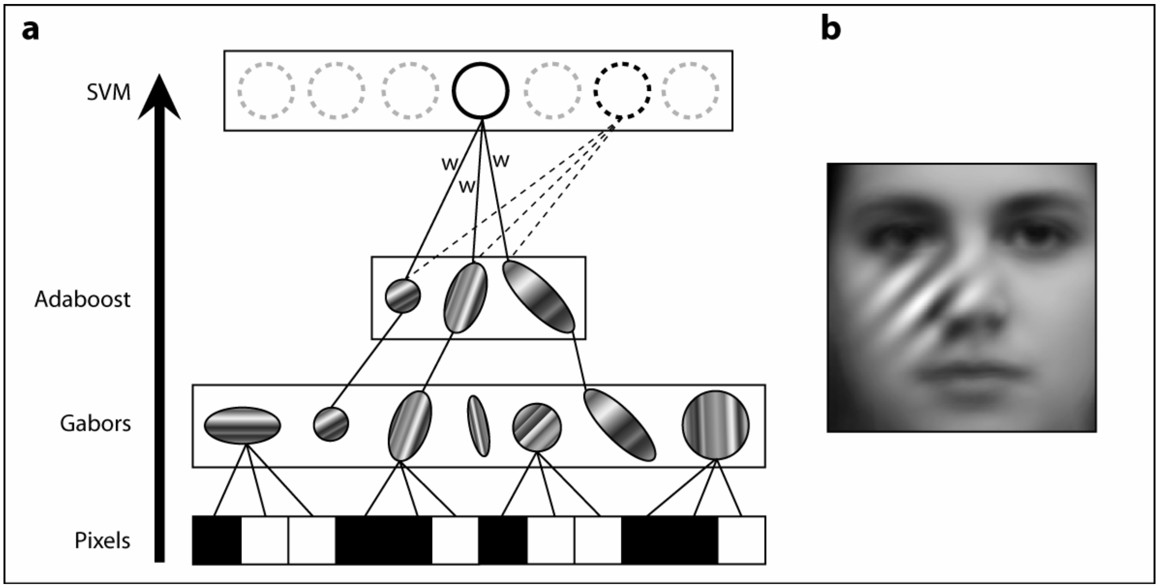

Figure 1. a) Diagram of support vector machine computer model showing the flow of information processing through multiple representational layers: from the pixel level to preprocessing by Gabor feature selection with AdaBoost and ending with seven-output SVM classification. SVM classifications represent 6 emotion types plus neutral. Separate weight connections for two SVM classifiers are depicted in black and dashed lines. b) Example of a Gabor filter projected onto a face image.

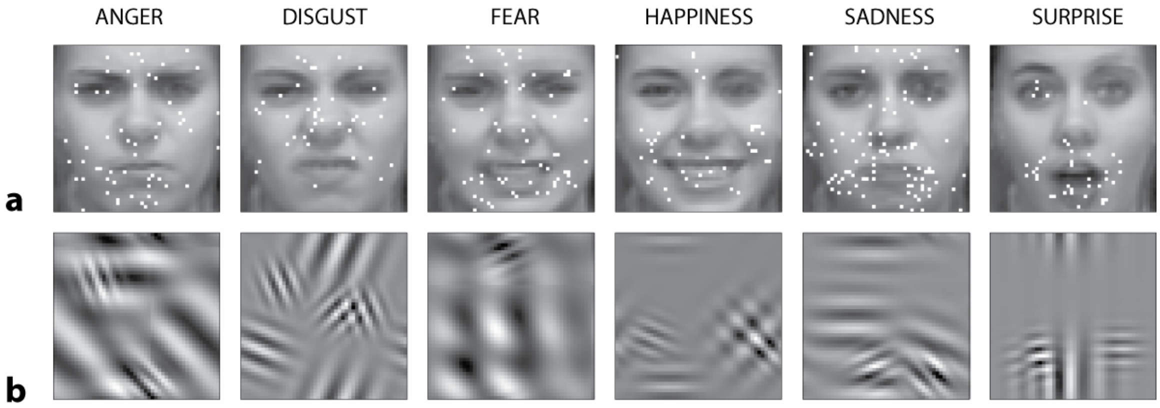

Figure 2. Illustration of Gabor features selected for each expression. a. Center locations of the first 50 Gabor features are indicated by white dots. b. Receptive fields of the first 10 Gabor features projected onto image space showing the preferred spatial frequency, orientation, and location.

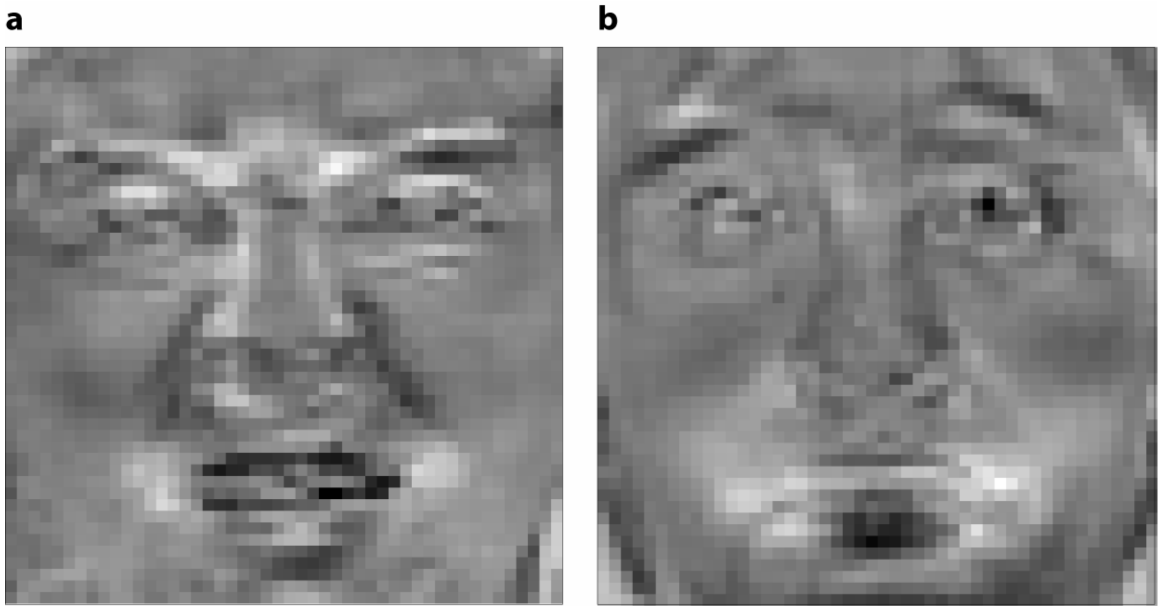

Figure 3. In order to visualize the weights, this figure employed a linear SVM trained directly on the image pixels rather than the Gabor representations. The weights shown in this figure were trained to discriminate two specific emotions, a) anger versus disgust and b) sadness versus fear. Positive weights are shown in white and negative weights in black.

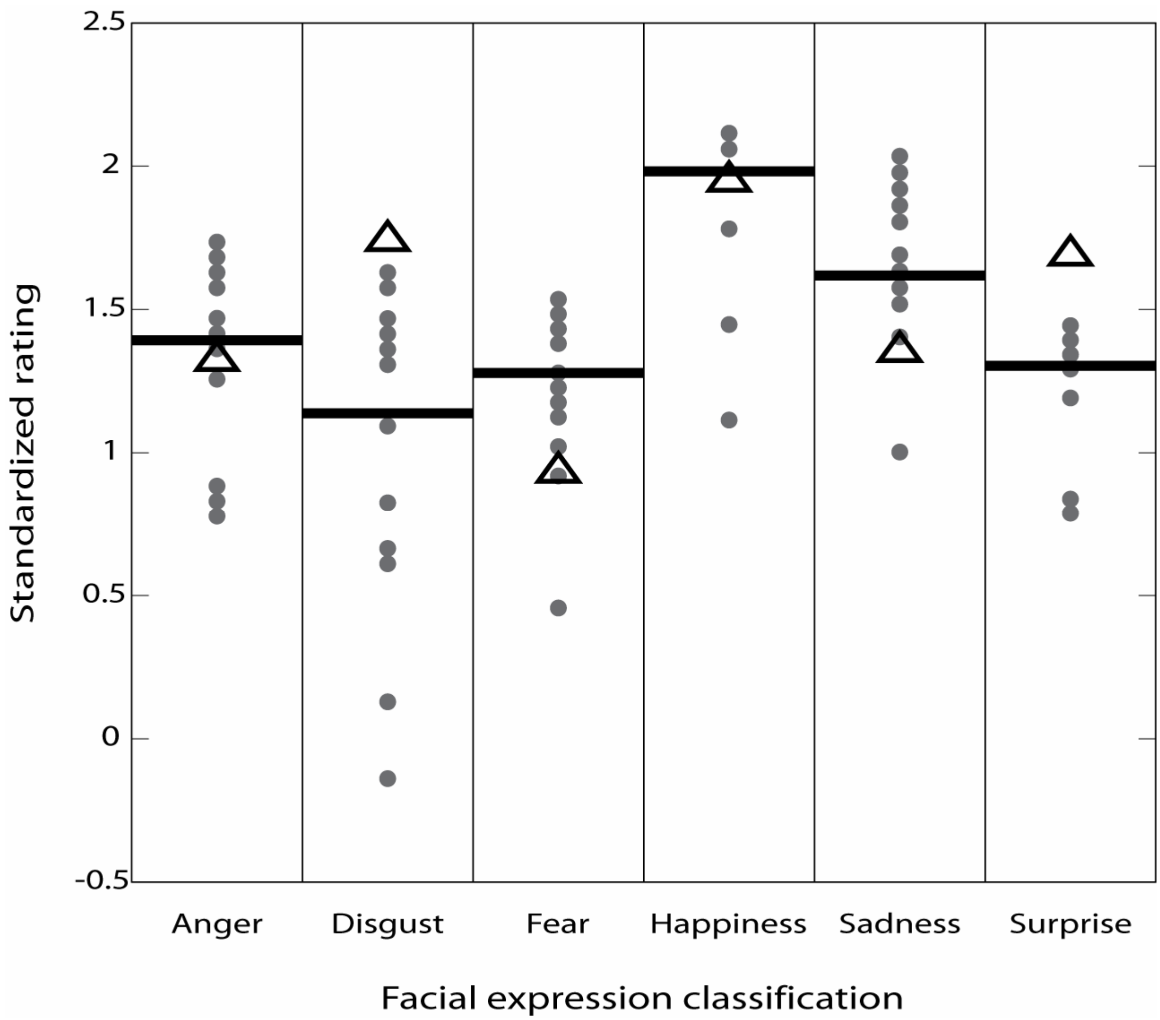

Figure 4. Standardized target emotion ratings (e.g. anger ratings for angry faces) for human subjects and SVM activations for the computer model averaged over exemplars. Means for each subject are plotted as points and the overall human subject mean is represented by a horizontal line. Mean standard ratings for the computer model are indicated by a triangle.

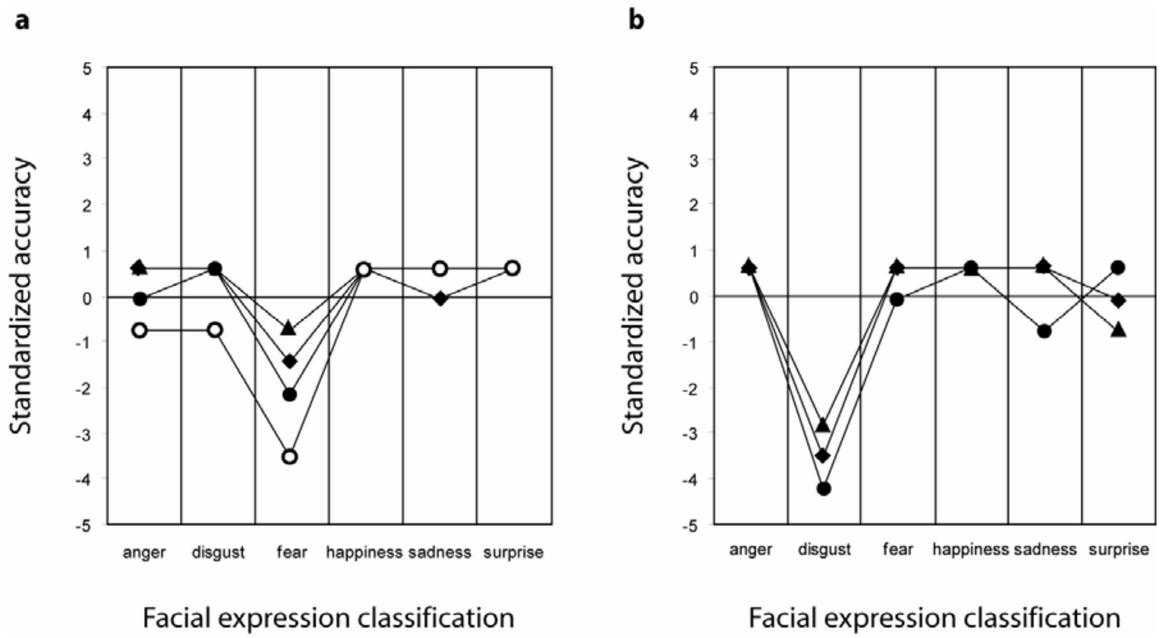

Figure 5. Target emotion forced choice accuracy for two clusters of human subjects identified by MDS. Each human subject is represented by a different filled shape. a) Depicts subjects that consistently rate fear lower than the other expressions. The computer model (open circles) fits this accuracy pattern. b) Unlike the model are subgroups of subjects who consistently rate disgust lower than the other expressions.

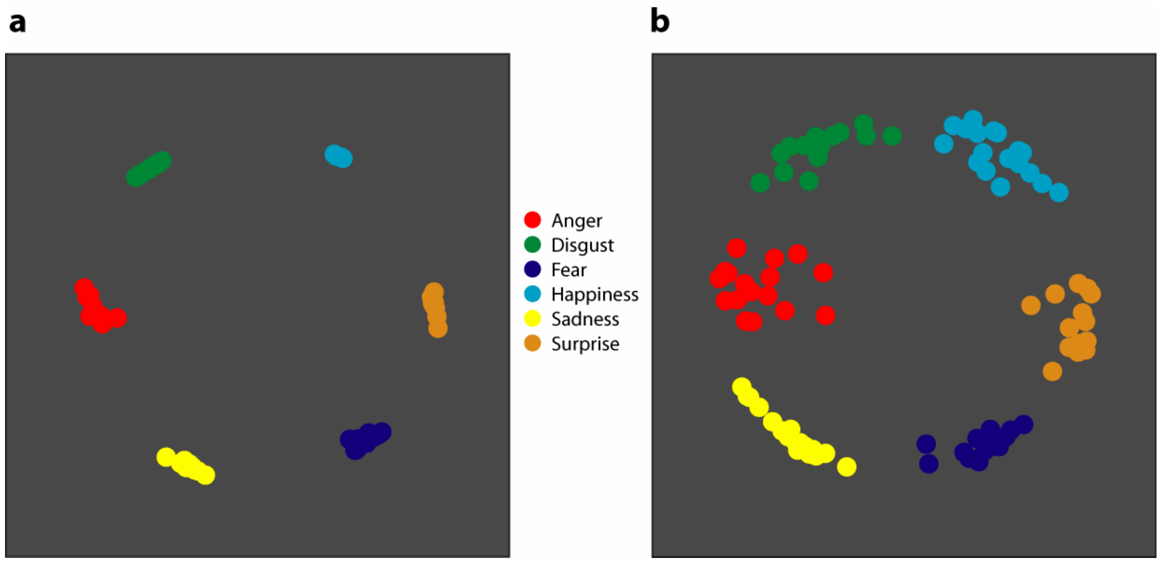

Figure 6. MDS plots of similarity between exemplars of different emotions from the POFA training dataset. a) Human rating norms. b) Computer model activations.

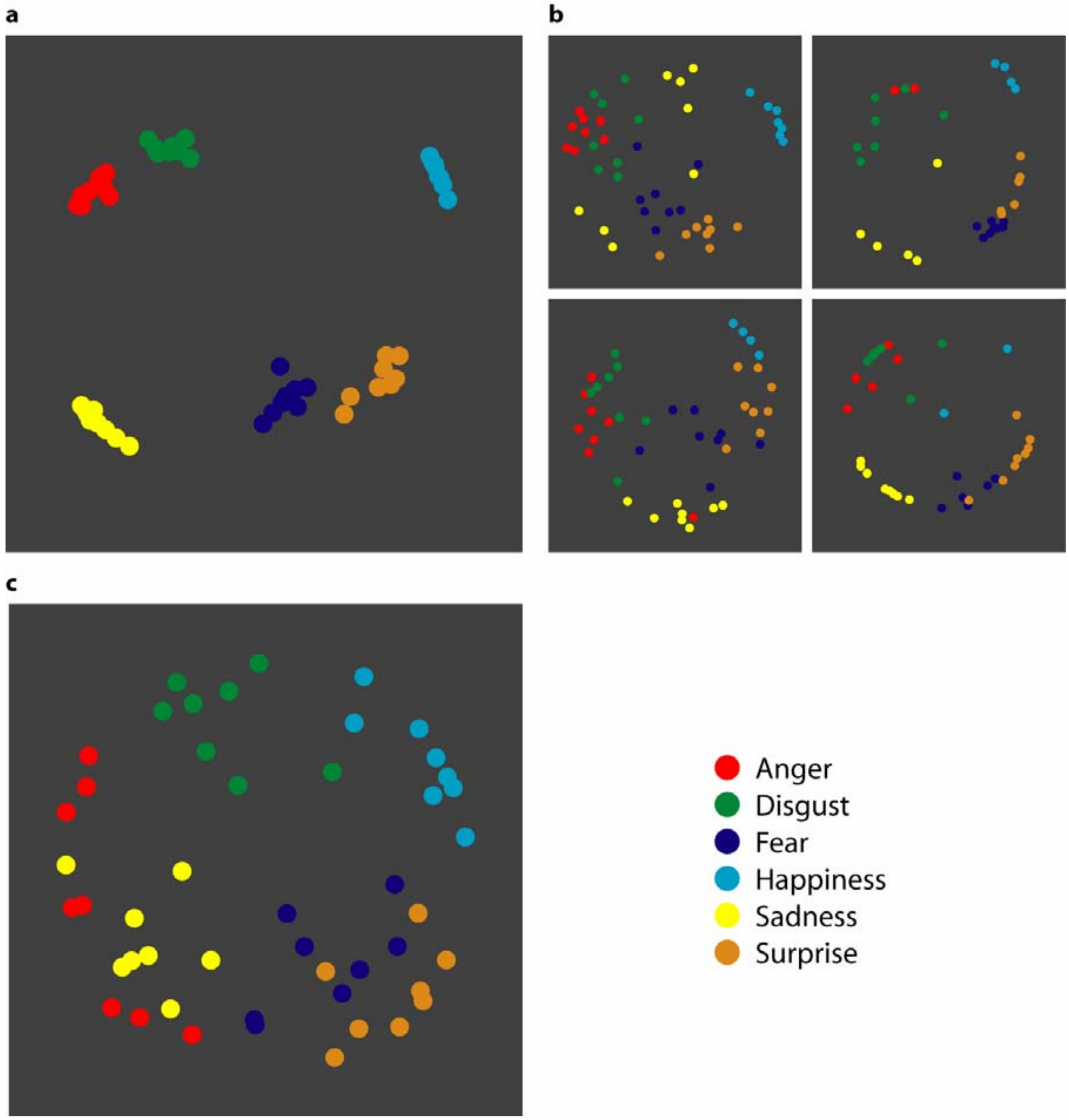

Figure 7. MDS plots of similarity between exemplars of different emotions from the JACFEE dataset. a) Human ratings averaged across all 23 subjects. b) Human ratings for subjects in two characteristic clusters of subject rating patterns (see Figure 5). The first column shows ratings for two subjects with low accuracy for fear. The second column

shows ratings for two subjects with low accuracy for disgust. c) Computer model activations.

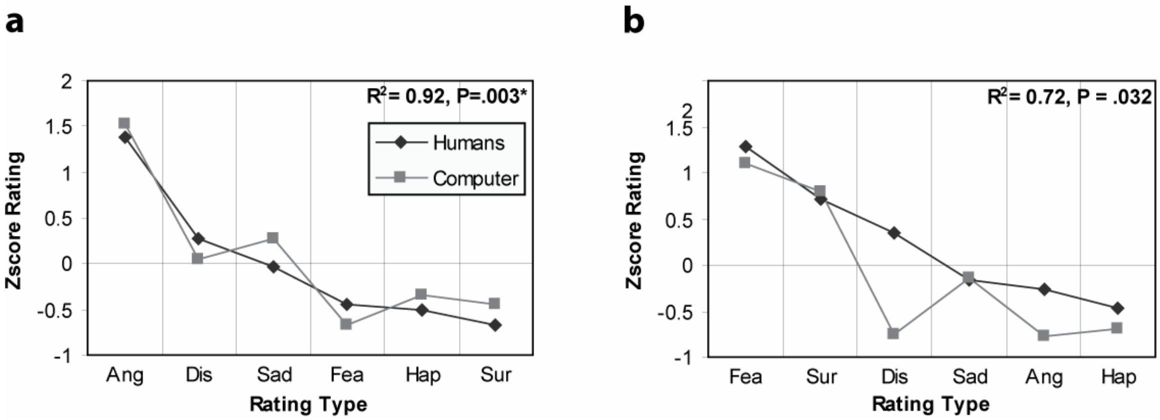

Figure 8. Comparison of human and computer rating profiles. a) Profile comparison averaged over anger exemplars. b) Profile comparison averaged over fear exemplars. The x-axis is rank-ordered by human ratings and thus the label order in (a) and (b) differ.