Neuropsychologia 45 (2007) 152–162
NEUROPSYCHOLOGIA
www.elsevier.com/locate/neuropsychologia

Human and computer recognition of facial expressions of emotion

J.M. Susskind [a], G. Littlewort [b], M.S. Bartlett [b], J. Movellan [b], A.K. Anderson [a,c,*]

[a] *Department of Psychology, University of Toronto, Canada*
[b] *Machine Perception Laboratory, Institute of Neural Computation, University of California, San Diego, United States*
[c] *Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, Ont. M6A 2E1, Canada*

Available online 12 June 2006

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 SVMs were trained to develop fully automatic optimal recognition of one of six 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.

Keywords: Computational model; Expression similarity; Discrimination; Specialized experts; Support vector machines; Visual features

1. 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 that had 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 non-linearities consistent with categorical perception, implying the existence of discrete expression categories (Calder, Young, Perrett, Etcoff, & Rowland, 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 but leaves 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, Christoff, Panitz, De Rosa, & Gabrieli, 2003; Phillips et al., 1998). Similarly, discrete neural representations have recently been proposed for recognition of anger in the ventral striatum (Calder, Keane, Lawrence, &

* Corresponding author at: Department of Psychology, University of Toronto, 100 St. George Street, Toronto, Ont. MS5 3G3, Canada. Tel.: +1 416 946 0207; fax: +1 416 978 4811.
E-mail address: anderson@psych.utoronto.ca (A.K. Anderson).

0028-3932/$ – see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.neuropsychologia.2006.05.001

Manes, 2004). Finding such dissociations in recognition for a variety of basic prototypes would provide further evidence for their status as primaries on which emotional experience and communication depend.

The alternative view suggests that emotion space is characterized by lower order dimensions, such that emotions are fuzzy categories clustered on axes such as valence, arousal, or dominance (Russell, 1980; Russell & Bullock, 1986; Schlosberg, 1952). Thus 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 some emotion types (e.g., anger and disgust) are closer than others (e.g., 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 emotion 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 the extent to which recognition of emotional expressions directly reflects the statistical structure of the images to which humans are exposed. 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 et al., 2004). 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 is 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 of a small number of underlying dimensions. We addressed this issue in the present study by comparing human behavioural data to a computer model that was trained to make a seven-way forced choice between basic expressions plus neutral faces. 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, Cottrell, Padgett, & Adolphs, 2002), SVMs learn an optimal decision boundary between two labeled classes by focusing on difficult training examples (Burges, 1998). 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.

2. Methods

2.1. Computer model details

The system we tested was developed at UC San Diego's Machine Perception Laboratory (Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2004). The software is currently distributed as part of the MPT/MPTX library (available online at http://www.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 to fit it to human perceptual data.

The computer model (see Fig. 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, Fortenberry, and Movellan (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 320×240 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 96×96 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 eight orientations and five 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). Fig. 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° (Movellan, 2002). Converting face images to Gabor magnitudes results in image