Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Psychology

Visual perception of facial expressions of emotion

Aleix M Martinez

Facial expressions of emotion are produced by contracting and relaxing the facial muscles in our face. I hypothesize that the human visual system solves the inverse problem of production, that is, to interpret emotion, the visual system attempts to identify the underlying muscle activations. I show converging computational, behavioral and imaging evidence in favor of this hypothesis. I detail the computations performed by the human visual system to achieve the decoding of these facial actions and identify a brain region where these computations likely take place. The resulting computational model explains how humans readily classify emotions into categories as well as continuous variables. This model also predicts the existence of a large number of previously unknown facial expressions, including compound emotions, affect attributes and mental states that are regularly used by people. I provide evidence in favor of this prediction.

Address
The Ohio State University, Columbus, OH 43201, USA

Corresponding author: Martinez, Aleix M (martinez.158@osu.edu)

Current Opinion in Psychology 2017, **17**:27–33

This review comes from a themed issue on **Emotion**

Edited by **Batja Gomes de Mesquita** and **Lisa Feldman Barrett**

http://dx.doi.org/10.1016/j.copsyc.2017.06.009

Introduction

Researchers generally agree that human emotions correspond to the execution of a number of computations by the nervous system. Some of these computations yield facial muscle movements, called Action Units (AUs) [1]. Specific combination of AUs defines facial expressions of emotion, which can be visually interpreted by observers.

Here, I hypothesize that the human visual system solves the inverse problem of production, that is, the goal of the visual system is to identify which AUs are present in a face. Crucially, I show how solving this inverse problem allows human observers to effortlessly infer the expresser's emotional state.

This hypothesis is in sharp contrast to the categorical model, which assumes that the visual system identifies emotion categories rather than AUs from images of facial expressions, Figure 1. The categorical model propounds that our visual system has an algorithm aimed to categorize facial expressions of emotion into a small number of canonical expressions [2]. This model has, in recent years, included six emotion categories: happiness, surprise, anger, sadness, disgust and fear [3]. The claim is that the visual system knows which image features code for each one of these emotion categories, allowing us to interpret the expresser's emotion [4].

A major problem with the categorical model is its inability to provide a fine-grained definition of the expresser's emotion, beyond the six canonical expressions listed above [5••]. Also, and crucially, the search for the brain's region of interest (ROI) or ROIs responsible for the decoding of these emotion categories has come up empty [6•,7]. This has prompted researchers to propose alternative models [8–10]. These models suggest that, rather than emotion categories, facial expressions transmit either continuous variables, such as valence and arousal, or affective attributes and mental states, such as dominance and worry.

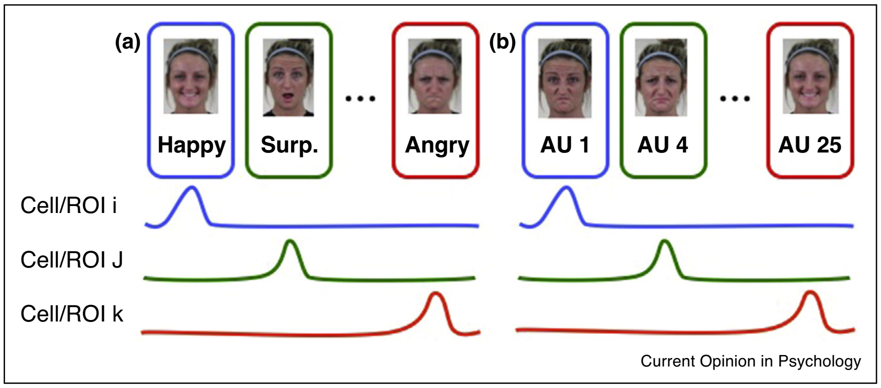
Which is the correct model? This paper provides converging computational, behavior and imaging evidence in support of the hypothesis that the visual system is tasked to decode AUs from face images, Figure 1b. I show that once AUs have been successfully decoded from faces, the brain can effortlessly extract high-level information, including canonical and fine-grained emotion categories (e.g., disgusted and happily disgusted), continuous affect variables (e.g., valence and arousal), and affect attributes and mental states (e.g., dominance and worry).

Visual recognition of action units

Which are the computations performed by the human visual system to decode AUs? Facial muscles are hidden under our skin and are, hence, not directly visible to us. The human visual system needs to infer their activation from observable image features.

When we move our facial muscles, the distances between major facial components (chin, mouth, nose, eyes, brows, and so on) change. For example, when people produce a prototypical facial expression of anger, the inner corners of their brows lower (which is labeled AU 4), their lids tightened (AU 7) and their upper and lower lip press against one another (AU 24). If you practice these movements in front of a mirror, you will see that the distance between the inner corners of your brows and mouth decreases and that your face widens. Conversely, when creating a prototypical facial expression of sadness, the

Figure 1



(a) The categorical model posits there must be a group of cells, region of interest (ROI), ROIs or brain networks that differentially respond to specific emotion categories. **(b)** The model proposed in the present paper postulates the existence of an ROI dedicated to the decoding of Action Units (AUs) instead. That is, cells in this ROI decode the presence of AUs, not emotion category.

combination of AUs (1, 4 and 15) leads to a larger than normal distance between brows and mouth and a thinner face. These second-order statistics (i.e., distance variations) are called configural features.

We have shown that these configural features are extremely accurate when used to visually detect the activation of AUs in images [2,11••]. For example, activation of AUs 4 and 24 can be successfully detected with 100% accuracy using a single configural feature — the distance between the inner corners of the brows and mouth (Supplementary Material). But, this algorithm sometimes assumes AUs are active when they are not, that is, a false positive. This happens when we observe someone who has a brow to mouth distance significantly shorter than the majority of people.

This effect is illustrated in Figure 2. The left image is consistently perceived as expressing sadness by human subjects. The right image is consistently categorized as expressing anger. But these images correspond to neutral expressions, that is, a face that does not display any emotion [11••,12]. Why then do we perceive emotion in them? Because our visual system assumes that AUs 1 and 15 on the left image and AUs 4 and 24 on the right image are active. The visual system reaches this conclusion because the configural features that define these AU activations are present in the image. This effect overgeneralizes to other species and drawings of facial expressions as shown in Figure S1 and S2, that is, we anthropomorphize.

Of course, very few people have such an uncanny distribution of facial components on their faces and, hence, the number of false positives is small. Furthermore, the brain can use contextual information to correct some, if not most, of them.

Computational model

The configural features described in the preceding section define the dimensions of the proposed computational model, Figure 3. Note that this model is norm-based. That is, the perception of AU intensity increases with the degree of activation, since this increases/decreases the value of the corresponding configural feature [11••].

But, why use these image features? Are other shape features better determinants of AU activation? To test this, we performed a computational analysis [5••]. In this study, the shape of all external and internal facial components was obtained. Then, machine learning algorithms were used to identify the most discriminant shape features of AU active versus inactive. The results demonstrated that the configural changes of our model are indeed the most discriminant image features.

Additional proof of the use of these configural features comes from the perception of AU activation and emotion in face drawings and schematics (Figure S2). Furthermore, a simple inversion eliminates the percept; if you rotate Figure 2 180°, the perception of anger and sadness will disappear [12]. This is a well-known consequence of configural processing [13]. Also, computer vision algorithms that use these features attain extremely accurate recognition of AUs (Figure S3).

These results thus support our hypothesis that the visual system solves the inverse problem of production by identifying which AUs construct an observed facial expression. Yet, if this model is correct, there must be a neural mechanism which implements these computations. Indeed, using multivariate pattern analysis on BOLD (blood-oxygen-level dependent) fMRI (functional Magnetic Resonance Imaging), we have identified a small ROI in posterior Superior Temporal Sulcus (pSTS)