Acta Psychologica 132 (2009) 286–292

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Task effects, performance levels, features, configurations, and holistic face processing: A reply to Rossion

Maximilian Riesenhuber *, Brian S. Wolff

Department of Neuroscience, Georgetown University Medical Center, Research Building, Room WP-12, 3970 Reservoir Rd. NW, Washington, DC 20007, USA

ARTICLE INFO

Article history:
Received 11 January 2009
Received in revised form 8 July 2009
Accepted 10 July 2009
Available online 7 August 2009

Keywords:
Face perception
Computational modeling
Inversion effect
Holistic
Object recognition

ABSTRACT

A recent article in *Acta Psychologica* ("Picture-plane inversion leads to qualitative changes of face perception" by Rossion [Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica (Amst), 128*(2), 274–289]) criticized several aspects of an earlier paper of ours [Riesenhuber, M., Jarudi, I., Gilad, S., & Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proceedings of the Royal Society of London B (Supplements), 271,* S448–S450]. We here address Rossion's criticisms and correct some misunderstandings. To frame the discussion, we first review our previously presented computational model of face recognition in cortex [Jiang, X., Rosen, E., Zeffiro, T., Vanmeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using FMRI and behavioral techniques. *Neuron, 50*(1), 159–172] that provides a concrete biologically plausible computational substrate for holistic coding, namely a neural representation learned for upright faces, in the spirit of the original simple-to-complex hierarchical model of vision by Hubel and Wiesel. We show that Rossion's and others' data support the model, and that there is actually a convergence of views on the mechanisms underlying face recognition, in particular regarding holistic processing.

1. Introduction

Faces are an object class of significant interest for many areas of cognitive neuroscience, including object recognition, decision making, social cognition, and perceptual learning. As with many other aspects of cognition, the effortlessness with which most of us perceive faces belies the complexity of the underlying neural processes. Indeed, the richness of behavioral phenomena associated with faces has stimulated much exciting research. Foremost among these behavioral phenomena is the so-called face-inversion effect (Yin, 1969), FIE, referring to the observation that people are substantially worse at discriminating faces presented upside-down than right side-up, whereas inversion usually has less of an impact on the discrimination of objects from other classes. Subsequent research has established that this advantage for upright faces vs. inverted faces might be due to the fact that faces are processed "holistically," i.e., that whole faces are processed more efficiently than their component parts (Tanaka & Farah, 1993).[1]

2. Is face processing "special?

A key question and bone of contention has been whether *quantitative* differences in the recognition of inverted faces relative to upright faces and of isolated face parts relative to the same parts embedded in whole (upright) faces necessarily imply that there need be a *qualitative* difference in the way faces are processed relative to other objects (making faces "special"), or whether face perception can be seen as a particular case of "generic" object recognition, relying on the same kinds of neural mechanisms underlying the recognition also of non-face objects, but refined through extensive experience with a particular object class, namely faces.

One approach to answering this question is to assume that faces are indeed "special" and then try to define qualities that might make faces "special" relative to other object classes. For instance, faces differ from non-face objects in that they usually contain two eyes, a mouth, and a nose, and that there are regularities in how these parts are arranged, e.g., the eyes are usually above the mouth and the nose is in-between, giving rise to theories that face recognition may be based on recognizing individual face parts (usually called "features," see footnote 1) and then computing their "configuration." This raises the question of what the particular face parts are supposed to be and what should make up their "configuration." While early studies (e.g., Haig, 1984) defined "features" as "eyes, mouth, and nose," other studies have used more elaborate

* Corresponding author. Tel.: +1 202 687 9198; fax: +1 202 784 3562.
E-mail address: mr287@georgetown.edu (M. Riesenhuber).

[1] Note that in this paper, we refer to named parts of the face, such as eyes, mouth, and nose, and eyebrows as face "parts" to contrast them to visual "features" a term we use to refer to the preferred stimuli of neurons below the holistic face neuron level (see Fig. 1 and below), which may or may not correspond to face "parts". If we need to refer to "face features" in their traditional sense in the face perception literature, i.e., as eyes, mouth, nose, etc., then we will use the term in quotes.

0001-6918/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2009.07.004

feature sets, e.g., by breaking up the eye "feature" into eyeball and eyebrow (Goffaux & Rossion, 2007). Correspondingly, "configuration" has been defined in a variety of ways, ranging from simple ("the second order spatial configuration" between features (Carey & Diamond, 1986)) to more complicated schemes including up to 29 different distance measurements including eyebrow-hairline distance, lip thickness and inter-nostril separation (Young & Yamane, 1992). While this freedom in the definition of the features and spatial relationships underlying face perception provides a flexible means of describing stimuli and accommodating results obtained in a particular experiment, the fact that many of these part definitions (as well as the spatial measurements) overlap (e.g., changing an eyebrow in an eyeball-eyebrow part-based parameterization changes the "eye" part in a parameterization that does not make the eyeball-eyebrow split, and moving the eyebrow changes the eye-eyebrow *configuration* in the former, but the "eye" *part* in the latter case) has lent an element of arbitrariness to these models, and there is no consensus as to what precisely the "parts" and the "configuration" might be that the human brain according to these models presumably calculates when perceiving faces (Tsao & Freiwald, 2006). A further challenge for "feature"/"configuration" models is that no quantitative neurobiologically plausible computational model has been put forward that can take a photographic face image, calculate face "features" and their arrangement, and generate quantitative predictions for neuronal tuning and behavior. This lack of a computational implementation makes it exceedingly difficult to falsify these verbal "feature"/"configuration" models, considering that face inversion is first and foremost a *quantitative* deficit: while subjects are impaired at discriminating inverted faces, they are generally very well able to do so above chance.

An approach alternative to that of pre-supposing that faces are special and then trying to determine what could make faces special relative to objects from other classes is to start with the opposite assumption, i.e., that face recognition utilizes the same *mechanisms* (albeit not necessarily the same neurons, see below) used to recognize non-face objects, and then see whether these mechanisms are insufficient to account for the experimental data. Apart from its parsimony, this approach has two key advantages: for one, there are computational models of "generic" object recognition in cortex (e.g., Fukushima, 1980; Perrett & Oram, 1993; Riesenhuber & Poggio, 1999b; Wallis & Rolls, 1997), based on a "simple-to-complex" hierarchical organization of visual processing, with succeeding stages being sensitive to image features of increasing complexity, and stimulus-driven learning shaping the selectivities of neurons at different stages of the processing pathway (Riesenhuber & Poggio, 2000), leading to neurons with tuning to complex, real-world objects at the highest stages in the visual processing pathway (Freedman, Riesenhuber, Poggio, & Miller, 2003; Jiang et al., 2007; Logothetis, Pauls, & Poggio, 1995). This class of models has

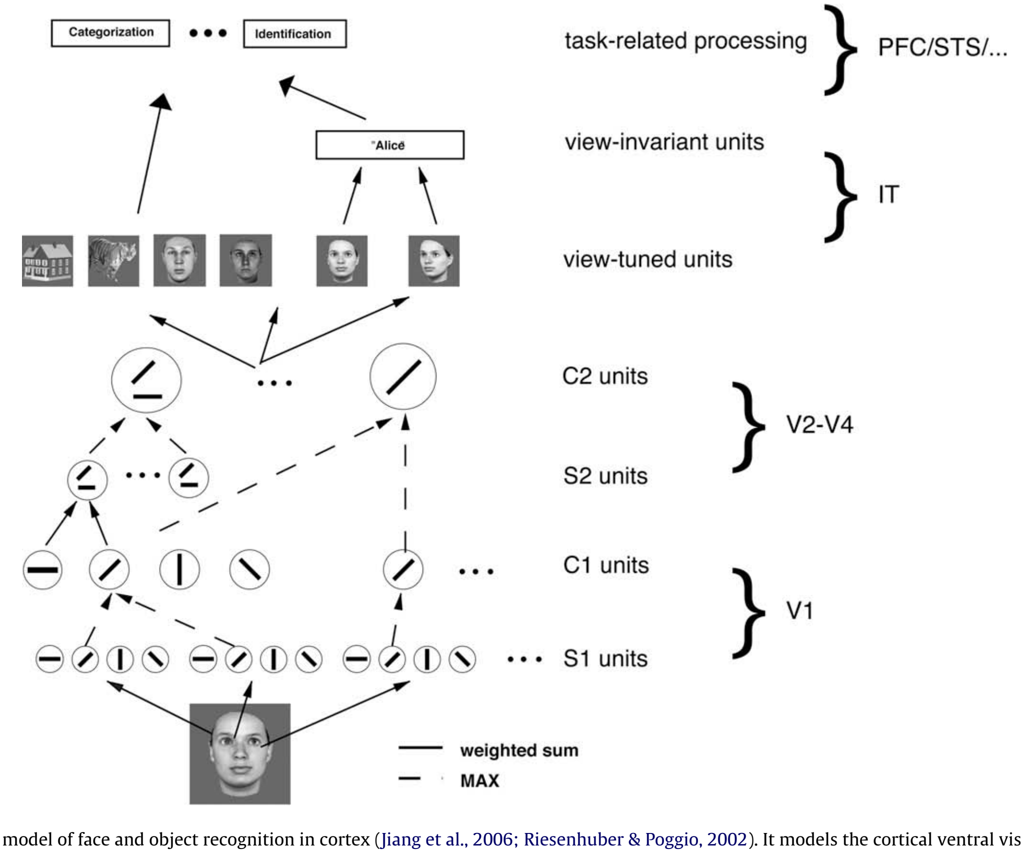

Fig. 1. Scheme of our model of face and object recognition in cortex (Jiang et al., 2006; Riesenhuber & Poggio, 2002). It models the cortical ventral visual stream (Ungerleider & Haxby, 1994) running from primary visual cortex, V1, over extrastriate visual areas V2 and V4 to inferotemporal cortex, IT. Starting from V1 simple cells, neurons along the ventral stream show an increase in receptive field size as well as in the complexity of their preferred stimuli. At the top of the ventral stream, in anterior IT, cells are tuned to complex stimuli such as faces. The bottom part of the model (up to view-tuned units in IT) consists of a view-based module (Riesenhuber and Poggio, 1999b), which is an hierarchical extension of the classical paradigm of building complex cells from simple cells. In the model, C1 cells pool S1 inputs through a MAX-like operation (dashed green lines), where the firing rate of a pooling neuron corresponds to the firing rate of the strongest input, which improves invariance to local changes in position and scale while preserving stimulus selectivity (Riesenhuber and Poggio, 1999b). At the next layer (S2), cells pool the activities of earlier neurons with different tunings, yielding selectivity to more complex patterns as found experimentally, e.g., in V4 (Cadieu et al., 2007). The underlying operation is in this case more "traditional": a weighted sum followed by a sigmoidal (or Gaussian) transformation (solid lines). These two operations work together to progressively increase feature complexity and position (and scale) tolerance of units along the hierarchy, in agreement with physiological data. The output of the view-based module is represented by view-tuned model units (VTUs) that exhibit tight tuning to rotation in depth but are tolerant to scaling and translation of their preferred object view, with tuning properties quantitatively similar to those found in IT (Logothetis et al., 1995; Riesenhuber and Poggio, 1999b). These units can then provide input to task-specific circuits located in higher areas, e.g., prefrontal cortex (PFC).