Vision Research 126 (2016) 9–18

Contents lists available at ScienceDirect

Vision Research

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

Putting low-level vision into global context: Why vision cannot be reduced to basic circuits

Michael H. Herzog [a,*], Evelina Thunell [a], Haluk Ögmen [b]

[a] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[b] Department of Electrical and Computer Engineering, Center for Neuro-Engineering and Cognitive Science, University of Houston, TX, USA

ARTICLE INFO

Article history:
Received 14 February 2015
Received in revised form 28 July 2015
Accepted 18 September 2015
Available online 17 October 2015

Keywords:
Crowding
Non-retinotopic processing
Grouping
Pooling models

ABSTRACT

To cope with the complexity of vision, most models in neuroscience and computer vision are of hierarchical and feedforward nature. Low-level vision, such as edge and motion detection, is explained by basic low-level neural circuits, whose outputs serve as building blocks for more complex circuits computing higher level features such as shape and entire objects. There is an isomorphism between states of the outer world, neural circuits, and perception, inspired by the positivistic philosophy of the mind. Here, we show that although such an approach is conceptually and mathematically appealing, it fails to explain many phenomena including crowding, visual masking, and non-retinotopic processing.

1. Introduction

Understanding a visual scene usually requires taking information across the entire visual field into account. Most models of vision try to break down the complexity of vision by a hierarchical, feedforward filtering approach. Based on the findings by Hubel and Wiesel (1959), neurons in the primary visual cortex (V1) extract lines by combining inputs from LGN neurons as a first step. The outputs of these neurons serve as building blocks for the next step: V1 neurons project to V2 neurons and so on. With each step, receptive field sizes increase and information is *pooled* across larger and larger regions of the visual field, eventually allowing higher cortical neurons to code for high-level features such as shapes and objects (see for example DiCarlo, Zoccolan, & Rust, 2012; Riesenhuber & Poggio, 1999; Thorpe, Fize, & Marlot, 1996). In a cartoon version, a "square neuron" receives input from neurons sensitive to the vertical and horizontal lines that make up the square.

At each stage, neurons are thought to process stimuli in a highly stereotyped fashion. For example, the response of a V1 neuron to a vertical line is independent of whether the line is presented alone or as part of a square because, to reduce complexity and keep processing simple, there are neither lateral nor top-down interactions (except for very local ones as shown in Fig. 1). Therefore, the neuron is "blind" to the horizontal lines of the square and to the overall shape. Low-level determines high-level analysis, but not the other way around. This highly stereotypical, hierarchical and feedforward processing is mathematically appealing, breaking down a seemingly impossible problem into treatable subproblems. Accordingly, research of the last 50 years has focused on understanding the detailed characteristics of low-level visual processing. The subjective aspects of perception, such as perceptual grouping and Gestalt, are assumed to emerge naturally in the subsequent stages, combining the outputs of low-level processing.

Within this framework, there is an important implicit assumption: There is an isomorphism between external world states, their neural presentations, and the corresponding percepts. For example in surround suppression, visibility of a grating patch increases monotonically with the patch size up to a certain point, beyond which visibility *decreases*. How can this non-monotonic dependency of performance on patch size be explained? Fig. 1 shows three tentative circuits. The circuits differ mainly in at which stage inhibition comes into play. The circuits have in common that there is a neuron, whose output determines perception. The neurometric function matches the psychometric function, and because of this isomorphism, subjective terms can be eliminated. In this sense, vision research is well in the tradition of philosophical eliminativism (see for example Churchland, 1981).

Similar circuits are omnipresent in all fields of low-level vision, for example in crowding. In crowding, performance on a target strongly deteriorates when adding elements next to the target (Fig. 2). Crowding has been explained by pooling models, where

* Corresponding author.
E-mail address: michael.herzog@epfl.ch (M.H. Herzog).

http://dx.doi.org/10.1016/j.visres.2015.09.009

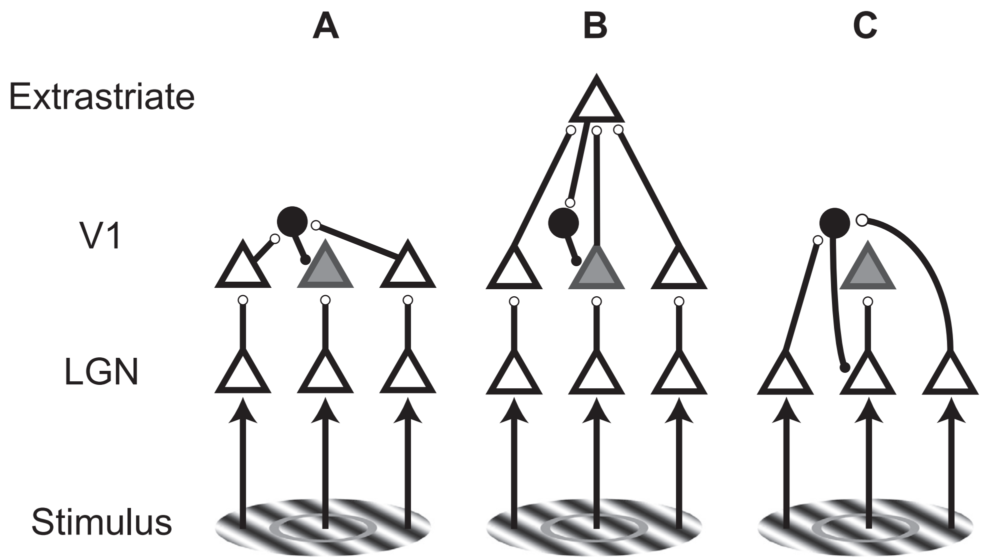

Fig. 1. Three possible mechanisms for surround suppression (Smith, 2006). An output neuron (gray triangle) is activated by the central part of the grating, which falls in the classical receptive field of the neuron. The neuron is inhibited by the surround. The white triangles represent excitatory neurons and the black disks represent inhibitory neurons. (A) Lateral connection model. V1 neurons with receptive fields centered on the surround suppress the central neuron via lateral inhibition. (B) An extrastriate neuron receives input from many V1 neurons, and then inhibits the central neuron via an inhibitory interneuron. (C) Surround suppression is generated within LGN. Similar circuits have been proposed for all types of low-level visual processing. Importantly, changes in the firing of an output neuron are isomorphic to changes in perception. Reproduced from Smith (2006).

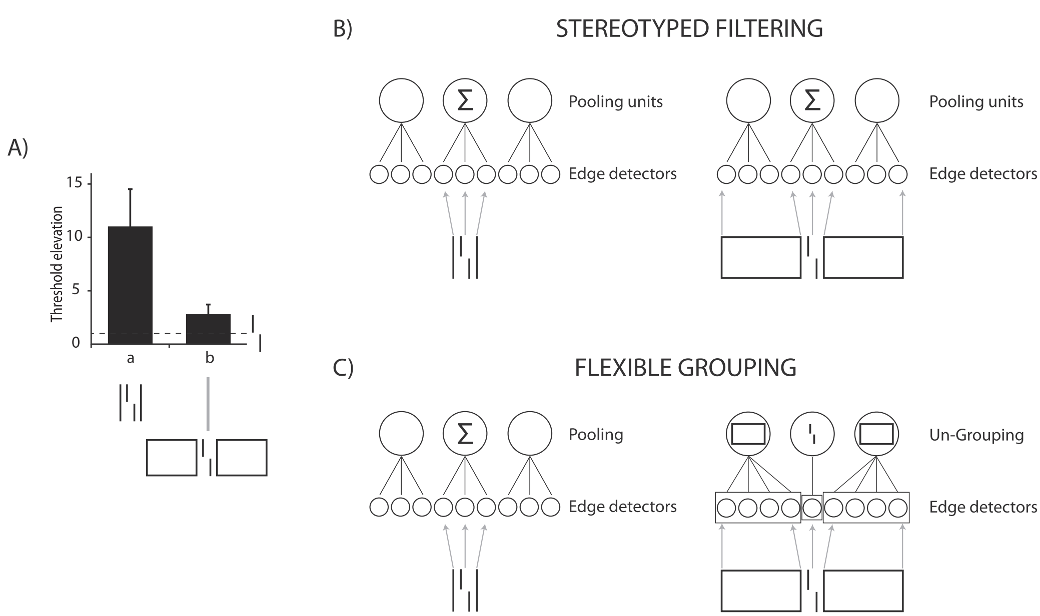

Fig. 2. (A) Observers indicated whether a vernier was offset to the left or to the right (the vernier is shown next to the dashed line). We determined the offset size for which 75% correct responses were obtained (threshold). Results for the flanked conditions are plotted in terms of threshold elevation compared to this unflanked vernier condition (dashed line), i.e., thresholds divided by the threshold of the unflanked condition. A threshold "elevation" of 1.0 indicates no crowding; values larger than 1.0 indicate crowding. Vernier offset discrimination deteriorates when a single line is added on each side (a). When the vernier is flanked by rectangles, performance improves even though the flankers of (a) are part of the rectangles. Adapted from Sayim et al. (2010) and Manassi et al. (2012). (B) In classic models of object recognition, a visual stimulus is analyzed by sets of filter banks. In the first step of filtering, lines and edges are extracted. In the subsequent step, neurons coding for nearby elements combine the outputs of the lower level neurons. Usually, pooling is thought to occur only for neurons coding for similar features, for example vertical but not horizontal lines, in agreement with most psychophysical results (e.g., Kooi et al., 1994). For this reason, the configuration with the vernier plus flanking lines gives a very similar output of filtering as the vernier plus boxes configuration. However, such pooling cannot explain the large effects of configuration shown in A. (C) We propose that our results can only be explained by a flexible grouping stage that determines which elements interfere with each other (by whatever mechanism).

signals of neurons with smaller receptive fields are integrated by neurons with larger receptive fields, causing target irrelevant information from the flankers to be averaged with the target signal (Fig. 2; Pelli, Palomares, & Majaj, 2004; Wilkinson, Wilson, & Ellemberg, 1997). These are examples of classic hierarchical and feedforward models, with pooling as the mechanism for combining information. In this sense, crowding is a reflection of the unavoidable limitations of the visual system. The brain simply cannot do better because it needs to pool information to allow for object recognition. As mentioned, similar circuits have been proposed in all fields of low-level vision with the rationale that once all basic circuits have been understood in great detail and the outputs of the circuits are combined properly, vision will be understood.