Cognition 170 (2018) 228–244

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Attention to distinguishing features in object recognition: An interactive-iterative framework

Orit Baruch*, Ruth Kimchi, Morris Goldsmith

University of Haifa, Israel

ARTICLE INFO

Keywords:
Object recognition
Object categorization
Visual attention
Distinguishing features
Interactive
Iterative
Top-down processes
Bottom-up processes

ABSTRACT

This article advances a framework that casts object recognition as a process of discrimination between alternative object identities, in which top-down and bottom-up processes interact—iteratively when necessary—with attention to distinguishing features playing a critical role. In two experiments, observers discriminated between different types of artificial fish. In parallel, a secondary, variable-SOA visual-probe detection task was used to examine the dynamics of visual attention. In Experiment 1, the fish varied in three distinguishing features: one indicating the general category (saltwater, freshwater), and one of the two other features indicating the specific type of fish within each category. As predicted, in the course of recognizing each fish, attention was allocated iteratively to the distinguishing features in an optimal manner: first to the general category feature, and then, based on its value, to the second feature that identified the specific fish. In Experiment 2, two types of fish could be discriminated on the basis of either of two distinguishing features, one more visually discriminable than the other. On some of the trials, one of the two alternative distinguishing features was occluded. As predicted, in the course of recognizing each fish, attention was directed initially to the more discriminable distinguishing feature, but when this feature was occluded, it was then redirected to the less discriminable feature. The implications of these findings, and the interactive-iterative framework they support, are discussed with regard to several fundamental issues having a long history in the literatures on object recognition, object categorization, and visual perception in general.

1. Introduction

Object recognition is of fundamental importance for the perception of and interaction with our environment. Despite extensive research, however, there is still no complete and comprehensive theory that can explain how we recognize objects, and some of the most basic characteristics of the recognition process continue to be a subject of debate.

One controversial issue concerns the role of top-down versus bottom-up processing. Despite many other differences, classic theories of object recognition (e.g., Biederman, 1987; Marr & Nishihara, 1978; Poggio & Edelman, 1990; Reisenhuber & Poggio, 1999; Tarr & Bülthoff, 1995; Tarr & Pinker, 1989; Ullman, 1989) are generally united in what might be called the orthodox view—that object recognition is based primarily on a bottom-up analysis of the visual input; recognition is achieved when some temporary representation of the input image matches a stored object representation. The functional architecture of the visual cortex—the increase in the receptive field size and in representational complexity from lower to higher areas in the cortex (Maunsell & Newsome, 1987; Vogels & Orban, 1996)—has been pointed to as consistent with the bottom-up view. Also, findings that the processes involved in object recognition are sometimes remarkably fast, occurring within 100–200 ms of stimulus presentation (e.g., Thorpe, Fize, & Marlot, 1996), have been taken by some researchers as evidence that object recognition can occur largely with feed-forward processing alone (e.g., Wallis & Rolls, 1997; but see Evans & Treisman, 2005).

Some proposals, however, have challenged the orthodox view, emphasizing the need for both bottom-up and top-down processing (e.g., Bar, 2003; Bullier, 2001; Ganis, Schendan, & Kosslyn, 2007; Humphreys, Riddoch, & Price, 1997; Lee, 2002; McClelland & Rumelhart, 1981; Schendan & Maher, 2009; Schendan & Stern, 2008; Ullman, 1995). For example, Bar (2003), Bar et al., 2006), inspired by Ullman's (1995) model, proposed that partially processed visual data based on low spatial frequencies of the input is transmitted from the initial areas of the visual stream directly to the orbito-frontal cortex. This low-spatial-frequency representation invokes initial hypotheses regarding the identity of the input, which subsequently facilitate the identification process by constraining the number of possibilities that have to be inspected (see also

* Corresponding author.
E-mail address: oritb@research.haifa.ac.il (O. Baruch).

http://dx.doi.org/10.1016/j.cognition.2017.10.007
Received 15 September 2016; Received in revised form 3 October 2017; Accepted 6 October 2017

Peyrin et al., 2010). Similarly, Bullier (2001) proposed a model of visual processing by which initially, information from the visual stimulus is transferred rapidly via magnocellular, dorsal pathways. Results from this first-pass computation are then sent back by feedback connections and used to guide further processing of parvocellular and koniocellular information in the inferotemporal cortex. The existence of massive projections from higher to lower areas of the visual pathways (e.g., Bullier, 2001; Lamme & Roelfsema, 2000) suggests that the involvement of top-down processing in object recognition is physiologically viable. Top-down influences on object recognition are also implicated in behavioral studies. For example, advance information about the target in RSVP experiments improves target detection (Intraub, 1981), priming by category names substantially improves object identification (Reinitz, Wright, & Loftus, 1989), and objects are recognized better in expected than in unexpected contexts (e.g., Bar & Ullman, 1996; Biederman, 1972, 1981).

Several models propose that top-down and bottom-up information might be integrated via an iterative error-minimization mechanism, where top-down predictions are matched to processed bottom-up information in recursive, interacting loops of activity (Friston, 2005; Hinton, Dayan, Frey, & Neal, 1995; Kveraga, Ghuman, & Bar, 2007; Mumford, 1992; Ullman, 1995).

2. Role of attention in object recognition

Partly related to the preceding issue is an ongoing controversy regarding the role of attention in object recognition. Some researchers have provided evidence suggesting that object recognition can be carried out in the near absence of attention (e.g., Li, VanRullen, Koch, & Perona, 2002; Luck, Vogel, & Shapiro, 1996). Other researchers, however, hold that attention plays a central role (e.g., Ganis & Kosslyn, 2007; Hochstein & Ahissar, 2002; Treisman & Gelade, 1980). Most notably, the highly influential Feature Integration Theory (Treisman & Gelade, 1980) holds that attention is crucial for the perception of an integrated object, as it operates to bind featural information represented in independent feature maps. In contrast, the more recent Reverse Hierarchy Theory (Hochstein & Ahissar, 2002) holds that whereas the initial perception of coherent conjoined objects can be achieved “at a glance” under spread attention, based on feedforward processing alone, top-down focused attention must subsequently be invoked to consciously identify specific details such as orientation, color, and precise location.

An additional role for attention in object recognition has emerged from the view of visual perception as a process of hypothesis testing (Gregory, 1966; von Helmholtz, 1867), by which attention is directed to diagnostic feature information that is used to decide between alternative hypotheses regarding object identity (e.g., Baruch, Kimchi, & Goldsmith, 2014; Ganis & Kosslyn, 2007; Ganis et al., 2007). This view, advanced in the present research, is outlined in the following section.

3. Interactive-Iterative attentional framework for object recognition

The present work was guided by a framework that views object recognition as a process of discrimination between probable alternatives—a process in which bottom-up and top-down processes interact, iteratively when necessary, with attention playing a crucial role in this interaction. We outline here the set of principles that comprises this framework (a more concrete schematic depiction appears as Fig. 13 in General Discussion)—essentially, a synthesis of ideas that have been proposed previously, from which specific predictions can be derived and empirically examined.

3.1. Object recognition begins with expectations based on past experience and present context

Object recognition undoubtedly requires an analysis of visual data. Yet, contrary to the conventional view, we suggest that the recognition process actually begins at the top. Everyday situations generally evoke expectations about probable objects, based on world knowledge, context, and goals (e.g., Bar, 2004; Biederman, 1972; Norman & Bobrow, 1976; Palmer, 1975). Even in the laboratory, expectations are evoked by the experimental task. Pure data-driven recognition—where an object could be anything—are presumably quite rare, and can be seen as a special case in which the probable alternatives are all objects known to the observer. A similar view of perception has recently been revived in several models using Bayesian inference, in which top-down priors help to disambiguate noisy bottom-up sensory input signals (e.g., Epshtein, Lifshitz, & Ullman, 2008; Friston & Kiebel, 2009).

3.2. The initial visual input is inherently limited

The initial information extracted from the visual scene in a data-driven (bottom-up) manner is inherently partial. In natural scenes, portions of objects—those on the side away from the viewer—are hidden from view and surfaces may undergo occlusion; sometimes the viewing conditions are poor, and at other times the relevant diagnostic information is subtle and cannot be acquired at a glance. Moreover, even under optimal viewing conditions, the initial information may be partial (e.g., coarse information carried by low spatial frequencies; Bar, 2003; Fabre-Thorpe, 2011; Hughes, Nozawa, & Kitterle, 1996). Although, depending on context, the initial partial information may sometimes suffice for recognition, in many cases object recognition will require additional processing.

3.3. Perceptual hypotheses guide the allocation of attention to distinguishing features

It was suggested long ago (Gregory, 1966; von Helmholtz, 1867) that perception is essentially a hypothesis-assessment process. Building on this idea, and in line with more recent ideas concerning the “predictive brain” (e.g., Bubic, von Cramon, & Schubotz, 2010; Enns & Llreas, 2008), we assume that the observer’s expectations—whether formed prior to or in interaction with the visual input—evoke a set of alternative hypotheses regarding the possible identity of the observed object. These hypotheses are expressed as the activation of internal representations of candidate objects, giving special weight to diagnostic features[1] (e.g., Gillebert, Op de Beeck, Panis, & Wagemans, 2009; Schyns & Rodet, 1997; Sigala & Logothetis, 2002; Wagar & Dixon, 2005) that discriminate between competing hypotheses. Attention is then directed to these distinguishing features in order to facilitate the extraction of the relevant information (see also Ganis & Kosslyn, 2007; Kosslyn, 1994).

The specific claim that attention is directed to distinguishing features in object recognition has been empirically addressed in relatively few studies, most of which used eye tracking as an indirect measure of spatial attention. For example, Rehder and Hoffman (2005a, 2005b; see also Blair, Watson, Walshe, and Maj, 2009) found that during visual object category learning, diagnostic features were fixated more often than non-diagnostic features, and that the proportion of correct responses correlated with the time diagnostic features were fixated. Using a more direct measure of spatial attention in the context of word and

[1] Note that the notion of features (and hence, distinguishing features) as conceived here is very broad, and refers to any aspect of an object that can serve to discriminate between the set of probable alternatives. Such aspects may include, for example, structural or configural features (e.g., *geons*; Biederman, 1987), surface features (e.g., color or texture), global features (e.g. global shape: elongated vs. round), or localized features and parts (e.g., the shape or color of a beak).