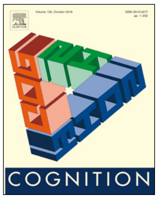
Cognition 180 (2018) 158–164

Contents lists available at ScienceDirect

Cognition

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



Original Articles

Context mitigates crowding: Peripheral object recognition in real-world images

Maarten W.A. Wijntjes[a,*], Ruth Rosenholtz[b]

[a] Perceptual Intelligence Lab, Industrial Design Engineering, Delft University of Technology, The Netherlands
[b] Dept. of Brain & Cognitive Sciences, CSAIL, MIT, United States

ARTICLE INFO

Keywords:
Object recognition
Crowding
Context
Integration
Shrink-wrap

ABSTRACT

Object recognition is often conceived of as proceeding by segmenting an object from its surround, then integrating its features. In turn, peripheral vision's sensitivity to clutter, known as visual crowding, has been framed as due to a failure to restrict that integration to features belonging to the object. We hand-segment objects from their background, and find that rather than helping peripheral recognition, this impairs it when compared to viewing the object in its real-world context. Context is in fact so important that it alone (no visible target object) is just as informative, in our experiments, as seeing the object alone. Finally, we find no advantage to separately viewing the context and segmented object. These results, taken together, suggest that we should not think of recognition as ideally operating on pre-segmented objects, nor of crowding as the failure to do so.

1. Introduction

How do we recognize an object? Traditional theories of visual perception suggest that the visual system must segment the object from the background, and piece together or "integrate" features of increasing size and complexity in order to recognize the object. There exist a number of explicit examples of this theory (e.g. Biederman, 1987; Kosslyn, 1987; Marr, 1982; Neisser, 1967; Palmer & Rock, 1994a, 1994b). The idea is also implicit in a number of theories. Selfridge (1959), for instance, describes matching an object in memory to the "observed object" without regard for how the latter might be distinguished from the background or surrounding clutter.

It seems at first glance almost a logical necessity that object recognition ignores spurious features outside the object. If this view is correct, then a fundamental issue consists of how to integrate the parts that belong to the object and ignore the parts that do not. Some researchers have suggested that this is a role for attention: that attention "selects" the target, in essence "shrink-wrapping" it so that the visual system can respond to its features and not those of surrounding image regions (Moran & Desimone, 1985).

In the fovea, object recognition is relatively robust and effortless. However, the visual system has trouble recognizing objects in the peripheral visual field in the presence of nearby flanking stimuli, a phenomenon known as crowding (Whitney & Levi, 2011; Levi, 2008; Pelli & Tillman, 2008). Crowding is characterized by a critical distance within which clutter greatly disrupts recognition of the target object (Bouma, 1970). Across a range of stimuli, the critical distance equals approximately half the eccentricity, i.e. the distance between the target and the point of fixation (Pelli & Tillman, 2008). Crowding has been attributed to a failure of object recognition mechanisms to limit integration of features to the object of interest, known as "excessive integration": (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli, Palomares, & Majaj, 2004; Pelli & Tillman, 2008; Chakravarthi & Cavanagh, 2009; Bernard & Chung, 2011). Some researchers have further suggested that the excessive integration might be due to limited attentional resolution (He, Cavanagh, & Intriligator, 1996; Intriligator & Cavanagh, 2001; Yeshurun & Rashal, 2010; and related to the more general notions of competition in Desimone & Duncan, 1995), such that the peripheral visual system cannot "select" only the object of interest for further processing.

These theories, both of normal object recognition mechanisms isolating the target object, and of crowding as a failure to do so, presume that ideally the visual system should shrink-wrap the target, integrating features over only its area. However, in everyday life, objects tend to appear in certain environments and not others. These regularities mean that context, i.e. the surrounding scene, provides cues for object recognition. Oliva and Torralba (2007) eloquently demonstrated this theoretical point by collecting a large number of images of a given type of object, centering them on that object, and averaging them. If context were uninformative, the result would be a uniform gray field

[*] Corresponding author at: Industrial Design Engineering, Delft University of Technology, Landbergstraat 15, 2628CE Delft, The Netherlands.
E-mail address: m.w.a.wijntjes@tudelft.nl (M.W.A. Wijntjes).

https://doi.org/10.1016/j.cognition.2018.06.015
Received 11 July 2017; Received in revised form 19 June 2018; Accepted 20 June 2018

everywhere except at the location of the object. Instead, the average images show considerable structure: keyboards tend to appear below computer monitors and on top of desks; faces tend to appear above a body and near the horizon; a fire hydrant sits on the ground plane; and boats lie in the water near other boats (http://people.csail.mit.edu/torralba/gallery). Nor does context only inform perception at the level of object recognition. The same image regularities that lead to Gestalt grouping mean that neighboring image regions are often informative as to the features of a given region. A particular edge segment, for instance, tends to co-occur with neighboring edges of certain locations and orientations, and not with others (Geisler & Perry, 2009).

The visual system clearly can make use of contextual information. A letter is better recognized within a meaningful word than in isolation (Reicher, 1969; Wheeler, 1970). When a gray mask hides an object, observers can correctly guess that object's category on their first try more than 60% of the time (Greene, Oliva, Wolfe, & Torralba, 2010). The Fusiform Face Area shows as much fMRI activation to a face implied by contextual cues (a body) as it does to a face alone (Cox, Meyers, & Sinha, 2004), although others have argued that this may be an artifact of low-resolution fMRI scanning (Schwarzlose, Baker & Kanwisher, 2015).

Given the potential importance of contextual information, does crowding point to a puzzling failure of peripheral vision to shrink-wrap the target? Or is such shrink-wrapping not ideal in real-world vision? We ask observers to recognize objects in real images, with naturally occurring correlations between the object and other scene elements, and natural amounts of nearby clutter. By varying the window through which observers view the peripheral object, we examine the relative importance of shrink-wrapping and integrating contextual information.

2. Experiment 1

We asked observers to identify peripheral objects, and varied the size of the surrounding aperture. The smallest aperture just fit the target; the largest aperture was five times the object size (Fig. 1A).

Fig. 1B shows several possible outcomes. Typical crowding experiments utilize arrays of items against a blank background, such as a triplet of letters. By design, the letters flanking the target are completely uninformative as to the identity of the target. In such experiments, performance typically drops as the flankers move to within the critical distance of the target. Based upon typical crowding experiments, what results might we expect when we vary the aperture size? Small aperture sizes are similar to wide target-flanker spacing, in the sense that no flankers appear within critical distance of the target. On the other hand, for larger aperture sizes, clutter lies within the critical distance of the target. Similarly, in traditional crowding experiments clutter lies within this window when target-flanker spacing is less than the critical distance. If our results were like classic crowding, we would expect performance to drop as the aperture size increased beyond the size of the object, asymptoting as it reached Bouma's critical distance (blue curve). On the other hand, to the extent that the visual system makes use of informative context, we would expect larger apertures to facilitate recognition, at least partially mitigating negative effects of crowding. Performance might follow a dipper function (yellow), in which for small apertures crowding dominates, but at larger aperture sizes contextual facilitation takes over. Crowding and contextual effects might balance, at least for small apertures (green). Or contextual information might more than compensate for detrimental effects of clutter (red). Of course, what happens in practice will depend heavily on the difficulty of the object recognition task (e.g. basic level categorization vs. subordinate level), and the degree of correlation between object identity and context in a particular dataset. Our goal here is to see what happens if we pick a natural image dataset, and a collection of common objects (i.e. no cherry-picking of either objects or their context), and ask for a straightforward and natural basic-level categorization.

2.1. Methods

2.1.1. Participants

Five observers participated, all male students (mean age 21). All had normal or corrected-to-normal vision and were native Dutch speakers. This number of observers was chosen based on power calculations as follows: As each observer views a given object at only one of five window sizes, we combine across observers to compare performance for different apertures. Five observers gives us 656 trials per condition. For a binomial distribution, the estimated confidence interval (CI) is largest at a probability of 0.5. For n = 656 at p = 0.5, the estimated CI is ± 0.04, which we deemed sufficiently precise to reveal important differences between the conditions (note that for a typical crowding experiment performance varies from near chance for the smallest

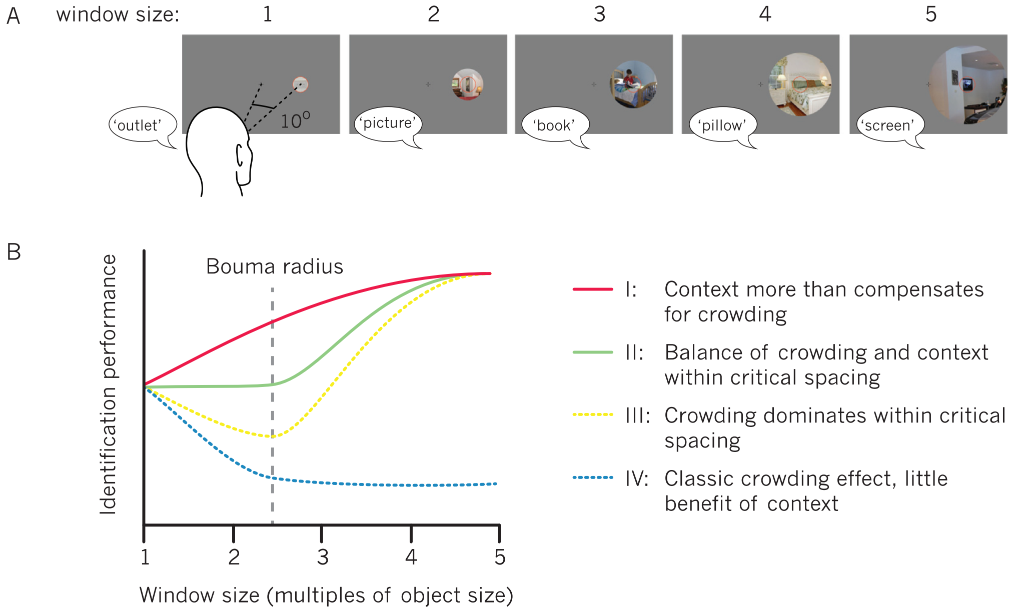

A
window size:
1
2
3
4
5
'outlet'
10°
'picture'
'book'
'pillow'
'screen'
B
Bouma radius
Identification performance
1
2
3
4
5
Window size (multiples of object size)
I: Context more than compensates for crowding
II: Balance of crowding and context within critical spacing
III: Crowding dominates within critical spacing
IV: Classic crowding effect, little benefit of context

Fig. 1. Experiment 1, methodology and predictions. (A) Each target appears at 10° eccentricity, within 5 possible aperture sizes. (B) The effects of informative context and visual crowding work in opposition. A number of outcomes are possible, depending upon the relative strength of these two factors (see text).