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
Available online 25 July 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

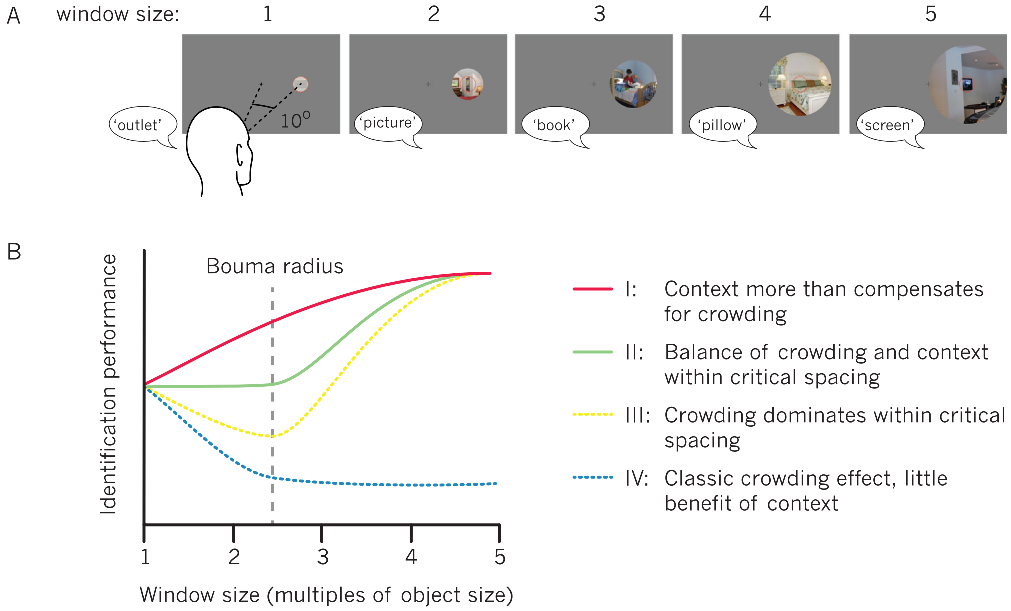
A
window size:
1
2
3
4
5
'outlet'
10°
'picture'
'book'
'pillow'
'screen'
B
Bouma radius
Identification performance
1
2
3
4
5
Window size (multiples of object size)
I: Context more than compensates for crowding
II: Balance of crowding and context within critical spacing
III: Crowding dominates within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

target-flanker spacing to 80% correct or more for an unflanked target).

2.1.2. Stimuli

Stimuli derived from the SUN2012 database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). It contains a diverse collection of object and scene categories, with a large number of images per category. The database also includes hand-labeled masks for individual objects, gathered using the LabelMe toolbox (Russell, Torralba, Murphy, & Freeman, 2008).

For each of the 100 most common objects in the SUN2012 database – excluding background elements such as wall, sky, ceiling, etc. – we randomly selected image-object pairs to satisfy the following constraints: (1) the object must have sufficient resolution, subtending at least 100 pixels in width or height; (2) all five apertures must fit completely within the image; (3) we eliminated any occluded objects. The final selection contained a total of 656 stimuli, consisting of 85 unique object labels. The number of stimuli within the same category ranged from 1 to 24, with an average of 7.7 stimuli per object category. Examples are shown in 1B.

We scaled each image so that the target subtended 4° in its largest dimension, and centered the image such that target objects appeared at 10° eccentricity. The aperture surrounding the target varied in size from 4° to 20° diameter, in steps of 4°. The largest aperture extended from the central fixation point to the edge of the screen (see Figure A, far right). At 10° eccentricity, crowding typically occurs when flankers lie within a critical spacing of approximately 5°, i.e. within a 10° diameter aperture (Bouma, 1970). The classical critical spacing, then, lies midway between the 2nd and 3rd-smallest aperture sizes.

2.1.3. Procedure

Each observer saw all 656 objects in one of five possible aperture sizes. The order of objects and aperture sizes were random and the aperture sizes were balanced across observers. Thus, each object was presented within five different aperture sizes to five different observers, and each data point in Fig. 2A consists of results from 656 trials.

For each trial, the observer fixated an isolated central cross, and then pressed the space bar to start the presentation. The observer was seated on one (long) side of a table looking at the presentation screen, while the experimenter (the first author, in experiments 1 and 2) was seated at the other side, looking at a flanking screen facing the opposite direction. A webcam under the observers' screen streamed a close-up of the observers' eyes to the experimenters' screen. During a trial, the experimenter watched this stream closely until an answer was given, which he logged on a spreadsheet in his screen. The observer was also instructed to report fixation violations. A violation reported by either the observer or experimenter result in the trial being discarded; in practice nearly all reported violations were noticed by the experimenter (i.e. it rarely happened that the observers reported a violation while the experimenter did not), while observers sometimes did not notice their own deviations, as one would expect. Variability among observers was rather large (4, 8, 35, 63, 16 fixation violations were counted for the five observers, respectively) amounting to an average of 3.8% of the trials. In the appendix an overview of all fixation violations per conditions for all experiments is presented. While the method we used to detect fixation violations is not as robust as one utilizing an eye tracker, the violation data in the appendix do not suggest that undetected violations drive any of the reported results (under the assumption that undetected violations for a given condition are proportional to detected violations).

In typical crowding experiments one can simply instruct the observer to identify the middle object in an array. However, for real scenes it can be difficult to determine which object lies in the center of a sizeable aperture, even when centrally viewed. We used a red circle, the size of the smallest aperture, as a precue, to clarify the task. The scene faded in (approx. 600 ms), to minimize spontaneous saccades due to sudden peripheral change. Presentation time was not limited but observers were encouraged to respond promptly. The observer responded by pressing a space bar to terminate the presentation, and then verbally indicated the basic-level object category. Answers were made in Dutch by native Dutch speakers, and logged by the (Dutch-speaking) experimenter.

2.1.4. Data analysis

We viewed the original photo while evaluating the correctness of each response, in order to be robust to labeling errors or ambiguities in the SUN2012 database. This led to a fairly liberal criterion for correct identification – rather than requiring a match to the SUN2012 label, we accepted observer labels that appeared to match the target object. Although we instructed observers to give a basic-level categorization (e.g. car), they sometimes responded with a subordinate level (e.g. convertible). We scored such responses as correct. Superordinate levels (e.g. vehicle) counted as incorrect. We computed average accuracy across observers, for each aperture size.

2.2. Results

To assess which of the qualitative models shown in Fig. 1 would apply, we tested for significant differences between subsequent window sizes. Because the accuracy data is binomially distributed, we used two-sample proportion tests converting the differences to z-scores. We used the binomial approximation to quantify 95% confidence intervals.

Accuracy monotonically increased from 45.2% (95% CI: 41.2–49.0) for the smallest aperture to 70.1% (CI: 66.5–73.6) for the largest aperture (Fig. 2A). Performance appears to saturate at around the third

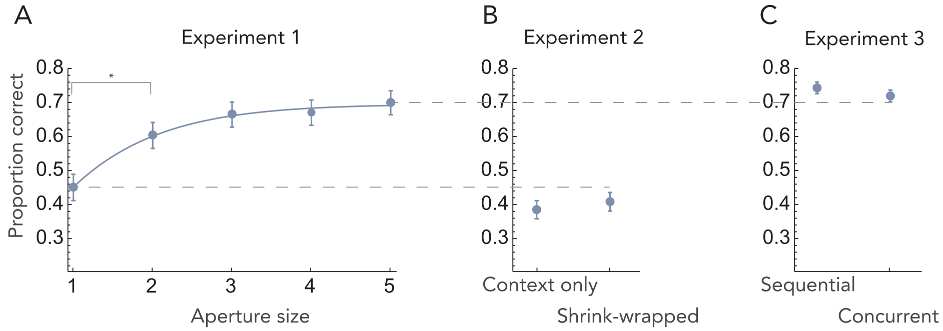

Fig. 2. Peripheral recognition accuracy in all three experiments. (A) Accuracy as a function of aperture size. Shown with best-fit logistic function. Error bars indicate 95% confidence intervals on the data points. The plot clearly shows that accuracy rises rapidly between the first and second aperture size and asymptotes around the third aperture size. (B) Context provides nearly as much information as the object alone. Performance for the shrink-wrapped target is similar to that for the smallest window size in (A), suggesting the poor performance with a small aperture was not due to poor shrink-wrapping. (C) Performance is similar when context and object are presented together as when the context is presented first, followed by the shrink-wrapped target. The stimulus presentation of the Concurrent condition is similar (but different observers) to the largest apertures size in (A), which is reflected by the similar accuracies.

aperture size, i.e. just beyond Bouma's critical distance. Planned comparisons for each of the four consecutive aperture pairs revealed that aperture size 2 was significantly better than size 1 (proportion test, $z = 5.48$, $p < 0.05$), no other consecutive aperture size pair showed significant differences ($p > 0.05$).

We also fit a logistic function $f(x)=(1-c)/(1+e^{-(x-a)/b})$ to the data. Here, x denotes the aperture size, $(1-c)$ denotes the asymptotic behavior for increasing x, a specifies the horizontal translation of the whole function, and b partly (together with c) determines the slope. We found that the best fit parameter of c to be 0.304, (95% CI: 0.245–0.330), for an asymptotic recognition rate of approximately 70% correct. Given that c is less than 1, the sign of b distinguishes between a monotonically increasing or decreasing function. The best fit parameter for b was 0.809 (95% CI: 0.520–1.595), indicating that performance significantly increases rather than decreases with increased aperture size.

Except for model I, the models shown in Fig. 1B all yield equal or worse performance as aperture sizes increase from the size of the target up to Bouma's radius. The data shows instead an increase in performance within this regime, as confirmed statistically both by the planned comparisons and the best fit parameters of the logistic function. Including the background in object recognition leads to better rather than worse performance. Fixation violations were relatively constant over the various aperture sizes (23, 27, 34, 22, 20), indicating that the pattern of better performance for larger apertures is unlikely to have arisen from observers being more likely to violate fixation for those conditions.

2.3. Discussion

Context clearly provides information for peripheral object recognition in real scenes. More context, i.e. a larger aperture around the target, is better. The benefit of context outweighs any degradation due to not shrink-wrapping the target.

If crowding is a failure to select a target from its surround, then helping select it should improve identification. Instead, the opposite occurs; smaller apertures yield poorer performance.

3. Experiment 2

These results naturally raise the question of how well observers would identify the object if shown only the context and not the object itself. In addition, the smallest aperture may still contain some clutter; would performance improve if we more carefully shrink-wrapped the object to its silhouette?

3.1. Methods

Four observers (two males) participated in the second experiment. One male was 58, the other three observers were students with a mean age of 21. All had normal or corrected-to-normal vision and all were native Dutch speakers.

We used the same set of 656 images from SUN2012 as in the first experiment. The experimental procedure was also similar, except that it included only two apertures: 'context-only' and 'shrink-wrapped' (see Fig. 3). In the Context-only condition, the largest aperture size was used, with the circular target area blanked out with the same mid-gray as the background. Trials were blocked by aperture type, with block order balanced across participants. Each observer saw a given scene only once, in only one of the two presentation conditions. Therefore, 1312 trials underlie analysis of each of the two conditions. Fixation violations again varied substantially between observers (44, 0, 68, 28 times, respectively), amounting to discarding 5.3% of the trials. For all observers, the violations were higher or similar in the Shrink-wrapped condition.

3.2. Results

In both the Context-only and Shrink-wrapped conditions, observers correctly identified about 40% of the stimuli (Fig. 2B). Chance is at most 1.2% correct (there were 85 object types present in the experimental set, although observers were naïve as to this number). The difference between the two conditions was small and not significant ($z(x) = 1.2027$, $p = 0.23$, Context-only: 38.6% (CI: 35.9–41.2), Shrink-wrapped: 40.9% (CI: 38.2–43.6)).

We also conducted proportion tests to compare the current results with Experiment 1. Performance with the largest aperture size was significantly better than with context alone ($z = 12.9697$, $p < 0.05$). Though context alone proves quite valuable in identifying the (non-visible) object, recognition is better if the object is visible. The difference between the smallest aperture size condition and the Shrink-wrapped condition was not significant ($z = 1.75969$, $p = 0.078$); careful shrink-wrapping does not improve performance with respect to the smallest aperture size from Experiment 1.

3.3. Discussion

With context alone, observers attain 40% correct performance identifying (invisible) peripheral objects. Context alone is as good as viewing only the target. These results would surely vary with the particular choice of images, targets, and task (e.g. basic-level vs. subordinate-level categorization). Nonetheless, with a plausible sampling of real image-object pairs, context is clearly highly important for basic-level categorization of objects in the periphery. Exactly what drives this perhaps surprisingly good performance at identifying unseen objects? In some images, nearby objects may actually have the same identity as the target; for example, cars tend to appear on the street near other cars. Scene category clearly affects the likelihood that a given target appears. Pairs of objects may also tend to co-occur, independent of scene category (Draschkow & Võ, 2017); a spoon may lie near a teacup regardless

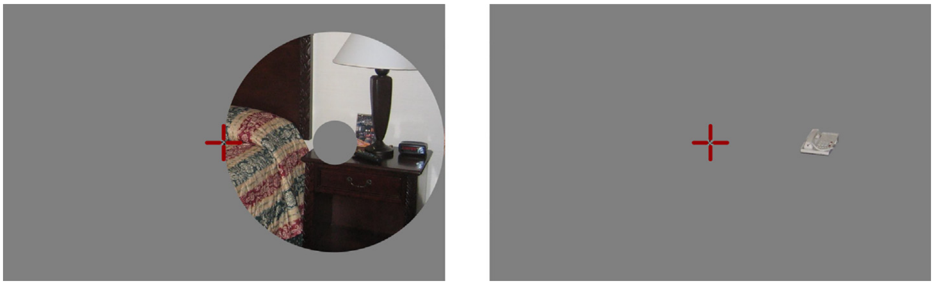

Fig. 3. Screenshots of the Context-only and Shrink-wrapped conditions in Experiment 2. As in Experiment 1, observers had to fixate the central fixation cross (shown in the figure enlarged by a factor of 3 and red for visibility). In one block only the context was presented by showing the largest aperture size with the target covered by a gray disk. In another block a shrink-wrapped version of the object was shown, using the polygonal masks given in the SUN2012 database. The stimulus pair shown here is complementary: the object and surround originate from the same image. For a given original image, each observer saw either the context alone or the object alone, but not both.

of where one takes one's tea. Because our stimulus set was chosen relatively randomly, rather than controlled, it is difficult to make strong inferences about the relative contribution of these factors without more systematic study.

4. Experiment 3

Clearly contextual information is useful for peripheral object recognition. However, perhaps shrink-wrapping remains optimal for the object recognition system, so long as some other mechanism extracts contextual information, such as the gist of the scene. Then a decision process could combine object and contextual information to categorize the object. If so, we would expect to see an advantage to presenting context and object separately, compared to presenting them together. Presenting the isolated object would overcome peripheral vision's failure to shrink-wrap the target, while presenting the context would normalize the available contextual information. The magnitude of the benefit for separate presentation provides a measure of the cost of peripheral vision failing to select the target.

4.1. Methods

We compared presenting context and target one after the other (Sequential condition) to presenting them at the same time and in the same image (Concurrent condition). Eight observers participated. All had normal or corrected-to-normal vision. All were native (American) English speakers.

Stimuli and procedure were similar to Experiments 1 and 2. For both Sequential and Concurrent position we used the largest aperture size. In the Sequential condition, the context faded in over 600 ms, was present for 1000 ms, after which the shrink-wrapped stimulus faded in over 600 ms and remained on until a response (as previous experiments). In the Concurrent condition, observers saw an aperture containing the target object, similar to the largest aperture condition in Experiment 1. An illustration of the procedure is shown in Fig. 4. Each observer saw each of the 656 stimuli in only one condition. The conditions were blocked and counterbalanced: four participants first viewed 328 stimuli in the Sequential condition followed by the remaining 328 stimuli in the Concurrent condition, and other the four participants first viewed the Concurrent condition. A total of 2624 trials underlie analysis of each of the two conditions. Fixation violations (7, 7, 25, 20, 35, 13, 25, 9 times for each of the 8 observers, respectively) amounted to 2.7% of the trials being discarded.

4.2. Results

Performance in the Sequential condition was 74.4% (CI: 72.7–76.1), very similar to the 72.1% (CI: 70.3–73.8) correct performance in the Concurrent condition (Fig. 2C). A two-proportion z-test found no significant difference between the two conditions (z = 1.93, p = 0.054).

4.3. Discussion

Even when one normalizes the amount of context available, performance is still not appreciably better with a shrink-wrapped target. Even if one considers the difference marginally significant, the impact on performance of viewing the target in context – the cost of failing to select the target – is very small. Sequential viewing provided a benefit of only 2.3% correct over performance in the Concurrent condition. If crowding were a critical failure to select the target, then we would expect to see substantially better performance when viewing shrink-wrapped targets in the Sequential condition.

5. General discussion

We found that peripheral object recognition performance monotonically increases as a function of the size of the viewing aperture around the target object. Viewing the context alone, without a visible target, led to as good of performance as viewing the shrink-wrapped target. Finally, there was no great advantage to viewing the context and object separately, compared to viewing the object within its natural context.

It is important to discuss these results within the context of classic crowding experiments. Previous studies using carefully controlled stimuli have provided great insight into peripheral mechanisms. Peripheral vision appears to integrate information over sizeable regions, often leading to significant degradation in performance when those regions contain irrelevant clutter. Crowding is often taken to mean that "objects in the world, unless they are very dissimilar, can be recognized only if they are sufficiently separated" (Pelli, 2008). This would predict that peripheral object recognition in the real world would generally be quite poor. However, Experiment 1 showed that in real images, observers recognize peripheral objects quite well. Furthermore, crowding has been described as a failure to select only the target, so as to integrate only the features of that object for recognition. When we "select" the object for the observer, however, performance is worse. This result differs from typical crowding experiments, in which (effectively) cutting out the target and displaying it apart from the flankers greatly improves performance. Note, however, that our results do not suggest that crowding mechanisms do not operate when viewing real-world scenes; presumably they do (Rosenholtz, 2014). In Experiment 1, peripheral object recognition performance asymptotes at only about 70% correct. Crowding is likely a major factor in this poor performance. Although we did not control for other possible losses in peripheral vision, such as reduced acuity, acuity losses are considerably smaller, and likely of less importance in identifying real-world stimuli

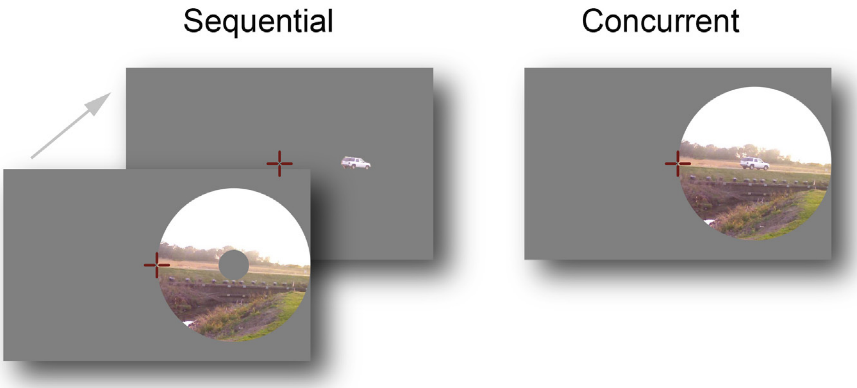

Fig. 4. Screenshots of Experiment 3. In the Sequential condition observers were first presented with the context for 1 s and then shown the shrink-wrapped object. In the Concurrent condition the object and context were shown simultaneously, similar to presentation of the largest aperture size from Experiment 1.

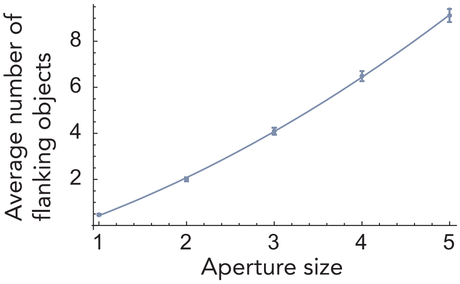

Fig. 5. Number of flanking objects within the aperture, as a function of aperture size. An object was considered present if the centroid of its LabelMe mask lay within the aperture. This definition likely underestimates the degree of crowding of target objects in our stimulus set, since crowding can occur when the target is flanked by parts of objects, or by clutter other than objects. Nevertheless, this plot gives some intuition about the amount of object clutter in our scene stimuli. The classic critical distance of crowding lies between aperture sizes 2 and 3, within which there lie approximately 3 flanking objects, on average.

than crowding (Rosenholtz, 2016).

Nor do real-world scenes contain less clutter than classic crowding experiments. When we randomly sampled targets in real-world scenes, on average approximately 3 flanking objects lay within critical distance of the target (Fig. 5). This number almost certainly underestimates the amount of crowding-inducing clutter, since it counts only flanking *objects*, narrowly defined. In comparison to classic crowding experiments, this should provide sufficient clutter to induce crowding.

Rather, peripheral object recognition in real-world images is likely better than expected because normally occurring context provides a cue to object identity that mitigates the effects of crowding. The informativeness of context means that rather than isolating an object, ideally one should "integrate" information from beyond its boundaries. In fact, we found that context alone can be as useful as the object itself. Interestingly, we found that the beneficial contribution of integration seems to saturate around Bouma's radius. Whereas the traditional crowding hypothesis would suggest that this integration sabotages recognition, we find the opposite effect. Intuitively, it makes sense that the benefit of context would, on average, plateau for larger window sizes. While a distant car may make a target object somewhat more likely to be a car, or distant water make the target more likely a boat, in general an object's identity likely correlates more strongly with nearby objects and materials. A nearby bed helps identify a clock radio, but a bookcase elsewhere in the room provides little additional information.

Of course, in the real world a target object may also be flanked by uninformative clutter. An uninformative flanker effectively "occludes" relevant contextual information. Object recognition needs tolerance to occlusion, but such tolerance is generally thought to be accomplished through robust inference rather than through shrink-wrapping the object (Yuille & Kersten, 2006).

We started with the notion that object recognition involves integration of features belonging to the object, and that crowding is the failure of peripheral processes to restrict integration to only the object. However, several lines of research challenge this view of object recognition. Many successful computational models of object recognition (from the fields of both human and computer vision), utilize features both from within the object and from outside of it (Riesenhuber & Poggio, 1999; Fink & Perona, 2003; Torralba, Murphy, & Freeman, 2004; Dalal & Triggs, 2005; Heitz & Koller, 2008; Zhu, Bichot, & Chen, 2011; Krizhevsky, Sutskever, & Hinton, 2012; Yamins et al., 2014). In addition, considerable research from the study of figure-ground segregation has challenged the notion that segmentation precedes recognition (Peterson, Harvey, & Weidenbacher, 1991; Peterson & Gibson, 1991, 1993, Peterson and Gibson, 1994a, 1994b; Peterson, 1994; Vecera & Farah, 1997; Navon, 2011). Nonetheless, the segmentation-first view of object recognition has persisted in various forms.

Previous theories have often focused on integration beyond object boundaries as the culprit in crowding. However, crowding phenomena may instead result from the particular features integrated by the visual system, and from the information lost in the process. Our results support the latter hypothesis. We have demonstrated that one should not conceptualize crowding as a failure to select only the object, since such selection is not ideal. Nor does it seem that some other mechanism extracts contextual information and provides it to recognition mechanisms that process the isolated object. We find only a weak and non-significant benefit to presenting object and context separately; the "cost" of not shrink-wrapping the target is small or non-existent. Rather, mechanisms with large receptive fields, extending beyond object boundaries, may jointly process features of both object and context. The detrimental effects of crowding then arise from the nature of the "integration", i.e. how peripheral vision encodes its inputs. Given the importance of context, integration over a sizeable region, rather than being a failure of peripheral vision, actually makes sense.

Acknowledgments

The authors would like to thank Emily Park and Lavanya Sharan with helping data collection and processing of Experiment 3. The work described here was funded in part by IIS-1607486 to Ruth Rosenholtz, as part of the NSF/NIH/ANR/BMBF/BSF Collaborative Research in Computational Neuroscience Program award to R.R. and Christoph Zetzsche. Furthermore, Maarten Wijntjes was supported by a grant from the Netherlands Organisation for Scientific Research (NWO).

Author contributions

M.W.A. Wijntjes and R. Rosenholtz developed the study concept. Both authors contributed to the study design. M.W.A. Wijntjes performed testing and data collection in Experiment 1 and 2, R. Rosenholtz oversaw data collection by E. Park (student research assistant) in Experiment 3. M.W.A. Wijntjes performed data analysis. Both authors drafted the manuscript and both approved the final version of the manuscript for submission.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cognition.2018.06.015.

References

Bernard, J.-B., & Chung, S. T. L. (2011). The dependence of crowding on flanker complexity and target-flanker similarity. *Journal of Vision, 11*(8), 1. https://doi.org/10.1167/11.8.1.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*(5241), 177–178.

Chakravarthi, R., & Cavanagh, P. (2009). Recovery of a crowded object by masking the flankers: Determining the locus of feature integration. *Journal of Vision, 9*(10) 4, 1–9.

Cox, D., Meyers, E., & Sinha, P. (2004). Contextually evoked object-specific responses in human visual cortex. *Science, 304*(5667), 115–117.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR '05) (pp. 886–893).

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193–222.

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memory, and speeds search. *Scientific Reports, 7*, 16471.

Fink, M. & Perona, P. (2003). Mutual boosting for contextual inference. In Proc. NIPS.

Greene, M., Oliva, A., Wolfe, J., & Torralba, A. (2010). What's behind the box? Measuring scene context effects with Shannon's guessing game on indoor scenes. *Journal of Vision, 10*(7), 1259.

Geisler, W. S., & Perry, J. S. (2009). Contour statistics in natural images: Grouping across occlusions. *Visual Neuroscience, 26*(1), 109–121.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383*(6598), 334–337.

Heitz, G. & Koller, D. (2008). Learning spatial context: Using stuff to find things. In Proc.

ECCV.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology, 43*(3), 171–216.

Kosslyn, S. M. (1987). Seeing and imagining in the cerebral hemispheres: A computational approach. *Psychological Review, 94*, 148–175.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Levi, D. (2008). Crowding – An essential bottleneck for object recognition: A mini-review. *Vision Research, 48*(5), 635–654.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in extrastriate cortex. *Science, 229*, 782–784.

Navon, D. (2011). The effect of recognizability on figure–ground processing: Does it affect parsing or only figure selection? *The Quarterly Journal of Experimental Psychology, 64*(3), 608–624. https://doi.org/10.1080/17470218.2010.516834.

Neisser, U. (1967). *Cognitive psychology.* New York: AppletonCentury-Crofts.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520–527.

Parkes, L., Lund, J., Angelucci, A., Solomon, J., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Palmer, S. E., & Rock, I. (1994a). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review, 1*, 29–55.

Palmer, S. E., & Rock, I. (1994b). On the nature and order of organizational processing: A reply to Peterson. *Psychonomic Bulletin & Review, 1*, 515–519.

Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology, 18*(4), 445–451.

Pelli, D., & Tillman, K. (2008). The uncrowded aperture of object recognition. *Nature Neuroscience, 11*(10), 1129–1135.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12), 12.

Peterson, M. A., Harvey, E. M., & Weidenbacher, H. J. (1991). Shape recognition contributions to figure-ground reversal: Which route counts? *Journal of Experimental Psychology: Human Perception and Performance, 17*(4), 1075.

Peterson, M. A., & Gibson, B. S. (1991). The initial identification of figure-ground relationships: Contributions from shape recognition routines. *Bulletin of the Psychonomic Society, 29*, 199–202.

Peterson, M. A., & Gibson, B. S. (1993). Shape recognition contributions to figure-ground organization in three-dimensional displays. *Cognitive Psychology, 25*, 383–429.

Peterson, M. A., & Gibson, B. S. (1994a). Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. *Perception & Psychophysics, 56*, 551–564.

Peterson, M. A., & Gibson, B. S. (1994b). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science, 5*, 253–259.

Peterson, M. A. (1994). The proper placement of uniform connectedness. *Psychonomic Bulletin and Review, 1*, 509–514.

Reicher, G. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*(2), 275–280.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Rosenholtz, R. (2014). Texture perception. In J. Wagemans (Ed.), Oxford Handbook of Perceptual Organization. Oxford, U.K.: Oxford University Press. Online publ., July 2014.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*(1), 437–457.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1), 157–173.

Schwarzlose, R. F., Baker, C. I., & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience, 25*(47), 11055–11059.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In D. V. Blake, A. M. Uttley, editors, Proceedings of the symposium on mechanisation of thought processes, London (pp. 511–529).

Torralba, A., Murphy, K., & Freeman, W. (2004). Contextual models for object detection using boosted random fields. In Proc. NIPS.

Vecera, S. P., & Farah, M. J. (1997). Is visual image segmentation a bottom-up or an interactive process? *Perception & Psychophysics, 59*(8), 1280–1296.

Wheeler, D. (1970). Processes in word recognition. *Cognitive Psychology, 1*, 59–85.

Whitney, D., & Levi, D. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences, 15*(4), 160–168.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 3485–3492).

Yamins, D., Hong, H., Cadieu, C., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yeshurun, Y., & Rashal, E. (2010). Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision, 10*, 1–12.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.

Zhu, C., Bichot, C. E., & Chen, L. (2011). Visual object recognition using daisy descriptor. In Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME 2011), Barcelona, Spain, 1–6.