Do target detection and target localization always go together? Extracting information from briefly presented displays

Ann J. Carrigan^{1,2,3} · Susan G. Wardle^{1,2} · Anina N. Rich^{1,2,3}

Published online: 19 June 2019
© The Psychonomic Society, Inc. 2019

Abstract

The human visual system is capable of processing an enormous amount of information in a short time. Although rapid target detection has been explored extensively, less is known about target localization. Here we used natural scenes and explored the relationship between being able to detect a target (present vs. absent) and being able to localize it. Across four presentation durations (~ 33–199 ms), participants viewed scenes taken from two superordinate categories (natural and manmade), each containing exemplars from four basic scene categories. In a two-interval forced choice task, observers were asked to detect a Gabor target inserted in one of the two scenes. This was followed by one of two different localization tasks. Participants were asked either to discriminate whether the target was on the left or the right side of the display or to click on the exact location where they had seen the target. Targets could be detected and localized at our shortest exposure duration (~ 33 ms), with a predictable improvement in performance with increasing exposure duration. We saw some evidence at this shortest duration of detection without localization, but further analyses demonstrated that these trials typically reflected coarse or imprecise localization information, rather than its complete absence. Experiment 2 replicated our main findings while exploring the effect of the level of “openness” in the scene. Our results are consistent with the notion that when we are able to extract *what* objects are present in a scene, we also have information about *where* each object is, which provides crucial guidance for our goal-directed actions.

Keywords Visual perception · Scene perception · Object recognition

Vision is fast: As soon as our eyes open, we get an impression that we can see everything around us. Early findings suggested that the basic meaning of natural scenes (e.g., classification as outdoor vs. indoor scenes) can be extracted after an exposure of only 100 ms (Potter, 1976; Potter & Faulconer, 1975). Further studies using backward masking to precisely control display duration showed that observers are above chance at categorizing scenes at the superordinate (e.g., natural vs. manmade) and basic (e.g., coast vs. city) levels after exposure durations as short as 20 ms (Greene & Oliva, 2009; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). In addition,

when primed with an object category (e.g., animal or truck), these objects can be detected when observers are shown scenes for only 20–25 ms, albeit with no backward mask (Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). It has been proposed that this remarkable ability is due to extensive experience with domain-specific types of scenes (Drew, Evans, Vö, Jacobson, & Wolfe, 2013).

Although there is evidence that considerable information, such as scene category and target/object detection, is extracted in the initial glance at an environment, goal-directed actions also require information about *where* target objects are within the environment. Despite the importance of location information for our successful interactions with the world, we know much less about the early stages of localization than of detection in brief displays. In particular, there is debate about the extent to which detection and localization are separable. This is important because localization information is thought to require selective processing, which has severe capacity limits, whereas at least in studies involving medical images as stimuli detection may be possible on the basis of “gist” or nonselective processing (e.g., Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Evans, Haygood, Cooper, Culpan

✉ Ann J. Carrigan
ann.carrigan@mq.edu.au

¹ Perception in Action Research Centre & Department of Cognitive Science, Macquarie University, Sydney, Australia

² ARC Centre of Excellence in Cognition & Its Disorders, Macquarie University, Sydney, Australia

³ Centre for Elite Performance, Expertise, and Training, Macquarie University, Sydney, Australia

& Wolfe, 2016; Wolfe, Võ, Evans, & Greene, 2011). Thus, determining the relative time courses of information processing that lead to detection versus localization is an important empirical question with implications for major theories of visual search.

The term “gist” is not entirely clear. The general sense in the natural-scenes and medical-imaging literatures, however, is that gist is something that is extracted rapidly, in a global fashion, without requiring selection (Brennan, Gandomkar, Ekpo, Tapia, Trieu, et al., 2018; Oliva, 2005). If an object can be detected but not localized, this implies that initial early detection might be based on global, nonselective processing (Evans et al., 2016). In contrast, it seems unlikely that specific location information is available in this global signal. Indeed, decades of visual search research has demonstrated that localizing (and detecting) targets that do not have unique features (i.e., are not “pop-out” targets) requires attentive processing (Wolfe, 1994; Wolfe & Horowitz, 2017). Instead, it has been suggested that “gist” may *guide* the way observers view the scene (Loftus & Mackworth, 1978; Oliva, Torralba, Catehano, & Henderson, 2003) and therefore facilitate efficient object recognition by guiding selective attention to target locations (Davenport & Potter, 2004).

Although there has been little study of localization specifically in natural scenes, claims of a dissociation between detection and localization information have been made in the medical-imaging literature. Some studies have reported that radiologists can detect, but not locate, a cancer in a mammogram in 250 ms (Brennan et al., 2018; Evans et al., 2013; Evans et al., 2016). The basis of these findings, however, is a *lack* of significant performance on a localization task with small but significant “above-chance” performance on a detection task. This might be problematic for at least two reasons: First, frequentist statistics do not allow interpretation of a null finding ($p > .05$) as support for the null hypothesis; second, there is no image-level analysis to determine the degree to which a few images might contribute to the apparent dissociation.

We previously have addressed these issues in a study in which we tested radiologists with mammograms presented for 250 ms (Carrigan, Wardle, & Rich, 2018). We found that radiologists could both detect and localize a mass “above chance,” but, more importantly, we demonstrated that on trials in which it appeared detection was successful and localization was unsuccessful, there were alternative explanations for the apparent dissociation: Either there was coarse information about location (responses clustered around the border of the mass) or a distractor area of the image had been mistakenly identified as a target (responses clustered within another, non-target area of the image). Although these results might provide an alternative explanation for some of the medical-imaging results, in other studies there was no specific mass to be located, yet radiologists were above chance in classifying the associated images as “abnormal” (e.g., Evans et al., 2016). In

these cases, only global, non-location-specific signals could be driving the effect, as there was nothing to localize. Overall, within the medical imaging literature, considerable debate has surrounded the question of whether, when there is sufficient information to detect a target, there is also information about its location, or whether these tasks can be dissociated.

A study on change detection in faces has provided evidence of a dissociation between detection and localization (Howe & Webb, 2014). Howe and Webb showed observers a photograph of a face for 1.5 s, followed by a 1s blank, and then another version of the same photograph with a single changed feature (e.g., removal of glasses). Observers were asked to indicate whether a change had occurred and, if so, to select the change from a list of nine possible options. The results showed that observers could sometimes detect that a change had occurred without identifying the specific change, even when taking into account potential correct guesses (an important innovation of this study). The authors suggested that the apparent lack of information about the identity of the change might reflect low precision in the location. In contrast, other change detection studies have shown that the detection of a change is accompanied by knowledge of the change location, and that this performance is driven by feature salience (Mitroff & Simons, 2002; Rensink, O’Regan, & Clark, 1997).

One challenge with using either medical images or faces to study detection and localization is that the images themselves carry information that could guide attention to particular locations. For example, in a mammogram, an expert might be more likely to attend to specific regions that are more likely to contain a mass. In the natural-scenes literature, this type of guidance has been called “scene-based guidance” and suggests that our search for objects is influenced by their expected locations (e.g., a toaster on the kitchen counter; Davenport & Potter, 2004; Wolfe et al., 2011). Thus, in these paradigms, it is hard to study the initial brief processing of a stimulus separate from rapid guidance mechanisms that we automatically use to increase our efficiency at perceiving visual displays. Here we addressed this question by using an artificial target that was not related to the image.

The overarching goal of the present study was to investigate the time course of target detection and localization in brief presentations. We used natural scenes in order to retain the expectation that the information could be extracted rapidly from a brief display, and so that we could independently verify that the “gist” had been processed, but we used a Gabor as a target, to avoid any scene-based guidance. Our first aim was to evaluate whether the localization information about a target is accessible at short durations. We then used another feature of natural scenes, usually defined as “openness,” to test whether this localization information was only extracted when the Gabor was not embedded within complex visual information.

To validate our range of durations, we first confirmed that participants could do scene categorization (natural vs.

manmade) for the background scenes at the shortest experimental duration (Exp. 1A). We then compared detection and localization performance for a Gabor target embedded in a range of natural scenes at brief exposure durations between 33 and 199 ms (Exp. 1B). To test the idea that location information might be present but less precise at brief durations (Howe & Webb, 2014), we included a left (L)-versus-right (R) localization task as well as the instruction to “click on the location with the mouse.” Finally, in Experiment 2, we explored the effect of openness by testing target detection and localization in “open” versus “closed” scenes.

Experiment 1

Experiment 1A was a scene categorization task (natural vs. manmade), designed to verify that the overall “gist” of the background scenes could be extracted at the shortest experimental duration (~33 ms) used in our paradigm. The aim of Experiment 1B was to test whether durations between 33 and 199 ms¹ resulted in sufficient processing to support detection along with localization of a Gabor target embedded in the natural scenes. Using a two-interval forced choice (2IFC) paradigm, natural scenes were presented at one of four durations (33–199 ms) on each trial, with a Gabor target randomly located within one of the two scenes. The participants were asked to report which scene contained the target and then where the target was located within the scene.

Method

Participants

Thirty participants (22 females, 8 males; age range 19–55 years, $M = 31.47$ years, $SD = 8.81$) were recruited from Macquarie University. All participants gave informed consent, reported normal or corrected-to-normal vision, and were financially reimbursed for their time. The study was approved by the Macquarie University Human Research Ethics Committee (Medical Sciences). The data for two observers were excluded due to technical issues, leaving 28 datasets for analysis.

Stimuli and apparatus

Natural-scene stimuli were classified into the scene categories defined by Oliva and Torralba (2001), which are available at <http://cvcl.mit.edu/database.htm> (see Fig. 1). A total of 160 photographic images of natural scenes comprising two superordinate categories (natural and manmade) were selected from

an internet search using Google Images. The natural and manmade categories comprised four basic-level categories (20 images in each): coast, mountain, open country, and forest for the natural superordinate category, and tall building, highway, city center, and street for the manmade superordinate category. The images were converted to grayscale and downsized to subtend $23^\circ \times 15^\circ$ of visual angle.

The target was a Gabor patch with the following parameters: orientation 45° , spatial frequency 0.5 cycles/deg, diameter 3.8° , Michelson contrast 0.2. The target image appeared in a different random location within a scene (constrained to fully appear within the borders of the display) and was present on all trials in either Scene 1 or Scene 2 (see Fig. 2).

The participants sat at a viewing distance of approximately 70 cm in a dimly lit, windowless laboratory at Macquarie University, Sydney. The stimuli were presented with MATLAB 8.2 using Psychtoolbox 3 (Kleiner, Brainard, & Pelli, 2007) and were displayed on a 27-in. Samsung SyncMaster SA950 LCD monitor ($1,920 \times 1,080$, 120 Hz).

Procedure

Experiment 1A: Scene categorization task This task was to verify that scene categorization (natural vs. manmade) was possible for the background scenes at the shortest experimental duration. We used a single-factor (superordinate category: natural, manmade) within-subjects design. Each trial began with a fixation point for 498 ms, followed by a scene from one of the superordinate categories, displayed in the center of the screen for 33 ms. This was followed by a backward $1/f$ noise mask for 249 ms. Participants categorized the scene by its superordinate category (manmade vs. natural) with a key press as accurately and quickly as possible (see Fig. 3). Participants were given ten practice trials at a longer scene presentation duration of 398 ms, to familiarize them with the task, before completing the experimental trials.

Experiment 1B: Target detection and localization task For the main experimental task, we used a 2 (location task: exact click, L vs. R) $\times$ 4 (duration: 33, 58, 116, 199 ms) within-subjects design. Initially the participants were shown a picture of the target, to familiarize them with a Gabor, and then were given eight practice trials (two per duration) with feedback. Each trial began with a fixation point for 498 ms, followed by Scene 1 (33–199 ms, constant within a block) and a backward $1/f$ noise mask (249 ms), and then by Scene 2 (same duration as Scene 1) and a $1/f$ noise mask (249 ms). Observers made a 2IFC decision with a key press regarding whether the target had been present in Scene 1 or Scene 2. Following this detection response, they were presented with a blank screen and asked one of two localization questions (in separate blocks, order counterbalanced across participants). The same images were presented for both localization tasks, and target location

¹ These durations represent the mean durations as a function of screen refresh rate, rather than gray–gray timings. Thus, they are approximate only.

Fig. 1 Example scenes for the superordinate categories and for the four basic categories within each

was independent of image identity (randomly shuffled). In one block of trials, participants were asked to click on the exact location of the target on the blank screen using the mouse. The location of the Gabor was random and was only constrained to appear within the borders of the display. In the other localization task, they were instead asked whether the target had appeared on the left or the right side of the screen, and they responded using a key press. Here the Gabor was not only constrained to appear within the borders of the display, it was also constrained to fully appear within two invisible bounding boxes on either the L or R of the screen. This localization task required a coarser judgment of the target's location in order to answer correctly, as compared to the more difficult exact-click task. The response keys for the L/R localization task were the same as the keys used for the detection task (the “z” and “m” keys; see Fig. 4).

On each trial, both scenes were selected from the same superordinate category (e.g., natural or manmade), but the basic category was random (e.g., both could be from the same category or from different categories within the superordinate category). Fifty percent of trials had the Gabor in Scene 1, and 50% in Scene 2, randomly interleaved within a block. The target location was randomized, with the restriction that it

was not clipped by the screen edge and that it appeared in the left half for 50% of trials, and in the right for the other 50% of trials. Duration order was blocked and counterbalanced across participants. The participants performed 160 experimental trials for each localization task. The experiment was self-paced, and the participants initiated each trial with a key press. The observers saw the same images in each task, but in a different randomized order (80 natural and 80 manmade scenes in each version of the task), giving a total of 320 trials across the experiment. They were instructed to respond as accurately as possible, and there was a minimum 15-s rest period every 40 trials. Participants were not provided with any feedback during the experimental tasks (see Fig. 4).

Results and discussion

Analysis

All analyses were performed using the Statistical Software for the Social Sciences (IBM SPSS version 25). The 95% confidence intervals (CIs) were calculated as $\bar{X} \pm 1.960(\sigma/\sqrt{n})$.

Fig. 2 Example of a city (manmade) scene with the embedded target Gabor (diameter = 3.8°)

Fig. 3 Example of a trial for the superordinate scene categorization task in Experiment 1A. The participant responded whether the image was a manmade or natural scene

Experiment 1A: Scene categorization The purpose of this experiment was to independently verify that sufficient information (gist or scene statistics) about scene categories (natural vs. manmade) could be extracted using our parameters at our shortest duration (33 ms), as had been demonstrated by others (Greene & Oliva, 2009; Joubert et al., 2007). We used a measure of sensitivity, d' , as our dependent measure. The mean d' for the categorization tasks was 2.29 ($SD = 0.56$, range = 1.36–3.65). A single-sample t test on d' relative to chance ($d' = 0$) demonstrated that performance was better than chance in categorizing the scenes as manmade vs. natural at the shortest experimental duration, $t(27) = 21.8$, $p < .0001$. This

replicates previous findings that sufficient visual information to categorize scenes is available from < 50 ms presentations.

Experiment 1B: Detection performance Figure 5 shows detection performance for the Gabor target at exposure durations from 33 to 199 ms. We calculated d' as a function of target presence in Scene 1 or Scene 2. A two-way repeated measures analysis of variance (ANOVA) on d' with the factors localization task (exact click, L vs. R) and duration (33, 58, 116, 199 ms) revealed no main effect of localization task, $F(1, 27) = 0.38$, $p = .541$; a significant main effect of duration,

Fig. 4 Example of a trial for the target detection and localization tasks, with the Gabor target in Scene 1. Note that the order of the blocks was counterbalanced across participants and that the target was equally likely to appear in Scene 1 or Scene 2

Fig. 5 Detection performance, collapsed across location tasks. The dotted line represents chance, and the error bars represent 95% confidence intervals

$F(3,81) = 16.25$, $p < .0001$, $\eta_p^2 = .38$ (Greenhouse–Geisser-corrected); and no significant localization task $\times$ duration interaction, $F(3, 81) = 1.29$, $p = .27$, $\eta_p^2 = .046$ (Greenhouse–Geisser-corrected). The detection task was identical for the two location tasks, and detection was performed prior to the localization task. It is therefore not surprising that we see only an effect of improved performance as duration increased.

Our primary question for target detection was whether at each duration there would be sufficient information to support detection. We therefore collapsed the detection data across localization tasks and evaluated detection performance using single-sample t tests at each duration relative to a chance level of $d' = 0$. Figure 5 shows the data for detection performance (d') collapsed across localization tasks. To maintain an overall Type I error rate of .05, a Bonferroni correction was used (testwise alpha was set at $p = .0125$). Detection performance

was significantly above chance at each exposure duration [33 ms, $t(27) = 3.98$, $p < .0001$; 58 ms, $t(27) = 16.04$, $p < .0001$; 116 ms, $t(27) = 27.51$, $p < .0001$; 199 ms, $t(27) = 34.23$, $p < .0001$].

Localization performance

The results above show that the target Gabor could be detected in complex natural scenes with presentations even as brief as 33 ms. These results are consistent with the previous literature, which has shown accurate object detection within scenes at exposure durations between 20 and 25 ms (Thorpe et al., 1996; VanRullen & Thorpe, 2001). Next, we investigated whether the target could also be located at these very brief presentation durations. Our dependent variable was percentage of localizations correct. We analyzed the total percentage of localizations correct across all trials (regardless of whether detection was correct or incorrect), splitting the analysis by the two localization tasks, L vs. R (coarse localization) and exact click (fine localization), which were presented in separate blocks.

Localization in the coarse L vs. R task This localization task was a 2AFC: left or right. Thus, chance is 50%. Figure 6 shows a clear pattern of increasing localization performance with increasing duration for both localization tasks. Our key question related to whether there was sufficient information at each duration to support localization. We therefore used single-sample t tests on percentages of correct localization responses (Bonferroni-corrected, testwise alpha set at .0125). For the L vs. R task (chance = 50%), this showed that performance was above chance for all durations [Fig. 6, black line;

Fig. 6 Accuracy on the two localization tasks. The dotted lines represent chance (black: left vs. right; color: exact click). The error bars represent 95% confidence intervals. Due to the different features of each task, it is not possible to directly compare performance

33 ms, $t(27) = 5.07, p < .0001$; 58 ms, $t(27) = 11.10, p < .0001$; 116 ms, $t(27) = 27.49, p < .0001$; 199 ms, $t(27) = 15.87, p < .0001$].

Localization in the exact-click localization task This localization task required a precise mouse click on the target. We calculated chance on the basis of the number of possible non-overlapping locations of the target Gabor within the image (chance = 16.67%). To allow for some imprecision in reporting the remembered target location, we defined a region of acceptance (ROA) for scoring a mouse click as a correct localization of twice the Gabor diameter, or 7.6° centered on the Gabor location, defined as a square boundary around the Gabor (i.e., the original matrix size of the 2-D sine wave prior to applying the cosine window in Matlab). This method has been utilized in other applied perception studies to account for slight imprecision errors (e.g., Carrigan et al., 2018; Evans et al., 2013), although the exact ROA chosen is essentially arbitrary. Importantly, we prespecified what this ROA would be and used the same one for all analyses. Our dependent variable was the percentage of localizations correct. Again, our key question related to whether there was sufficient information at each duration to support localization, using single-sample t tests on percentages of correct localization responses (Bonferroni-corrected, testwise alpha was set at $p = .0125$). For the exact click task (chance = 16.67%), this showed that performance was above chance for all durations [Fig. 6; 33 ms, $t(27) = 6.3, p < .0001$; 58 ms, $t(27) = 14.78, p < .0001$; 116 ms, $t(27) = 26.75, p < .0001$; 199 ms, $t(27) = 51.21, p < .0001$].

The localization results show that the participants could accurately localize a Gabor target on some trials for

presentation durations as brief as 33 ms. Specifically, participants performed significantly better than chance for all durations, 33–199 ms, even for the precise-localization task (exact click). In Fig. 6 and the corresponding analysis, all localization trials were included, regardless of whether detection was correct. Figure 7 represents the relative proportions of trials across all four durations when detection and localization were both correct, when detection was correct and localization was incorrect, when localization was correct and detection was incorrect, and when detection and localization were both incorrect. Each experimental trial is represented once in the graph. The proportion of trials on which both detection and localization are correct clearly increases as a function of duration, as one would expect. There also seems to be an effect of duration on the proportion of trials on which detection appears to be correct but localization is incorrect (Fig. 7). This pattern is statistically reliable: One-way repeated measures ANOVAs for each localization task showed significant effects of duration (33, 58, 116, 199 ms) on the proportion of detection-only trials [exact click: main effect of duration, $F(3, 81) = 30.39, p < .0001, \eta_p^2 = .53$; L vs. R: main effect of duration, $F(3, 81) = 20.05, p < .0001, \eta_p^2 = .43$]. As duration increased from 33 to 199 ms, there was a decrease in the proportion of trials on which observers were correct on detection but not location.

At the shortest durations, accuracy on both detection and localization appears lower in absolute performance for the more precise localization task (exact click) than for the coarse localization task (L vs. R); however, note that differences in chance baseline between the two localization tasks (50% for L/R, 16.67% for exact click) limit a direct comparison of between-task performance. In Fig. 7a, for the L-versus-R task,

Fig. 7 Percentages of trials across all four durations, grouped as a function of response profile: trials on which both detection and localization were correct, trials on which detection was correct and localization was incorrect, trials on which localization was correct and

detection was incorrect, and trials on which both detection and localization were incorrect. (a) Left-versus-right task. (b) Exact-click task. The error bars represent 95% confidence intervals

this shows that observers were $\sim 20\%$ correct for “localization only” at the 33 ms duration. This could be due to a keyboard assignment issue, since the keys for the detection response matched those for the localization response (e.g., “z” = Scene 1 and left; “m” = Scene 2 and right). This seems likely to have caused occasional response conflict in the L vs. R task, in which participants might have accidentally pressed the button corresponding to the target location first, instead of the detection as Scene 1 versus Scene 2. We therefore do not interpret these localization-without-detection trials for the L/R task further.

Returning to the summary statistics, the overall finding that observers were greater than chance on both detection and localization shows that even at brief durations, a target embedded in a natural scene can often be spatially localized as well as detected. Therefore, the localization of targets at brief presentation durations may depend on how salient the target is within a particular scene or on features of particular types of scenes, such as the level of a scene’s openness. Previous research has demonstrated that target detection is difficult for target letters (e.g., Henderson, Chanceaux, & Smith, 2009) and for target Gabors located in geographical maps (Rosenholtz, Li, & Nakano, 2007). Here, we explored how a scene’s openness affects target localization at brief durations, by performing an exploratory post-hoc scene analysis on the effect of openness on localization performance within our diverse natural-scene image set.

Models within the computational literature emphasize the importance of global properties, or the distribution of basic features of a scene along with the scene’s spatial layout, for scene recognition. For example, the spatial envelope model (SEM; Oliva & Torralba, 2001) describes the “degree of openness” of a scene. Scenes vary with regard to their degree of openness, ranging from low, where a scene comprises many visual characteristics, to high, where scenes often contain a horizon and are vast, containing minimal visual items. For the predetermined categories from Oliva and Torralba, we grouped three categories with a low degree of openness as “closed” categories (mountain, forest, and city), and three categories with a high degree of openness as “open”

categories (coast, open country, and highway). Since this was a post-hoc analysis, there were unequal numbers of open versus closed scenes.

If localization is affected by a scene’s degree of openness, there should be a difference between localization performance on correct-detection trials in open versus closed scenes. We divided the scenes into categories according to their level of openness (open, closed) and conducted a two-way repeated measures ANOVA on the two localization tasks separately, with the factors scene (open, closed) and duration (33, 58, 116, 199 ms). See Fig. 8.

For the L vs. R task, we found a significant main effect of scene (open vs. closed), $F(1, 27) = 70.16, p < .0001, \eta_p^2 = .72$; a significant main effect of duration, $F(3, 81) = 176.13, p < .0001, \eta_p^2 = .87$ (Huynh–Feldt-corrected); and a significant scene $\times$ duration interaction, $F(3, 81) = 5.99, p = .001, \eta_p^2 = .18$ (Greenhouse–Geisser-corrected). Similarly, for the exact-click task, there was a significant main effect of scene (open vs. closed), $F(1, 27) = 87.49, p < .0001, \eta_p^2 = .76$; a significant main effect of duration, $F(3, 81) = 259.69, p < .0001, \eta_p^2 = .91$; and a significant scene $\times$ duration interaction, $F(3, 81) = 7.96, p < .0001, \eta_p^2 = .23$. The interactions suggest that openness does influence the degree to which location information is available. Experiment 2 was designed to follow up this initial analysis by experimentally manipulating the degree of openness in natural scenes, to systematically examine its effects on target detection and localization at brief durations.

Experiment 2

Experiment 2 was designed to systematically investigate the influences of openness in natural scenes on target detection and localization. We manipulated openness using the computational definition of the degree of openness for natural scenes, with the following superordinate categories: open (coast, open country, highway) and closed (forest, mountain, city) (Oliva & Torralba, 2001), as we outlined in Experiment 1. The post-hoc analysis of openness for Experiment 1 had been limited, due

Fig. 8 Localization accuracy as a function of exposure duration, partitioned into open versus closed scene categories in a post-hoc analysis. Note that only trials on which the target was correctly detected are included. The error bars represent 95% confidence intervals

to the small and unbalanced set of open and closed natural scenes across the durations, since the experiment had not been designed for this purpose. Thus, in Experiment 2 we increased the number of scenes in each category to be equal across open/closed scene types in each duration, and we examined detection and localization performance as a function of duration and scene type.

Method

Participants

Thirty participants (24 females, 6 males; age range 18–58 years, $M = 28$ years, $SD = 8.5$) were recruited from Macquarie University. Five had participated in Experiment 1 (15 months prior). All gave informed consent, reported normal or corrected-to-normal vision, and were reimbursed for their time. The study was approved by the Macquarie University Human Research Ethics Committee (Medical Sciences).

Stimuli

We used 240 natural-scene images in total. Half of the images of scenes were repeated from Experiment 1, and the other half were new images selected from the Web using Google Images, to reach a balanced number across the open/closed categories. The open (coast, open country, highway) and closed (city, forest, mountain) categories contained 40 exemplars for each basic category (see Fig. 9). The images were converted to grayscale and downsized, and they subtended $23^\circ \times 15^\circ$.

Procedure

The stimuli and setup were identical to those of Experiment 1, with the following modifications.

Target detection and localization task We used a scene (open, closed) $\times$ duration (33, 58, 116, 199 ms) within-subjects design. To reduce the effects of possible keyboard assignments, we changed the detection response keys to “left arrow” for Scene 1 and “right arrow” for Scene 2. We also set the possible target locations to be 12 locations, which were equally prevalent. Following the detection response, participants were presented with a screen with a 4×3 grid of squares and were asked to indicate the location of the target by using the mouse to select one of the 12 possible locations. Chance could therefore be calculated precisely as 8.3% ($1/12$), replacing the left/right and exact-click tasks used in Experiment 1. For each interval within a trial, the scenes were randomly selected from within the same open or closed category. Duration order was counterbalanced across participants. After 12 practice trials with feedback (three at each duration), participants were presented 240 experimental trials (60 at each duration; 30 trials per condition; see Fig. 10).

Results and discussion

Detection performance

First, we evaluated whether there were any differences in detection performance between the open and closed scenes. Figure 11 shows that performance was higher for open than for closed scenes, and that this difference increased with

Fig. 9 Example scenes for the open and closed categories

Fig. 10 Example of a trial in Experiment 2 for the target detection and localization tasks. A Gabor target is in Scene 1 of an open scene

longer duration. We calculated d' as a measure of detection performance and conducted a two-way repeated measures ANOVA with the factors scene (open, closed) and duration (33, 58, 116, 199 ms) on the mean d' values. There was a main effect of scene, $F(1, 29) = 202.24, p < .0001, \eta_p^2 = .88$; a significant main effect of duration, $F(3, 87) = 146.42, p < .0001, \eta_p^2 = .84$; and a significant scene $\times$ duration interaction, $F(3, 87) = 10.86, p < .0001, \eta_p^2 = .27$. A Bonferroni correction for multiple comparisons ($\alpha = .05/4 = .0125$) was applied to post-hoc analyses. The interaction was due to d' being significantly higher in the open than in the closed scenes at all durations ($p < .0001$) except for 33 ms ($p = .078$) (see Fig. 11).

We next examined whether the localization performance was influenced by scene type. As for Experiment 1, trials were partitioned into those in which detection and localization were correct or incorrect, to examine the response distribution as a function of duration and scene type. Figure 12 shows the total proportions of trials across all four durations when detection and localization were both correct, when only detection was

correct and localization was incorrect, when localization was correct and detection incorrect, and when both detection and localization were incorrect. Each experimental trial is represented once in the graph. For trials on which detection was correct but localization was incorrect, there appears to be an effect of duration (Fig. 12). This pattern is statistically reliable: One-way repeated measures ANOVAs for both scene categories showed significant effects of duration (33, 58, 116, 199 ms) on detection-only trials [closed scenes: main effect of duration, $F(3, 81) = 49.3, p < .0001, \eta_p^2 = .63$; open scenes: main effect of duration, $F(3, 81) = 97.17, p < .0001, \eta_p^2 = .77$]. The proportion of trials on which detection but not localization was correct decreased as a function of increasing duration.

Localization performance

To investigate the degree to which a briefly presented target can be spatially located when embedded in open versus closed scenes, we analyzed localization accuracy across all trials,

Fig. 11 Detection performance for open versus closed scenes. The dotted line represents chance, and the error bars represent 95% confidence intervals

Fig. 12 Percentages of trials across all four durations, grouped as a function of response profile: trials on which detection and localization were both correct, trials on which detection was correct and localization was incorrect, trials on which localization was correct and detection was

incorrect, and trials on which detection and localization were both incorrect. (a) Open scenes. (b) Closed scenes. The error bars represent 95% confidence intervals

regardless of whether detection was correct or incorrect. Figure 13 shows that localization accuracy appears to be higher for the open than for the closed scenes, and this improves with duration. To directly evaluate whether there were differences in target localization performance between open and closed scenes across durations, we conducted a two-way repeated measures ANOVA with the factors scene (open, closed) and duration (33, 58, 116, 199 ms) on the mean percentage localization accuracy values. We found significant main effects of scene, $F(1, 29) = 328.48$, $p < .0001$, $\eta_p^2 = .92$, and duration, $F(3, 87) = 374.88$, $p < .0001$, $\eta_p^2 = .93$, as well as a significant scene $\times$ duration interaction, $F(3, 87) = 16.43$, $p < .0001$, $\eta_p^2 = .36$. As in the Experiment 1 exploratory post-hoc analysis of openness, the interaction shows that scene category (open vs. closed) has a different effect on

Fig. 13 Localization performance for the open and closed scene categories. All trials were included in the analysis, regardless of whether target detection was correct. The dotted line represents chance (8.3%). The error bars represent 95% confidence intervals

localization accuracy depending on the duration. Although participants were able to localize a salient target with high levels of accuracy for both the open and closed scenes from around 116 ms, they were more accurate in both detection and localization for open than for closed scenes. Using a Bonferroni correction ($\alpha = .05/4 = .0125$), paired-sample t tests showed that localization accuracy was significantly higher in the open than in the closed scenes at all four durations [33 ms, $t(29) = 3.23$, $p = .003$; 58 ms, $t(29) = 7.42$, $p < .0001$; 116 ms, $t(29) = 13.48$, $p < .0001$; 199 ms, $t(29) = 8.97$, $p < .0001$]. To replicate the findings of Experiment 1, we also conducted single-sample t tests (Bonferroni-corrected) on localization accuracy. As compared to chance performance of 8.3%, observers were above chance at all durations for both the open and closed scenes [open scenes: 33 ms, $t(29) = 8.88$, $p < .0001$; 58 ms, $t(29) = 16.06$, $p < .0001$; 116 ms, $t(29) = 34.98$, $p < .0001$; 199 ms, $t(29) = 54.1$, $p < .0001$; closed scenes: 33 ms, $t(29) = 3.56$, $p = .001$; 58 ms, $t(29) = 11.48$, $p < .0001$; 116 ms, $t(29) = 18.38$, $p < .0001$; 199 ms, $t(29) = 30.23$, $p < .0001$]. Even at the briefest duration, localization performance was above chance for both the open and closed scenes, with a steady rise in performance with increasing presentation duration. These results demonstrate that detection and localization performance are both affected by the degree of openness in a scene.

In the medical perception literature, Evans et al. (2013) and Evans et al. (2016) argued that a global signal supports detection but not localization of abnormalities in a mammogram. It is difficult to determine precisely what apparent detection-without-localization trials represent, since many factors could

account for these data. In a study similar to Evans et al. (2013), we argued that to test this hypothesis we would need to go beyond the summary statistics (e.g., average d' ; Carrigan et al., 2018). We suggest that many of the apparent detection-without-localization trials (localization errors) can be explained by other factors.

Although it is possible that trials on which detection is correct and localization is incorrect might reflect a true dissociation between detection and localization, especially at the shortest duration, there are other plausible explanations. First, the location information may have lacked precision. This could reflect accurate information that was degraded over the maintenance period due to task demands or coarse-level information about location that was maintained perfectly. In either scenario, on these trials observers would respond correctly that the Gabor target was in Scene 1 or Scene 2, but would click on a location that only imprecisely reflected the correct location, resulting in an error.

To test whether our observers had any information about localization on trials on which only the detection response was correct, we compared their localization responses with those of a hypothetical observer who had no information about localization. For this naive hypothetical observer, the best-

performing option would be to click in the center of the screen on every trial. Figure 14 demonstrates the localization errors (when detection was correct) for both the (a) open and (b) closed scenes. The black lines show our participants' actual data as the mean error distances in pixels from the Gabor for each duration, where 0 = correct (x -axis). The color lines represent the hypothetical, naive observers, defined as having no localization knowledge.

Our observers, on average, clicked 200 pixels *closer* to the actual target location than did the hypothetical observer. This is consistent with the theory that coarse or less precise localization information exists in many of the apparent “incorrect” localization responses. This suggests that any true dissociation between detection and localization is a very small effect, and that most such trials can be plausibly explained by other factors, such as imprecise localization information.

General discussion

Observers can report the basic category of a scene (e.g., natural or manmade) and detect targets in both natural scenes (Greene & Oliva, 2009; Potter & Faulconer, 1975; Thorpe

Fig. 14 Distance of localization from the Gabor on correct detection-incorrect localization trials, for both (a) open scenes and (b) closed scenes at the four presentation durations. The color lines represent the mean error

distances in pixels for hypothetical, naive observers; the black lines represent the mean error distances in pixels for the actual observers. The error bars represent 95% confidence intervals

et al., 1996; VanRullen & Thorpe, 2001) and medical images (Carrigan et al., 2018; Evans et al., 2013; Evans et al., 2016; Kundel & Nodine, 1975), at very brief display durations. Although location information is crucial for guiding our interactions with the environment, some findings have suggested that detecting a target and localizing it might not always go together (Evans et al., 2013; Evans et al., 2016; Howe & Webb, 2014). This is consistent with our findings, in which we showed that observers can both detect and localize a Gabor target at brief durations around 33 ms.

In Experiment 1, even for the most difficult exact-click localization task, which requires precise location information, observers had sufficient information from a 33 ms presentation to be able to perform greater than chance. This suggests that fine-grained localization information can be extracted in as little as 33 ms. The results of Experiment 2 showed that this localization information is not as evident in complex scenes with low levels of openness, but observers still have some information about the target location. Together, our results suggest that at least some localization information is available at brief exposure durations that supports rapid target detection.

Empirically determining the relationship between detection and localization is a challenge. Although the results show that a target in a natural scene can be localized with brief displays, this was not the case on all trials; performance was far from ceiling. Additionally, on a proportion of trials detection was correct but localization was incorrect, especially at the briefest durations (33 and 58 ms) and for the more precise exact-click localization task. The challenge in understanding these particular results is that there are several possible interpretations. These trials might truly reflect a dissociation between information for detection and information for localization, as has been suggested in the medical literature, where target detection but not localization was above chance at 500 ms (Evans et al., 2013). However, apparent detection without localization could also occur for a number of other reasons that are not clear from summary statistics alone. First, observers might have had some localization information on these trials, but the information was too coarse to be counted as correct in the localization tasks. This is suggested by the numerically greater occurrence of apparent “detection-without-localization” trials for the more precise exact-click localization task than for the left/right localization judgment, and from our inspection of the localization errors (Exp. 2, Fig. 14). The latter analysis suggests that most of the detection-without-localization trials could be due to response imprecision. Second, on these trials participants might have guessed the correct response (was the target in Scene 1 or Scene 2?), without actually having seen the target on that trial. These “lucky guesses” would presumably be accompanied by a guess on the subsequent localization, which would have only a 1/12 chance of being correct (Exp. 2). We found that the proportion of detection-without-localization trials decreased with

increasing exposure duration, but this could be consistent with either the guessing account or a dissociation of detection from localization information. In our previous study, using medical images and a present/absent task, we were able to calculate a guessing correction, on the basis of a hypothetical observer who only had detection with localization (i.e., no dissociation was possible). In that study, in difficult images, we failed to find evidence of detection without localization beyond what could be accounted for as “lucky guesses” (Carrigan et al., 2018). We have yet to find an appropriate similar correction for the 2IFC task used here, but since we had only small proportions of trials that might have shown detection without localization, and since our analysis of the errors suggests that localization imprecision was responsible for many of these trials, we do not see strong evidence for a dissociation. Our main analyses demonstrate that information supporting both detection and localization can be extracted from very brief displays.

Brennan et al. (2018) proposed that two classes of gist provide, respectively, the “what” and the “where” of information. The first class involves the extraction of basic statistical information such as size and orientation, independent of image characteristics, which allows for detection. However, this class does not appear to be informative regarding where to allocate subsequent fixations. It was proposed that the second class of gist *does* contain spatial layout information, which constrains subsequent fixations. Given the evidence in the medical perception literature and the proposal that gist may also guide scene viewing (Loftus & Mackworth, 1978; Oliva et al., 2003), this raises the possibility that in natural scenes, at brief durations, these two classes of gist can be identified. If we accept the premise posited by Brennan and colleagues that there are two classes of gist, then our results fall into the second class: that localization information is available in the first glance. A further possibility is that there is “gist” without localization and detail, and in parallel a different localization process. Our evidence, from both the present study and Carrigan et al. (2018), suggests that once there is sufficient information to support detection of a target, there is also sufficient information to support localization, albeit with some noise (i.e., imprecision). Whether one calls this a second class of “gist” or a rapid selective process, we see little evidence that detection can be completed in dissociation from localization.

Our main findings demonstrate that a briefly viewed scene can be processed sufficiently for a target to be localized. In our experiments, for some scenes the Gabor would be salient (e.g., the open scenes; Fig. 13), guiding attention effectively. Performance was not at ceiling, however, suggesting that we did not reach the realm of “pop-out” in these displays (Borji & Itti, 2013; Rosenholtz et al., 2007). For other scenes (e.g., the closed scenes), there was less salient information about the target, but some information about localization was still

available. Using this paradigm allowed us to precisely control for low-level visual factors (e.g., target size, contrast, and spatial frequency) as well as semantic congruency. Although we found localization of a target in natural scenes with rapid viewing, it is possible that our displays were not brief enough to dissociate detection and localization, due to the speed of the visual processing of natural scenes. In addition, the alternative interpretations we offer above for why target detection might appear to occur without localization on some trials in our own results also apply to studies with medical images (see Carrigan et al., 2018).

The results of Experiment 2 are consistent with what we know about the concepts that covary (but are not intrinsically related) with enclosure: crowding and visual clutter. In free viewing, studies have shown that with increasing clutter or set size, performance decreases due to crowding (Adamo, Cain, & Mitroff, 2015; Asher, Tolhurst, Troscianko, & Gilchrist, 2013; Rosenholtz, Li, Mansfield, & Jin, 2005; Rosenholtz et al., 2007; Whitney & Levi, 2011). However, it is possible that in our study, the decrease in detection and localization accuracy for the closed scenes was driven by displays with more spatial frequencies along noncardinal directions, rather than because they were low on their degree of openness. A future study teasing apart these points would be beneficial.

Our rapid perception of the world around us is critical for successful interactions. Here we have explored the degree to which the information available in very brief presentations of natural scenes could support not just the detection of a target, but also knowledge of where that target is. Access to location information is crucial for guiding actions or further analysis (e.g., eye movements); we found a tight link between the information supporting detection and localization.

Author note We thank Patrick Nalepka for his assistance with the localization error analysis. This work was supported through an Australian Postgraduate Award to A.J.C. S.G.W. was supported by an Australian NHMRC Early Career Fellowship (APP1072245). This work also was partially supported by an ARC Discovery Project grant to A.N.R. (DP170101780). All authors contributed to the study design. S.G.W. programmed the experiments, and A.J.C. collected and analyzed the data, under the supervision of S.G.W. and A.N.R. A.J.C. drafted the manuscript, and S.G.W. and A.N.R. provided critical revisions. All authors approved the final version of the manuscript for submission.

References

- Adamo, S. H., Cain, M. S., & Mitroff, S. R. (2015). Targets need their own personal space: Effects of clutter on multiple-target search accuracy. *Perception*, 44, 1203–1214. <https://doi.org/10.1177/0301006615594921>
- Asher, M. F., Tolhurst, D. J., Troscianko, T., & Gilchrist, I. D. (2013). Regional effects of clutter on human target detection performance. *Journal of Vision*, 13, 1–15. <https://doi.org/10.1167/13.5.25>
- Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 185–207.
- Brennan, P. C., Gandomkar, Z., Ekpo, E. U., Tapia, K., Trieu, P. D., Lewis, S. J., . . . Evans, K. K. (2018). Radiologists can detect the “gist” of breast cancer before any overt signs of cancer appear. *Scientific Reports*, 8, 8717.
- Carrigan, A. J., Wardle, S. G., & Rich, A. N. (2018). Finding cancer in mammograms: If you know it’s there, do you know where? *Cognitive Research: Principles and Implications*, 3. <https://doi.org/10.1186/s41235-018-0096-5>
- Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, 15, 559–564. <https://doi.org/10.1111/j.0956-7976.2004.00719.x>
- Drew, T., Evans, K., Vö, M. L.-H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? *RadioGraphics*, 33, 263–274. <https://doi.org/10.1148/rq.331125023>
- Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review*, 20, 1170–1175. <https://doi.org/10.3758/s13423-013-0459-3>
- Evans, K. K., Haygood, T. M., Cooper, J., Culpan, A.-M., & Wolfe, J. M. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proceedings of the National Academy of Sciences*, 113, 10292–10297. <https://doi.org/10.1073/pnas.1606187113>
- Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58, 137–176. <https://doi.org/10.1016/j.cogpsych.2008.06.001>
- Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, 9(1):1–8. <https://doi.org/10.1167/9.1.32>
- Howe, P. D. L., & Webb, M. E. (2014). Detecting unidentified changes. *PLoS ONE*, 9, e84490. <https://doi.org/10.1371/journal.pone.0084490>
- Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorisation and object interference. *Vision Research*, 47, 3286–3297.
- Kleiner, M., Brainard, D., & Pelli, D. (2007). What’s new in Psychtoolbox-3? *Perception* 36(EGVP Abstract Suppl), 14.
- Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology*, 116, 527–532.
- Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 565–572. <https://doi.org/10.1037/0096-1523.4.4.565>
- Mitroff, S. R., & Simons, D. J. (2002). Changes are not localized before they are explicitly detected. *Visual Cognition*, 9, 937–968. <https://doi.org/10.1080/13506280143000476>
- Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, 696, 251–258.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene. A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145–175. Retrieved from <http://cvcl.mit.edu/database.htm>
- Oliva, A., Torralba, A., Castelano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing*

- (Vol. 1, pp. 251–253). Piscataway, NJ: IEEE Press. <https://doi.org/10.1109/ICIP.2003.1246946>
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509–522. <https://doi.org/10.1037/0278-7393.2.5.509>
- Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature*, 253, 437–438.
- Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8, 368–373. <https://doi.org/10.1111/j.1467-9280.1997.tb00427.x>
- Rosenholtz, R., Li, Y., Mansfield, J., & Jin, Z. (2005). Feature congestion: A measure of display clutter. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 761–770). New York: ACM Press.
- Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7, 17. <https://doi.org/10.1167/7.2.17>
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381, 520–522. <https://doi.org/10.1038/381520a0>
- VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artefactual objects. *Perception*, 30, 655–688.
- Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15, 160–168. <https://doi.org/10.1016/j.tics.2011.02.005>
- Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1, 202–238. <https://doi.org/10.3758/BF03200774>
- Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behavior*, 1, 0058. <https://doi.org/10.1038/s41562-017-0058>
- Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Science*, 15, 77–84. <https://doi.org/10.1016/j.tics.2010.12.001>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.