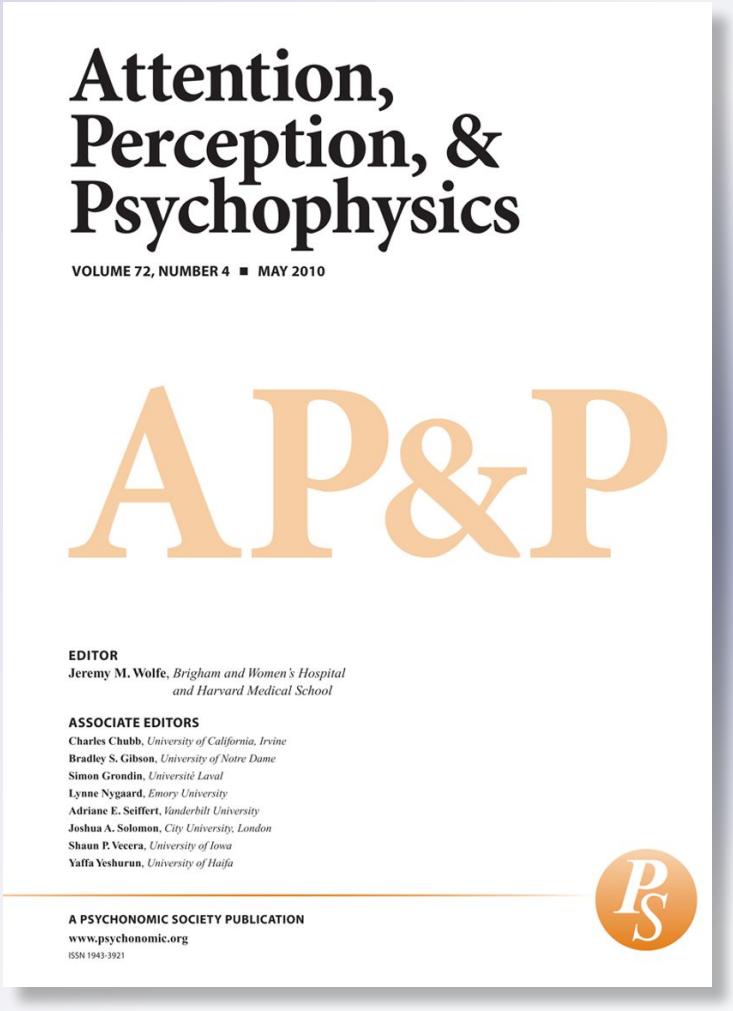
Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm

Melissa L.-H. Võ & John M. Henderson

Attention, Perception, & Psychophysics

ISSN 1943-3921
Volume 73
Number 6

Atten Percept Psychophys (2011)
73:1742-1753
DOI 10.3758/s13414-011-0150-6



Springer

Atten Percept Psychophys (2011) 73:1742–1753
DOI 10.3758/s13414-011-0150-6

Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm

Melissa L.-H. Võ · John M. Henderson

Published online: 24 May 2011

Abstract In the present study, we investigated the influence of object–scene relationships on eye movement control during scene viewing. We specifically tested whether an object that is inconsistent with its scene context is able to capture gaze from the visual periphery. In four experiments, we presented rendered images of naturalistic scenes and compared baseline consistent objects with semantically, syntactically, or both semantically and syntactically inconsistent objects within those scenes. To disentangle the effects of extrafoveal and foveal object–scene processing on eye movement control, we used the flash-preview moving-window paradigm: A short scene preview was followed by an object search or free viewing of the scene, during which visual input was available only via a small gaze-contingent window. This method maximized extrafoveal processing during the preview but limited scene analysis to near-foveal regions during later stages of scene viewing. Across all experiments, there was no indication of an attraction of gaze toward object–scene inconsistencies. Rather than capturing gaze, the semantic inconsistency of an object weakened contextual guidance, resulting in impeded search performance and inefficient eye movement control. We conclude that inconsistent objects do not capture gaze from an initial glimpse of a scene.

Keywords Eye movement control · Naturalistic scenes · Extrafoveal processing · Scene semantics and syntax · Object–scene inconsistency · Flash-preview moving-window paradigm

M. L.-H. Võ (✉)
Visual Attention Lab, Harvard Medical School,
Brigham and Women's Hospital,
64 Sidney Street, Suite 170,
Cambridge, MA 02139, USA
e-mail: mlvo@search.bwh.harvard.edu

J. M. Henderson
University of South Carolina,
Columbia, SC, USA

M. L.-H. Võ · J. M. Henderson
University of Edinburgh,
Edinburgh, UK

Introduction

When we explore our visual world, we tend to move our eyes from one location to another about three to four times per second. These very fast saccadic eye movements are necessary, since much of the visual information we want to process can be acquired only with our fovea, a small, high-resolution region of our retina. At the same time, there is substantial evidence that we can acquire sufficient scene information from the visual periphery that can modulate object processing without the need to foveally process all scene regions (e.g., Davenport & Potter, 2004; for a review, see Henderson & Hollingworth, 1999a). How much semantic processing can be done outside the fovea? Can objects that do not fit the gist of the scene capture gaze? Here, we investigated the extent to which object–scene relationships can be processed in the visual periphery by measuring their effects on eye movement control. In order for semantic inconsistency in the visual periphery (imagine a fire hydrant in the corner of your kitchen) to affect eye movements, the semantic fit between peripherally presented objects and the scene context has to be analyzed. This again requires that objects outside the fovea have to be, at least partially, identified prior to their fixation. While there is good evidence that the global gist of a scene can be extracted from only a short glance (e.g., Castelhano &

Henderson, 2008; Oliva & Schyns, 2000; Oliva & Torralba, 2006; Potter, 1975; Thorpe, Fize, & Marlot, 1996), it is less clear that sufficient computation of object identities can be accomplished without foveated object processing (Fei-Fei, Iyer, Koch, & Perona, 2007; Henderson & Hollingworth, 1999a, 1999b; Hollingworth & Henderson, 1998).

Evidence in support of semantic processing in the visual periphery has come from studies in which short presentations of scenes were used while effects on attention were measured as a function of critical objects embedded in these scenes (Gordon, 2004, 2006; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; VanRullen & Thorpe, 2001). Gordon (2004), for example, found effects of semantic inconsistencies on covert attention during the first fixation on a scene, using spatial probes. The results suggest that within approximately 150 ms of scene onset, participants attended preferentially to inconsistent objects (see also Gordon, 2006). Joubert et al. found that performance in a scene categorization task was impeded when the scene—flashed for only 26 ms—contained salient inconsistent objects. Furthermore, VanRullen and Thorpe investigated the time course of processing objects embedded in shortly flashed scenes, using event-related potentials (ERPs). Participants were asked to decide whether an animal or a vehicle was present in a 20-ms flash of a scene. They found that waveforms started to diverge between objects that either did or did not match the target category as early as 150 ms after scene onset. Altogether, these studies suggest that objects within scenes but outside the fovea can be semantically processed very early in scene viewing and might, therefore, be able to influence the deployment of attention within the first fixation on a scene.

One might conclude that if attention is preferentially drawn to inconsistent objects, gaze should be drawn there as well. However, experimental evidence concerning the effects of extrafoveal semantic processing on eye movement control has been contradictory. Some studies have suggested that semantic inconsistencies in the visual periphery can be detected early enough to attract eye movements (e.g., Becker, Pashler, & Lubin, 2007; Bonitz & Gordon, 2008; Loftus & Mackworth, 1978; Underwood & Foulsham, 2006; Underwood, Humphreys, & Cross, 2007; Underwood, Templeman, Lamming, & Foulsham, 2008), whereas other studies have shown no evidence that semantically inconsistent objects attract gaze prior to their fixation (e.g., De Graef, Christiaens, & d'Ydewalle, 1990; Gareze & Findlay, 2007; Henderson, Weeks, & Hollingworth, 1999; Võ & Henderson, 2009). One possible reason for the contradictory findings might be differences in the control and quality of the scene material used, which have included line drawings (e.g., De Graef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978), edited photographs (e.g., Rayner, Castelhano, & Yang, 2009; Underwood et al., 2007, 2008), and rendered images of naturalistic scenes (Võ & Henderson, 2009). For example, in their classic "octopus in farmyard" study, Loftus and Mackworth found not only earlier fixation of inconsistent objects, but also longer saccades (of about 6.5°–8° of visual angle) entering scene regions containing an inconsistent object. Since the authors had defined processing of information appearing within 2°–3° of visual angle as near-foveal vision, it was concluded that eye movements can be attracted toward semantic inconsistencies in the visual periphery. However, this study was subsequently criticized for the possible lack of control of visual saliency of the consistent and inconsistent target objects (e.g., Henderson & Hollingworth, 1999a, 1999b; Henderson et al., 1999; Underwood & Foulsham, 2006; Võ & Henderson, 2009).

In a more recent study measuring eye movements during free viewing of naturalistic scenes, Bonitz and Gordon (2008) found that semantically inconsistent objects were fixated earlier than their consistent counterparts. However, although inconsistent objects were fixated earlier, this effect occurred only after several eye movements. In this case, gaze is not initially captured by inconsistent objects in the visual periphery but might still be directed to inconsistent objects slightly earlier than to consistent objects in the course of scene viewing. As the eyes move through a scene, it becomes more likely that effects of object–scene consistency arise from fixations that land near the target objects. Thus, earlier fixations of inconsistent objects might be generated during fixations proximate to inconsistent objects. In line with this hypothesis, Rayner and colleagues (2009) found slightly earlier fixations on “weird” than on normal scene regions, but there was no evidence that object–scene inconsistencies attracted gaze from farther away. The same might be true for other studies that have reported results consistent with attraction of both attention and gaze toward object–scene inconsistencies (e.g., Underwood et al., 2007, 2008).

The present study was therefore designed to disentangle extrafoveal semantic processing across the entire scene, on the one hand, from foveal semantic processing of a small scene region near fixation, on the other. Following Loftus and Mackworth’s (1978) definition, we regard information within 2°–3° of visual angle of a current fixation as foveal or near-foveal and effects due to information outside of this range as extrafoveal.

In the majority of studies investigating the influence of object–scene processing on covert and overt attention, the scene has remained visible throughout ongoing inspection, making it difficult to differentiate between the effects of initial global and subsequent local semantic processing. In the present study, we sought to determine whether gaze is captured by object–scene relationships processed during the initial glimpse of the scene. At the same time, we wanted to

limit the influence of extrafoveal processing after the initial glimpse, while allowing participants to freely move their eyes during subsequent scene exploration. To accommodate both of these goals, we used the flash-preview moving-window paradigm introduced by Castelhano and Henderson (2007). This paradigm combines the brief tachistoscopic viewing method typically used in scene categorization experiments with the moving window technique typically used to investigate eye movements under restricted-viewing conditions.

In prior research, this method has been successfully applied to study the influence of the initial glimpse of a scene on subsequent eye movement control during search (Castelhano & Henderson, 2007; Võ & Schneider, 2010). In this paradigm, participants are first presented with a short preview of the search scene, followed by the presentation of a target word indicating which object they will be looking for. The scene is then presented again for search, but participants are able to view the scene only through a small gaze-contingent window. With this paradigm, object–scene relationships can be processed only during the initial glimpse or when the moving window includes the critical object because fixation has landed close to it. Note that because participants are able to see the entire scene only during the initial preview, they should also be motivated to process as much of the scene as possible in this initial glimpse, increasing the likelihood of extrafoveal object–scene processing, if it is possible.

So, if gaze is directed earlier toward inconsistent than toward consistent objects in the periphery, this effect would have to be due to extrafoveal semantic processing from the initial glimpse. On the other hand, if gaze is not directed to inconsistent objects any earlier than to consistent objects, this would imply that the degree of semantic object processing during an initial glimpse of a scene is insufficient to draw both covert and overt attention to inconsistent objects in the visual periphery.

To date, most of the evidence bearing on the effect of object–scene inconsistencies has come from one type of manipulation: the semantic violation of a scene's gist. However, a different way to produce object–scene inconsistencies relates to an object's structural relationship with other scene elements, or *scene syntax* (e.g., Biederman, 1981). We therefore manipulated both semantic and syntactic inconsistencies of objects in scene contexts. Semantic violations of the scene context were created by replacing a semantically plausible object within a scene (e.g., a pot in a kitchen) with an implausible object (e.g., a printer in the kitchen). For this manipulation, consistent objects were swapped across scenes to create the inconsistent conditions. We operationalized syntactic inconsistencies by violating the local scene structure—that is, by having objects that normally rest on surfaces float above those surfaces (referred to by Biederman, Mezzanote, & Rabinowitz, 1982, as a "support violation"). Thus, syntactically inconsistent objects violated the scene structure without affecting expectations about where an object can usually be found within a scene. In the first three experiments, we used the flash-preview moving-window paradigm to investigate whether semantic, syntactic, or combined semantic and syntactic object–scene inconsistencies presented in the visual periphery would affect eye movement control during scene search, while the fourth experiment investigated whether semantic object–scene inconsistencies would attract gaze in a nonsearch task.

General method

Stimulus material

The stimulus material used in all three experiments consisted of 20 images of real-world scenes rendered from 3-D models. The scenes were displayed on a 21-in. computer screen (resolution, 1,024 × 768 pixel; 140 Hz) subtending visual angles of 25.66° (horizontal) and 19.23° (vertical) at a viewing distance of 90 cm. Each scene was manipulated so that it conformed to one of the four experimental conditions. In the consistent–surface condition (the control condition), the object of interest was semantically consistent with the scene context and rested on a surface (e.g., a pot on a kitchen stove), whereas in the consistent–float condition, the same object was displayed hovering in midair above the surface. In the inconsistent–surface and inconsistent–float conditions, the semantically consistent object was replaced by a semantically inconsistent object (e.g., a printer on a kitchen stove) resting on a surface or hovering in midair, respectively. Target objects were presented at an average eccentricity of 7.13° (*SD* = 1.35°) of visual angle from the center of the screen and, on average, subtended 4.20° (*SD* = 1.84°) of visual angle in width and 4.16° (*SD* = 1.76°) of visual angle in height. On average, each scene contained 8.45 (*SD* = 1.23) objects, 2.90 (*SD* = 1.62) of which were closer to the fixation point than was the target object. Figure 1 displays a sample scene in its four versions. Note that all previews contained the search target and were identical to the subsequently presented search scene.

Scenes were paired so that each semantically consistent object in its scene was inconsistent in its paired scene (e.g., a pot on a stove and a printer on an office desk were swapped so that the pot appeared on the desk and the printer on the stove). Semantically consistent and inconsistent objects were matched for size and were placed in the same position within each scene and away from the initial fixation location at the center. The specific semantic and

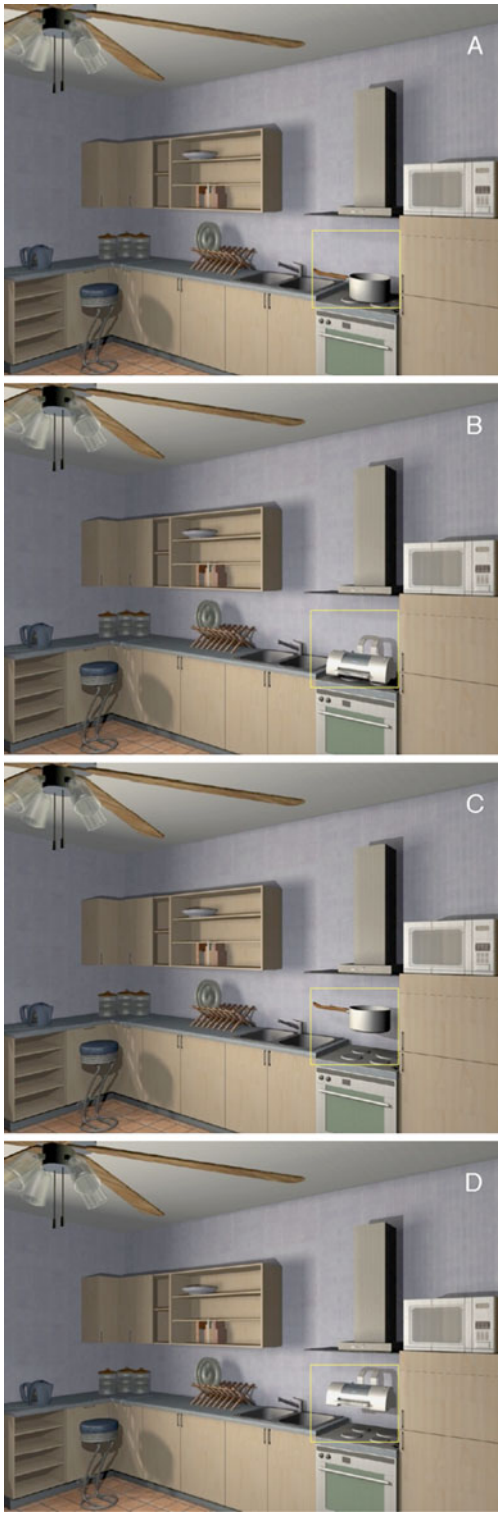


Fig. 1 Sample of four versions of a kitchen scene containing **a** a semantically consistent nonfloating object, **b** a semantically inconsistent nonfloating object, **c** a semantically consistent floating object, or **d** a semantically inconsistent floating object. Yellow rectangles indicate scoring regions and were not shown to participants

syntactic manipulations of objects within scenes used here have previously been shown to strongly affect eye movement measures, including gaze durations and number of fixations once fixated (Võ & Henderson, 2009). Furthermore, the scenes were analyzed using the Itti and Koch (2000) MATLAB Saliency Toolbox to determine the most salient regions according to low-level saliency calculations of brightness, color, contrast, and edge orientation. The rank order of saliency peaks, with rank 1 assigned to the most salient region of the scene, was used to control mean low-level saliency across consistent and inconsistent objects (M = 8.45, SD = 3.09 vs. M = 8.9, SD = 2.38, respectively; $p > .05$).

Furthermore, the detection rate for semantic and syntactic object–scene inconsistencies was tested on 32 participants who took part in two separate control experiments. Participants were instructed to fixate the center of the screen, while a 1,000-ms presentation of a location cue indicated the position of the critical object. Scenes were flashed for 250 ms and subsequently masked, while participants held their gaze on the center of the screen (controlled by recording their eye movements). Participants were asked to judge whether the indicated object had been consistent or inconsistent with the rest of the scene by pressing either the left or the right button of a joystick. Results showed that both semantic and syntactic inconsistencies were detected above chance [semantic detection, 70%, $t(15) = 10.97$, $p < .01$; syntactic detection, 64%, $t(15) = 7.67$, $p < .01$].

The consistent–surface condition served as a baseline for all four experiments, against which we contrasted semantically inconsistent objects resting on surfaces (Experiments 1 and 4), semantically consistent but floating objects (Experiment 2), and both semantically and syntactically inconsistent objects (Experiment 3).

Apparatus

Eye movements were recorded with an EyeLink1000 tower system (SR Research, Canada), which tracks eye position with a resolution of 0.01° of visual angle at a sampling rate of 1,000 Hz. The position of the right eye was tracked, although viewing was binocular. Experimental sessions were carried out on a computer running Windows XP. Stimulus presentation and response recording were controlled by Experiment Builder (SR Research).

Procedure

The procedure of the first three experiments closely followed the procedure of the flash-preview moving-window paradigm reported in Castelhano and Henderson (2007). The participants first received written instructions. They were informed that they would be shown a series of scenes in which they had to find predefined target objects as quickly as possible. They were also informed that short

previews of the scene would precede the display of the search scene and that they should attend to these previews since they could provide additional helpful information.

At the beginning of the experiment, the eyetracker was calibrated using a 9-point calibration and validation procedure. The participants' viewing position was fixed with a chin and forehead rest. As can be seen in Fig. 2, each trial sequence was preceded by a fixation check: To initiate the next trial, the participants had to fixate a cross centered on the screen for 200 ms. When the fixation check was deemed successful, the fixation cross was replaced by the scene preview for 250 ms, followed by a mask for 50 ms. A word indicating the identity of the target object was then displayed at the center of a gray field for 1,500 ms, followed by a fixation cross for 500 ms. The search scene was then shown through a 5°-diameter circular window that moved contingent on the participants' fixation location. The scene was visible within the window and was replaced with a gray field outside of the window. Thus, no peripheral vision was possible throughout the entire period of active visual search. Participants had to search the scene for the target object and indicate its detection by holding fixation on the object and pressing a response button. The search scene was displayed for 15 s or until buttonpress. Experiment 4 differed from the previous experiments only in that participants were not asked to search for objects after the flashed preview but to simply inspect the scene using the gaze-contingent window for a total of 15 s (see Experiment 4 for more details).

Two practice trials at the beginning of the experiment allowed participants to become accustomed to the task and the gaze-contingent window. The experiment lasted about 15 min.

Eye movement data analysis

Following Võ and Henderson (2009), the interest area for each target object was defined as the rectangular box that was large enough to encompass the consistent and inconsistent target objects when located on a surface, as well as when floating (see Fig. 1). Thus, the scoring regions were the same for all conditions across experiments to allow for better comparison. Incorrect responses and responses with a latency deviating more than two standard deviations from the within-participants mean were defined as outliers and were excluded from further analyses (11%–13% across all experiments, equally distributed across consistent and inconsistent conditions). Fixation durations of less than 90 ms and more than 1,000 ms were also excluded as outliers. The remaining raw data were subsequently filtered using SR Research Data Viewer.

Experiment 1

Experiment 1 investigated whether peripherally determined object–scene inconsistencies generated from an initial scene glimpse can attract gaze. If so, gaze should be more likely to move to the inconsistent than to the consistent objects. If, on the other hand, semantic inconsistencies are typically processed only when the eyes land on or very near to the object, no bias for earlier fixation of inconsistent objects should be observed.

Participants

Sixteen native English-speaking students (11 female) from the University of Edinburgh ranging in age from 18 to

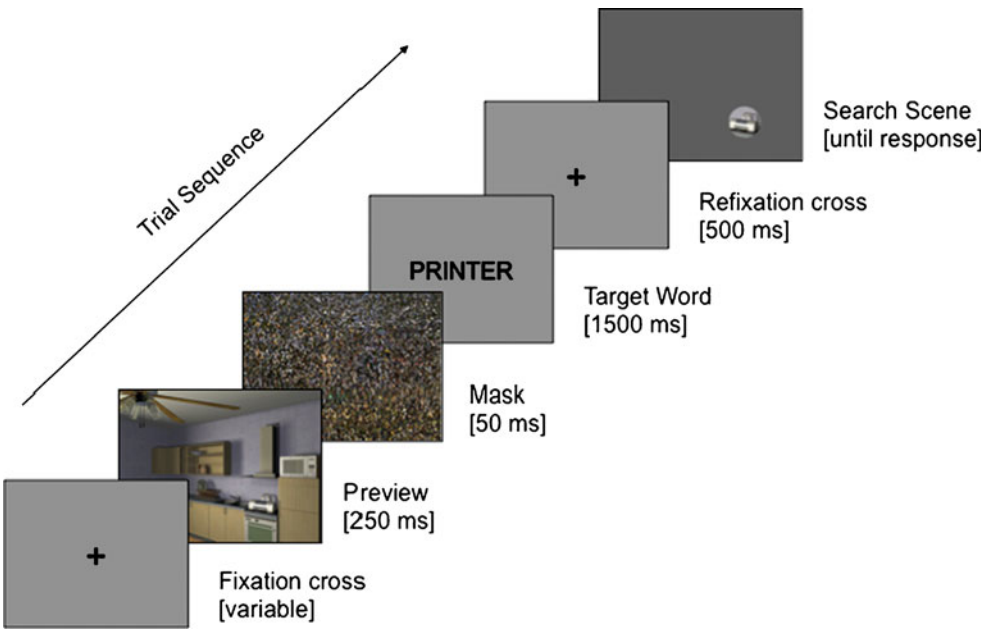


Fig. 2 Trial sequence of the flash-preview moving-window paradigm in Experiments 1–3

31 years (M = 21.8, SD = 3.15) participated in Experiment 1 for course credit or for £6/h. All participants reported normal or corrected-to-normal vision and were unfamiliar with the stimulus material. One participant had to be replaced due to unstable recording of eye position.

Results and discussion

The aim of the present study was to investigate whether initial eye movements during scene viewing can be modulated by the processing of peripheral semantic object–scene inconsistencies. To investigate whether these inconsistencies affect eye movements prior to their fixation, a set of measures was calculated (see Table 1): initial saccade latency, initialsaccade amplitude, entering saccade amplitude, latency to first target fixation, number of fixations to first target fixation, and response times (RTs).

Initial saccade latency Initial saccade latency was measured from scene onset until the initiation of the first saccade and averaged 286 ms across conditions. There was no effect of semantic inconsistency, $t(15) < 1$.

Initial saccade amplitude Initial saccade amplitude was the length of the first saccade after search scene onset and averaged 2.93° of visual angle across conditions. There was no effect of semantic inconsistency, $t(15) = 1.21$, $p > .05$.

Entering saccade amplitude Entering saccade amplitude was the length of the saccade entering the target interest area and averaged 3.73° of visual angle across conditions. There was no effect of semantic inconsistency, $t(15) < 1$.

Latency to first target fixation Latency was measured from search scene onset until the first fixation of the target object and averaged 1,489 ms across conditions. There was an effect of semantic inconsistency, $t(15) = 2.09$, $p < .05$, with increased latency for semantically inconsistent objects, contrary to the prediction that inconsistent objects attract gaze.

Number of fixations to first target fixation This measure was defined as the number of discrete fixations until the target object was first fixated. The measure included the initial scene fixation centered on the screen, but not the first fixation on the target object. On average, participants performed 5.26 fixations to the first fixation of the target object. There was again an effect of semantic inconsistency, with an increased number of fixations for semantically inconsistent objects, $t(15) = 2.11$, $p < .05$, contrary to the prediction that inconsistent objects attract gaze.

Response time RT was defined as the time elapsed from scene onset until buttonpress and averaged 2,726 ms across conditions. Reflecting the latency to fixation and number of fixations measures, RTs were longer to semantically inconsistent objects than to their consistent counterparts, $t(15) = 3.01$, $p < .01$.

Summary The data from Experiment 1 provided no evidence for gaze capture by semantically inconsistent objects in naturalistic scenes. Neither initial saccade latency nor initial saccade amplitude was affected by the semantic manipulation of the target objects. In fact, there was no evidence in any measure that semantically inconsistent objects attracted earlier fixations than did consistent objects. Instead, semantically inconsistent target objects were fixated later than their consistent counterparts. For semantically consistent objects, the short scene preview seems to have provided helpful information about the scene's spatial layout and possible target locations, facilitating search (Castelhano & Henderson, 2007; Malcolm & Henderson, 2010; Torralba, Oliva, Castelhano, & Henderson, 2006; Võ & Henderson, 2010), while semantically inconsistent target objects lacked such contextual guidance.

Experiment 2

The results of Experiment 1 suggest that semantic object–scene inconsistencies in the visual periphery do not attract

Table 1 Summary of mean values (with standard errors) in Experiment 1 regarding dependent variables as a function of semantics (consistent vs. inconsistent), including initial saccade latency, initial saccade amplitude, entering saccade amplitude, latency to first target fixation, number of fixations to target fixation, and response times

Measures	Semantic		*df*	*t*	*p*
	Consistent	Inconsistent			
Initial saccade latency (ms)	288 [13]	285 [13]	15	0.20	.85
Initial saccade amplitude (deg of visual angle)	2.80 [0.31]	3.06 [0.42]	15	1.21	.25
Entering saccade amplitude (deg of visual angle)	3.66 [0.37]	3.80 [0.38]	15	0.22	.82
Latency to first target fixation (ms)	1,255 [205]	1,723 [145]	15	2.10	.03
Number of fixations until target fixation	4.41 [0.71]	6.1 [0.42]	15	2.10	.03
Response time (ms)	2,390 [277]	3,061 [202]	15	3.01	<.01

eye movements even when the usefulness of initial extrafoveal processing is increased. Experiment 2 investigated whether syntactic, rather than semantic, inconsistencies would attract gaze. The logic of the study was the same as that in Experiment 1. If syntactic inconsistencies can be sufficiently processed within the first glimpse of a scene, eye movements should move to those syntactic inconsistencies more quickly than they move to consistent controls. If, on the other hand, syntactic inconsistencies are processed only after fixation lands near or on the object, no syntactic consistency bias should be observed. Note also that in this experiment, we did not expect a benefit of contextual guidance for syntactically consistent over inconsistent objects, because objects in both conditions were in a generally appropriate location in their scenes. Therefore, if the failure to observe an attraction by inconsistent objects in Experiment 1 was due to a countervailing influence of contextual guidance for consistent objects, inconsistent object attraction should be observed in this experiment where the degree of contextual guidance was similar between syntactically consistent and inconsistent objects.

Participants

Sixteen native English-speaking students (9 female) from the University of Edinburgh ranging in age from 19 to 34 years (M = 22.0, SD = 3.67) participated in Experiment 2 for course credit or for £6/h. All participants reported normal or corrected-to-normal vision, and none had taken part in Experiment 1.

Results and discussion

The same measures as those used in Experiment 1 were calculated. A summary of the mean values of these measures can be seen in Table 2.

Initial saccade latency Initial saccade latency averaged 288 ms across conditions. There was no effect of syntactic inconsistency, $t(15) < 1$.

Initial saccade amplitude Initial saccade amplitude averaged 3.69° of visual angle across conditions. There was no effect of syntactic inconsistency, $t(15) < 1$.

Entering saccade amplitude Entering saccade amplitude averaged 3.43° of visual angle across conditions. There was no effect of syntactic inconsistency, $t(15) = 1.63$, $p > .05$.

Latency to first target fixation Latency averaged 1,361 ms across all conditions. There was no effect of syntactic inconsistency, $t(15) < 1$.

Number of fixations to first target fixation On average, participants performed 4.87 fixations to the first fixation of the target object. There was no effect of syntactic inconsistency, $t(15) = 1.19$, $p > .05$.

Response time RT averaged 2,920 ms across conditions. There was no effect of syntactic inconsistency on RT, $t(15) < 1$.

Summary Initial eye movements were not preferentially directed to a syntactically inconsistent target object in a briefly flashed scene preview. Furthermore, the syntactic manipulation in Experiment 2 did not affect overall search performance; that is, syntactically inconsistent objects were not more difficult to find than their consistent counterparts, presumably because syntactically inconsistent objects were placed in expected scene locations despite the fact that they were floating. Therefore, the failure to observe gaze attraction for inconsistent objects cannot be due to countervailing contextual guidance for consistent objects.

Experiment 3

The results of Experiments 1 and 2 suggest that neither semantic nor syntactic object–scene inconsistencies attract gaze from an initial scene glimpse. A possible criticism of these experiments is that the inconsistency simply was not powerful enough to generate gaze capture. This explanation

Table 2 Summary of mean values (with standard errors) in Experiment 2 regarding dependent variables as a function of syntax (surface vs. float), including initial saccade latency, initial saccade amplitude, entering saccade amplitude, latency to first target fixation, number of fixations to target fixation, and response times

Measures	Syntax		*df*	*t*	*p*
	Surface	Float			
Initial saccade latency (ms)	288 [17]	287 [18]	15	0.11	.91
Initial saccade amplitude (deg of visual angle)	3.82 [0.28]	3.55 [0.29]	15	0.98	.34
Entering saccade amplitude (deg of visual angle)	3.68 [0.37]	3.19 [0.20]	15	1.62	.12
Latency to first target fixation (ms)	1,295 [164]	1,427 [101]	15	0.90	.38
Number of fixations untiltarget fixation	4.60 [0.52]	5.14 [0.42]	15	1.19	.25
Response time (ms)	2,927 [265]	2,913 [246]	15	0.05	.96

seems unlikely given that we have observed powerful effects of these manipulations in these scenes once the target objects have been fixated (Võ & Henderson, 2009). Nevertheless, Experiment 3 provided a more direct test of the hypothesis that a stronger manipulation would produce an inconsistency attraction effect by combining the two inconsistencies into a double inconsistency. The predictions were the same as those in Experiments 1 and 2.

Participants

Sixteen native English-speaking students (10 female) from the University of Edinburgh ranging in age from 19 to 28 years (M = 22.9, SD = 2.36) participated in Experiment 3 for course credit or for £6/h. All participants reported normal or corrected-to-normal vision, and none had taken part in Experiment 1 or 2. Two participants had to be replaced due to difficulties in tracking their eyes.

Results and discussion

A summary of mean values of the eye movement measures can be seen in Table 3.

Initial saccade latency Initial saccade latency averaged 304 ms across conditions. There was no effect of semantic/syntactic inconsistency, $t(15) < 1$.

Initial saccade amplitude Initial saccade amplitude averaged 3.50° of visual angle across conditions. There was no effect of semantic/syntactic inconsistency, $t(15) = 1.17$, $p > .05$.

Entering saccade amplitude Entering saccade amplitude was the length of the saccade entering the target interest area and averaged 3.47° of visual angle across conditions. There was no effect of semantic inconsistency, $t(15) = 1.16$, $p > .05$.

Latency to first target fixation Latency averaged 1,590 ms across conditions. As in Experiment 1, there was an effect of inconsistency, $t(15) = 1.92$, $p < .05$, in that the latency to the first fixation on the target was increased for the semantic/syntactic inconsistent objects.

Number of fixations to first target fixation On average, there were 5.53 fixations to the first fixation of the target object. There was again an effect of inconsistency, $t(15) = 1.94$, $p < .05$, with an increased number of fixations for the semantic/syntactic inconsistent objects.

Response time RT averaged 3,150 ms across conditions. RTs to the semantic/syntactic inconsistent objects were longer than those to their consistent counterparts, $t(15) = 2.84$, $p < .01$.

Summary Like Experiments 1 and 2, Experiment 3 provided no evidence of gaze capture by inconsistent objects in naturalistic scenes, despite the combination of semantic and syntactic inconsistencies. Similar to the semantic manipulation in Experiment 1, objects that were inconsistent in both semantics and syntax were fixated later than their consistent controls, as seen in the prolonged search times and increased number of fixations.

Even though Experiments 1–3 showed no signs of gaze attraction toward object–scene inconsistencies in the visual periphery, a remaining concern centered on the search task itself. The strong contextual guidance that controls eye movements during object search might have overshadowed possibly smaller effects of semantic inconsistency. However, it was not at all clear how contextual guidance would offset gaze attraction in the floating object condition (Experiment 2), since guidance would direct the eyes to the same (correct) location in both conditions. Nevertheless, the last experiment directly addressed this issue by employing a memorization rather than a search task to test whether semantic object–scene inconsistencies generated from an initial scene glimpse

Table 3 Summary of mean values (with standard errors) in Experiment 3 regarding dependent variables as a function of semantics–syntax (consistent–surface vs. inconsistent–float), including initial saccade latency, initial saccade amplitude, entering saccade amplitude, latency to first target fixation, number of fixations to target fixation, and response times

Measures	Semantic/Syntax		*df*	*t*	*p*
	Consistent–Surface	Inconsistent–Float			
Initial saccade latency (ms)	300 [14]	307 [15]	15	0.73	.48
Initial saccade amplitude (deg of visual angle)	3.39 [0.28]	3.60 [0.35]	15	1.17	.26
Entering saccade amplitude (deg of visual angle)	3.35 [0.23]	3.58 [0.21]	15	1.16	.26
Latency to first target fixation (ms)	1,336 [178]	1,843 [182]	15	1.94	.04
Number of fixations until target fixation	4.66 [0.65]	6.40 [0.63]	15	1.94	.04
Response time (ms)	2,701 [251]	3,598 [211]	15	2.84	<.01

might attract gaze in situations that were less constrained by scene context.

In addition, another factor that could have weakened a potential effect of object–scene inconsistency might be the long lag between the scene preview and the commencement of search (2,050 ms in Experiments 1–3). The employment of a memorization task in Experiment 4 also addressed this issue by reducing the lag to only 550 ms, since replacing the search task made the presentation of a target word superfluous.

Experiment 4

This last experiment was designed to motivate participants to process as much information from a flashed preview without the need to subsequently search, since contextual guidance might have counteracted possible attractions of gaze. Furthermore, we increased the number of participants to 25.

Method

Participants Twenty-five native English-speaking students (15 female) from the University of Edinburgh ranging in age from 18 to 26 years ($M = 23$, $SD = 4.15$) participated in Experiment 4 for course credit or for £6/h. All participants reported normal or corrected-to-normal vision and were unfamiliar with the stimulus material.

Procedure In Experiment 4, participants viewed the same scenes that had been presented in Experiment 1 containing either semantically consistent or inconsistent objects. The procedure differed from Experiment 1 only in that the search task was replaced by a memorization task (see Fig. 3). That is, participants were told that a short flash of a scene would appear, followed by 15 s of restricted viewing of the scene, during which they should memorize as much of the scene as possible for later memory test questions. They were further instructed that since the flash would be the only time the whole scene would be visible, they should try to gather as much information from the scene preview as possible.

The trial sequence of Experiment 4 consisted of a 250-ms scene preview, followed by a mask for 50 ms. Note that since there was no object to search for, no target word appeared and the mask was directly followed by a fixation cross for 500 ms. Thus, the lag between preview and scene viewing in Experiment 4 was 550 ms and, therefore, considerably shorter than the 2,050-ms lag in Experiments 1–3. The scene was then shown for a total of 15 s through a 5°-diameter circular window that moved contingent on the participant's fixation location. A memorization task was not administered.

Results and discussion

Since no responses were made during this 15-s viewing period and the trial could not be terminated earlier, we were able to investigate the influence of semantic inconsistency not only before, but also after the fixation of the critical object. Thus, we added the total gaze duration upon fixation to the set of measures. A summary of the mean values of these measures can be seen in Table 4.

Initial saccade latency Initial saccade latency averaged 284 ms across conditions. There was no effect of semantic inconsistency, $t(15) < 1$.

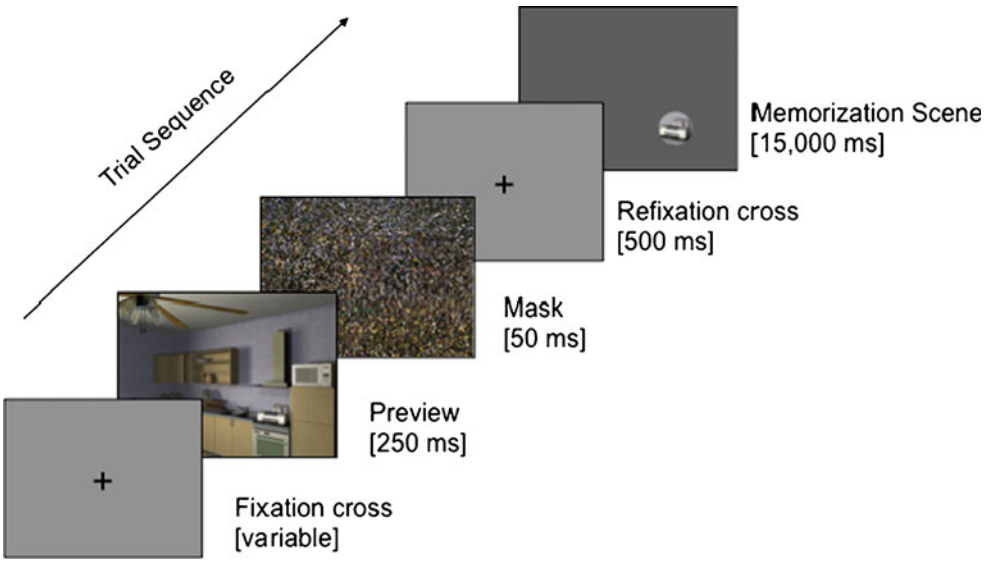


Fig. 3 Trial sequence of the flash-preview moving-window paradigm in Experiment 4

Table 4 Summary of mean values (with standard errors) in Experiment 4 regarding dependent variables as a function of semantics (consistent vs. inconsistent), including initial saccade latency, initial saccade amplitude, entering saccade amplitude, latency to first target fixation, number of fixations to target fixation, and total gaze duration

Measures	Semantic		*df*	*t*	*p*
	Consistent	Inconsistent			
Initial saccade latency (ms)	288 [12]	280 [14]	24	0.65	.53
Initial saccade amplitude (deg of visual angle)	2.33 [0.15]	2.12 [0.11]	24	1.32	.20
Entering saccade amplitude (deg of visual angle)	2.78 [0.16]	2.70 [0.16]	24	0.44	.66
Latency to first target fixation (ms)	4,639 [265]	4,290 [315]	24	0.92	.37
Number of fixations until target fixation	13.54 [0.73]	12.66 [0.84]	24	0.85	.40
Total gaze duration (ms)	1,014 [67]	1,209 [68]	24	2.08	.02

Initial saccade amplitude Initial saccade amplitude averaged 2.23° of visual angle across conditions. There was no effect of semantic inconsistency, $t(15) = 1.33$, $p = .20$.

Entering saccade amplitude Entering saccade amplitude was the length of the saccade entering the target interest area and averaged 2.74° of visual angle across conditions. There was no effect of semantic inconsistency, $t(15) < 1$.

Latency to first target fixation Latency was measured from search scene onset until the first fixation of the target object and averaged 4,465 ms across conditions. There was no effect of semantic inconsistency, $t(15) < 1$.

Number of fixations to first target fixation On average, participants performed 13.10 fixations to the first fixation of the target object. There was no effect of semantic inconsistency, $t(15) < 1$.

Total gaze duration Total gaze duration summarizes the time that a critical object was fixated in the course of scene viewing. The average time spent on the critical objects averaged 1,111 ms. Inconsistent objects were gazed at longer than consistent objects upon their fixation, $t(15) = 2.08$, $p < .05$.

Summary Despite employing a task that was less driven by contextual guidance than visual search, despite reducing the lag between scene preview and commencement of scene exploration from 2,050 to 550 ms, and despite increasing the number of participants, Experiment 4 nevertheless provided no evidence for gaze capture by semantically inconsistent objects in naturalistic scenes. However, upon fixation, semantic inconsistency led to prolonged gaze. Thus, we found no evidence that semantic inconsistencies in the visual periphery are initially processed to a degree that attracts gaze. In addition, we replicated previous findings that inconsistent objects hold gaze once fixated (e.g., De Graef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009).

General discussion

Do higher-level object–scene inconsistencies attract gaze within the first glimpse of a scene? That is, can semantically or syntactically inconsistent objects that appear in the visual periphery of a naturalistic scene be processed to a degree that modulates subsequent eye movement control? Previous studies have shown that, within a glance, inconsistent objects in scenes preferentially attract covert attention (e.g., Gordon, 2004, 2006; Joubert et al., 2007). These findings suggest that semantic object information might be processed in the visual periphery. However, experimental evidence concerning their effects on eye movement control has been mixed, with some studies suggesting that the eyes are attracted by inconsistent objects (e.g., Becker et al., 2007; Bonitz & Gordon, 2008; Loftus & Mackworth, 1978; Underwood et al., 2007; Underwood et al., 2008) and others suggesting that they are not (e.g., De Graef et al., 1990; Gareze & Findlay, 2007; Henderson et al., 1999; Võ & Henderson, 2009). Importantly, none of the studies conducted up until now have investigated the effects of initial scene processing isolated from subsequent ongoing visual processing. Thus, it has been unclear whether the effects sometimes found in prior studies resulted from initial scene processing or, rather, arose during later stages of scene processing when fixation happened to fall near a target object.

To investigate this issue, we used the flash-preview moving-window paradigm (Castelhano & Henderson, 2007). By presenting a scene preview for too brief a time to allow for eye movements and by limiting subsequent visual input to a gaze-contingent window, this paradigm provides a method for isolating the effects of initial scene processing from processing that takes place during later stages of scene viewing. Additionally, the flash-preview moving-window paradigm motivates participants to process as much information from the preview as possible. If an analysis of objects–scene consistency in an initial view can draw the eyes, the effect should be observed in this paradigm. Instead, Experiments 1–3 showed that the eyes either were equally likely to move to consistent and

inconsistent objects or were more likely to move to consistent objects first, since search for consistent objects can be guided by scene context (for a review, see Wolfe, Võ, Evans, & Greene, 2011). Even in a memorization task that does not draw as much on contextual guidance as a search task does, we found no attraction of gaze toward inconsistent objects (Experiment 4). Only upon fixation did the eyes linger on the inconsistent objects longer than on their consistent controls. Note that using the same stimulus material in a previous study, we also found strong effects of both semantic and syntactic manipulations on eye movement control upon fixation of the inconsistent object but no attraction of gaze before the eyes landed on the object (Võ & Henderson, 2009). Thus, the failure to find attraction of gaze toward inconsistent objects in this study cannot be attributed to a lack of strength of our inconsistency manipulations.

Therefore, our findings do not support the hypothesis that object–scene semantics are processed across the visual field during initial scene analysis (e.g., Gordon, 2004, 2006; Joubert et al., 2007). If at all, signals stemming from such a computation are not strong enough to modulate eye movement control. Rather, our results add to the growing number of studies that argue against the ability to process object–scene inconsistencies in the visual periphery (e.g., De Graef et al., 1990; Gareze & Findlay, 2007; Henderson et al., 1999; Võ & Henderson, 2009).

Why has evidence supporting attraction of gaze by inconsistent objects from the visual periphery sometimes been reported? We suggest that previously reported findings of earlier fixations on inconsistent objects have been due to later local, rather than initial global, scene processing. Accordingly, Underwood and Foulsham (2006) reported that the picture had been scanned for 2 or 3 s by the time an incongruent object was fixated, while Bonitz and Gordon (2008) found that odd objects were not fixated until about the sixth fixation in the scene. Another related reason for the diverging results might lie in differences regarding mean saliency rank values of the critical objects across studies. For example, Underwood et al. (2007) used objects that had a mean saliency rank value of about 3, as compared with the rest of the scene, while our objects ranked much lower in visual saliency (mean rank about 8.5). In our case, a mean rank value of about 8 implies that seven other regions in the scene were visually more conspicuous, while in the study by Underwood et al. (2007), on average, only two other regions were more conspicuous. Thus, at least during free scene viewing, the effect of scene inconsistencies might also depend on the relative visual salience of the inconsistent object with regard to the rest of the scene. While we took great care to control for relative saliency ranks, future studies might want to address this issue by directly manipulating the relative saliency ranks of inconsistent objects within a scene.

Conclusions

In this study, we tested whether object–scene inconsistencies in the visual periphery can be processed to such a degree that eye movements during scene viewing are modulated before their fixation. The flash-preview moving-window paradigm enabled us to disentangle the effect of object–scene inconsistencies during initial scene processing and the local processing of such inconsistencies during later stages of scene viewing. By limiting the visual input during search to a gaze-contingent window, we not only maximized the potential influence of initial scene processing, but also minimized the influence of inconsistency processing from proximate target fixations during later stages of scene viewing. In four experiments we clearly showed that neither semantic nor syntactic object–scene inconsistencies attract gaze from an initial glimpse of a scene.

Acknowledgements This project was supported by Grant RES-00-22-2721 from the Economic and Social Research Council of the U.K. to J.M. H. and Grant DFG: VO 1683/1-1 awarded to M.V.

References

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 20–30.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–263). Hillsdale: Erlbaum.

Biederman, I., Mezzanote, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica, 129,* 255–263.

Castelhano, M., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 753–763.

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the activation of scene gist. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 660–675.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559–564.

De Graef, P., Christiaens, D., & D'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52,* 317–329.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1, Art. 10), 1–29.

Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 618–637). Amsterdam: Elsevier.

Gordon, R. D. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 760–777.

Gordon, R. D. (2006). Selective attention during scene perception: Evidence from negative priming. *Memory & Cognition, 34,* 1484–1494.

Henderson, J. M., & Hollingworth, A. (1999a). High-level scene perception. *Annual Review of Psychology, 50,* 243–271.

Henderson, J. M., & Hollingworth, A. (1999b). The role of fixation position in detecting scene changes across saccades. *Psychological Science, 5,* 438–443.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127,* 398–415.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47,* 3286–3297.

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.

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real world scene search. *Journal of Vision, 10*(2, Art. 4), 1–11.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors meduate scene recognition. *Cognitive Psychology, 41,* 176–210.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155,* 23–36.

Potter, M. C. (1975). Meaning in visual scenes. *Science, 187,* 965–966.

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Viewing task influences eye movements during active scene perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 254–259.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59,* 1931–1949.

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency, and gist in the inspection of objects in natural scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 564–579). Amsterdam: Elsevier.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17,* 159–170.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13*(4), 454–461.

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3, Art. 24), 1–15.

Võ, M. L.-H., & Henderson, J. M. (2010). The time course of initial scene processing for guidance of eye movements when searching natural scenes. *Journal of Vision, 10*(3, Art. 14), 1–13.

Võ, M. L.-H., & Schneider, W. X. (2010). A glimpse is not a glimpse: Differential processing of flashed scene previews leads to differential target search benefits. *Visual Cognition, 18,* 171–200.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Sciences, 15,* 77–84.