Low-level visual saliency does not predict change detection in natural scenes

Jonathan A. Stirk
School of Psychology, University of Nottingham, Nottingham, UK

Geoffrey Underwood
School of Psychology, University of Nottingham, Nottingham, UK

Saliency models of eye guidance during scene perception suggest that attention is drawn to visually conspicuous areas having high visual salience. Despite such low-level visual processes controlling the allocation of attention, higher level information gained from scene knowledge may also control eye movements. This is supported by the findings of eye-tracking studies demonstrating that scene-inconsistent objects are often fixated earlier than their consistent counterparts. Using a change blindness paradigm, changes were made to objects that were either consistent or inconsistent with the scene and that had been measured as having high or low visual salience (according to objective measurements). Results showed that change detection speed and accuracy for objects with high visual salience did not differ from those having low visual salience. However, changes in scene-inconsistent objects were detected faster and with higher accuracy than those in scene-consistent objects for both high and low visually salient objects. We conclude that the scene-inconsistent change detection advantage is a true top–down effect and is not confounded by low-level visual factors and may indeed override such factors when viewing complex naturalistic scenes.

Keywords: change detection, visual attention, visual saliency, semantic saliency


Introduction

Our inability to detect changes to our visual world has been termed change blindness and is often interpreted as evidence against the maintenance of a lasting veridical representation of a visual scene across time (see Rensink, 2002, and Simons & Levin, 1997, for reviews). The conditions of difficult detection are important to our understanding of the role of visual memory and the role of attention in perception of scenes. Of course, when we naturally view a scene, our eyes are constantly foveating different regions of interest, and information during saccadic eye movements is suppressed, resulting in a temporal gap between successive views. Change blindness has often been shown to occur if a change is made to the scene during a saccade, suggesting that detailed visual information is not maintained from one view to the next (Grimes, 1996; Henderson & Hollingworth, 2003; McConkie & Currie, 1996). Induced change blindness has also been shown to be contingent upon a number of techniques that prevent our normal motion detection systems from drawing attention to the change location in an exogenous manner (occlusion-contingent and blink-contingent changes; e.g., O’Regan, Deubel, Clark, & Rensink, 2000). These methods disrupt change detection by masking the normally present local motion transients with much larger global visual transients, such as the introduction of a blank field between successive views (Rensink, O’Regan, & Clark, 1997). Regardless of the technique used to produce change blindness, even what seem very obvious changes, such as a person’s identity, may be missed (Simons & Levin, 1998).

Rensink et al. (1997) developed the repeated change technique known as the “flicker” paradigm, in which original (A) and modified (A’) visual scenes are interspersed with a blank mask. The sequence image A–blank–image A’–blank is repeated until the change between A and A’ is detected. This paradigm allows sufficient time for a coherent scene representation to be constructed and, unlike the one-shot change detection approach, allows accuracy and reaction time to be used as dependent measures. The induced change blindness often reported using this technique has been shown to not be caused by the visual masking of the change itself (O’Regan, Rensink, & Clark, 1999). The types of changes made during the flicker task have included existence changes (additions or deletions of objects), property changes (color, orientation, etc.), location changes, and identity changes. These differing change types are detected with different speeds and accuracies, suggesting that our prechange and postchange representations may be better, or more accessible, for some aspects of the object than others. For example, in their cued change blindness task,
Aginsky and Tarr (2000) concluded that location and existence changes are better represented in visual memory than surface property changes such as color.

In their initial experiments using this paradigm, Rensink et al. (1997) made changes to regions of the scene that had been previously rated as “central” or “marginal” interest and found that changes made to objects within the central interest region were more easily detected. This suggests that attention plays a major role in change detection and that our attention is initially allocated to more “interesting” or semantically informative regions of the scene first. Change detection is also attenuated when an exogenous precue is used to orient attention to a region of the scene that is to be changed (Scholl, 2000), again highlighting the importance of attention. Change blindness paradigms can therefore be useful for informing us which objects have received focused attention and what features of those objects have been represented. The efficiency of change detection within regions of a visual scene has been taken to reflect the allocation of attention and can be used to map spatial attention (Tse, Sheinberg, & Logothetis, 2003).

Despite the requirement of attention for accurate change detection, even changed objects that are the focus of attention may still remain undetected (Ballard, Hayhoe, & Pelz, 1995; Simons & Levin, 1998), supporting the idea that attention is necessary, but not always sufficient, for change detection (Rensink et al., 1997; Simons & Levin, 1997). Rensink (2000) argues that focused attention acts upon the low-level visual properties of the scene (proto-objects), which are extracted preattentively and held in a coherence field, which maintains object representations over brief interruptions. If focused attention leaves an object, then the coherence field dissolves, preventing change detection to that object. This model assumes that previously fixated objects should show no benefit in change detection compared with nonfixated ones; however, Hollingworth (2004) reported a serial-position effect on change detection accuracy when fixation scan paths were manipulated. Further studies by Hollingworth (see Hollingworth, 2006, for a review) suggest that change detection is supported by both visual short-term and long-term memory systems and that visual representations of a scene are not as transient as was suggested by the early change blindness literature and by Rensink’s coherence theory. Change blindness can be explained in terms of encoding failures, when the to-be-changed object has not been previously fixated, or retrieval and comparison failures, when the changed object in the new scene is not refixated and the object representations cannot be compared (Hollingworth, 2003; Hollingworth & Henderson, 2002). Without attention, such representations cannot, of course, be constructed.

Two important issues raised by the change blindness literature are what object properties attract attention in a scene and whether top–down and/or bottom–up factors influence our allocation of attention when perceiving natural scenes. Studies investigating scene perception have suggested that we acquire information about the general “gist”, or scene-schema, early on in viewing, within the first 100 ms (Biederman, 1972; Intraub, 1981; Potter, 1976; Venturino & Gagnon, 1992). Once the general scene-schema has been extracted, knowledge-based information can be used to help guide attention and control gaze (see Henderson, 2003). Scene-schema knowledge can provide information about categories of objects that might be expected within a specific scene, for example, shower gel, sponge, and soap are items that one might expect to see in a bathroom scene. It might also provide likely candidate regions for the locations of such objects (e.g., soap is likely to be on a surface, close to a sink). Some studies have indicated that objects that are inconsistent with the scene-schema (e.g., a can of beans in a bathroom scene) attract early fixations during viewing. Loftus and Mackworth (1978) presented participants with black and white line drawings of scenes sometimes containing scene-schema-inconsistent objects. These were not only fixated earlier than consistent ones but also received longer fixation durations, suggesting that attentional disengagement is more difficult (Henderson, Weeks, & Hollingworth, 1999). When photographs of real-world scenes are used, inconsistent objects attract eye fixations earlier than their scene-consistent counterparts do (Gordon, 2004; Underwood & Foulsham, 2006; Underwood, Humphreys, & Cross, 2007). Hollingworth and Henderson (2000) presented participants with pairs of gray-scale line drawings within the flicker task (Rensink et al., 1997), which contained changes to semantically consistent or inconsistent objects. Visual salience was partially controlled by swapping target objects between different scenes. Their results supported previous eye-tracking studies, showing that changes to inconsistent objects in the scenes were detected faster and with greater accuracy than changes to consistent ones. This pattern of results emerged even when the possibility of using a specific encoding strategy was discouraged.

Contrary to the findings discussed above, some studies have failed to find an inconsistent object advantage in scene perception. De Graef, Christiaens, and d’Ydewalle (1990) and Henderson et al. (1999) both failed to show early fixations of inconsistent objects in eye-tracking studies. Such contradictory findings require some explanation, and one potential candidate is low-level visual saliency. Although some previous studies have attempted to control for low-level visual salience (Hollingworth & Henderson, 2000), few have directly manipulated visual salience independently of semantic salience (however, see Underwood & Foulsham, 2006; Underwood, Foulsham, van Loon, Humphreys, & Bloyce, 2006; Underwood et al., 2007).

This brings us back to the question of what attracts attention while searching visual scenes for changes. Accurate change detection is supported by visual attention,
which may be controlled exogenously by stimulus features or endogenously by existing or accumulated knowledge about the scene. The inconsistent-object effect supports a role of top–down allocation of attention, but this can also be confounded by bottom–up factors such as visual salience. The Iti and Koch (2000) saliency map model predicts that attention will be allocated to regions of the scene based upon the ordering of saliency peaks, and therefore, such low-level stimulus properties must be carefully controlled when examining potential semantic (top–down) effects. Kelley, Chun, and Chua (2003) made changes (color changes or deletions) to two objects in colored photographs of complex visual scenes and measured reaction times to detect a change during a flicker paradigm. One of the objects was of high contextual significance (central interest) and the other was of low contextual significance (marginal interest). Both objects were matched for color in an attempt to control for low-level visual saliency; however, no objective measure of the overall visual salience of changed regions was used. Their finding that central changes, which are semantically more informative, were preferentially detected compared with marginal changes could therefore still be confounded by bottom–up visual processes. This study by Kelley et al. recognized that competing visual changes need to be matched for visual salience. However, their attempt to match objects for color does not guarantee equivalent overall salience. Background contrast differences and feature orientations of competing objects were not really measured objectively, and therefore, objects that were different in terms of scene interest may have also been nonequivalent in terms of visual salience.

This study addresses the potential confound of low-level visual salience being responsible for the inconsistent-object advantage by independently manipulating semantic (scene-schema consistency) and visual salience. This allows us to determine whether the inconsistent-object advantage reported by some is simply a by-product of low-level visual factors attracting early attention. If this is the case, then we would expect to find an interaction between visual and semantic salience when both are independently manipulated. That is, responses to inconsistent objects would be faster only when they are also highly visually salient. The present experiment made changes to both scene-consistent and -inconsistent objects, which were objectively rated as having high or low visual salience using a software derived from the computational model of Iti and Koch (2000). Based on this model, we might hypothesize that change detection will be faster and more accurate when changes are made to objects with high visual salience compared with those with low salience, as these objects would afford earlier fixations. This would be in alignment with the findings of Wright (2005), who reported that change detection in natural scenes could be predicted from a single overall saliency output based on a topographic saliency map (Itti & Koch, 2000). He found significant positive correlations between three subjective measures of salience and percentage change detection; however, when quantitative measures of visual components of the images were calculated (spatiotemporal contrast differences and channel response differences), none of these were found to be significant predictors of change detection. This highlights that change detection is not dependent on any individual visual characteristic of the scene but more likely on a combination of such characteristics, which, in addition, may be influenced by top–down processes. When semantic information is low, bottom–up processes may have a greater influence on the allocation of attention and, therefore, on change detection. When semantic information is high, as is often the case with complex scenes, low-level factors such as visual salience may be less important.

This article examines factors affecting the allocation of visual attention in real-world natural scenes using a change detection paradigm. Unlike many previous studies, we use an objective measure of visual saliency, which we manipulate independently of the scene-schema consistency of the object being changed.

**Methods**

**Participants**

A total of 24 participants (4 men and 20 women; \( M \pm SD \) age = 26.3 ± 4.8 years) were recruited by posters and word of mouth to take part in the experiment. All were paid an inconvenience allowance for taking part, and all had normal or corrected-to-normal vision.

**Stimuli and design**

The stimuli were digital photographs of everyday indoor scenes such as office desks, kitchen areas, bathrooms, and others. All photographs were JPEGs and were resized using Adobe Photoshop to 443 × 591 pixels (width × height), each having a resolution of 300 dpi. At an average viewing distance of 60 cm from the screen, the stimuli subtended 19.7° × 13.9° of visual angle (20.8 × 14.6 cm). Each scene was posed five times to provide one unchanged/original image and four changed images. For each of the four variants of the scene, one object from the original image was replaced by a different object occupying the same spatial location. Objects were matched for similar size. The changed object was carefully controlled for both low-level visual salience (high vs. low salience) and for scene-schema consistency (scene consistent vs. inconsistent). A saliency map for each image was computed using a software package based on the
computational model of visual attention of Itti and Koch (2000). This program predicts the order of fixations (scan paths) within an image, in a purely bottom–up fashion, based on variations of orientation, color, and intensity. This model does not factor in some low-level stimulus properties that have been shown to be associated with attention, such as contour length (Shashua & Ullman, 1988). Forty-two calculated individual feature maps are summed to provide an overall two-dimensional saliency map, which is thought to guide the focus of attention through a winner-take-all mechanism. Locations of features are not taken into account during the saliency calculations, and the model has no knowledge of what features combine to make a perceptual object. The changed objects were classified as having either high or low visual salience according to the Itti and Koch algorithm, and this was determined by examining the predicted fixation order of scene regions. When the changed object had a saliency peak within the three highest peaks, it was considered here to be of high salience, and when the object had a saliency value with an ordinal rank of eight or higher, it was considered to be of low salience. Mean salience ratings for the objects are presented in Table 1.

In addition to controlling the visual salience of the changed object, half of the changes were made by replacing an object with one that was not consistent with the general scene-schema (e.g., a mug in a kitchen scene was replaced by a bottle of shower gel). The two independent factors, visual salience and scene-schema consistency, were combined to give four within-subject conditions in the design.

The original image was paired with the changed image during the flicker sequence (see Procedure section), and scenes were rotated across the four conditions to produce four versions of the experiment, which were ran on four separate groups of participants. Forty different change scenes were used per version of the experiment (10 scenes per condition), resulting in 40 original–changed image pairs. During the flicker trial procedure, half of the pairs

Table 1. Mean salience ratings for objects across the four conditions of the design. Ratings represent the mean ordinal fixation number, which was calculated using the Itti and Koch (2000) algorithm. Values inside parentheses show SD.

<table>
<thead>
<tr>
<th>Salience</th>
<th>Consistent</th>
<th>Inconsistent</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>1.68 (0.73)</td>
<td>1.55 (0.68)</td>
</tr>
<tr>
<td>Low</td>
<td>13.25 (4.28)</td>
<td>13.70 (4.39)</td>
</tr>
</tbody>
</table>

Figure 1. (A) An example scene (shower) showing the presence of an inconsistent/high visually salient object (toilet roll) highlighted in circle (circle not present in original stimuli). (B) Output from the Itti and Koch (2000) algorithm, with toilet roll having the third highest saliency peak and therefore predicted to attract an early fixation. Images not shown in original size.
were presented in the original–changed image order, and half were presented in the changed–original image order. A further 40 scenes were paired with their identical image to be used as catch trials (trials in which no change occurred to the image during the flicker procedure). This resulted in 80 experimental trials, which were presented in a unique randomized order per participant. Eight new sets of image pairs (four change pairs [one image pair for each condition] and four no-change pairs) were constructed in the same way as above and were used in a practice block (see Procedure section). All images were finally converted to Windows BITMAP files, ready for use in the experimental software package. An example image, along with its saliency map output, can be seen in Figure 1. Some examples of image pairs can be seen in Figure 2.

Apparatus

The experiment was constructed and ran using the E-Prime experimental design software package. Stimuli were presented in the center of a 17-in. flat-screen monitor at a screen resolution of 1,024 × 768 pixels.

Procedure

Participants were informed that they were going to see two pictures in very quick succession, separated by a brief grey blank field. They were told that the sequence image–blank–image–blank would be repeated until they made a keyboard response. It was explained using on-screen instructions that the aim of the task was to determine whether the two successively presented images were the same or different, in other words, to detect whether a change occurred. Responses were made by pressing a key on the keyboard labeled “SAME” or “DIFFERENT” (the alphabetic Z and M keys were marked, respectively). Once the participant had read and understood the instructions, two blocks of trials were presented in succession. The initial practice block of eight trials included feedback on accuracy and reaction time to determine whether a change occurred between the images.
presented. Once the practice block was completed, instructions were redisplayed before the 80-trial experimental block began. No feedback was provided during the experimental block.

**The flicker procedure**

Each trial began with a fixation cross (+) presented for 3,000 ms in the center of the screen. The cross was presented in black against a white background using size 30 Courier New font in bold. This was followed by presentation of the first image against a white background in the center of the screen for 480 ms, which was replaced by a uniformly dark grey mask covering the entire screen (1,024 × 768 pixels) for 160 ms. This was followed by a 480-ms presentation of the second image, which was either the same as or different from (one object changed) the first image. Finally, another dark grey mask replaced the second image for 160 ms, before the sequence looped back to the start. The 480-ms image presentations are greater than the time proposed by Potter (1976) to process a scene and therefore allow ample time for storage of image properties in memory. The sequence was repeated until the participant made a “same” or “different” response. **Figure 3** shows the general structure of the flicker procedure. Response time to press the Z or M keys and accuracy were measured per trial. Timing started from the onset of the initial image, and any preemptive responses before the second image was displayed were removed. **Movie 1** shows an actual QuickTime demonstration of an example flicker.

**Results**

Reaction times (in milliseconds) and accuracy (% correct) to determine whether a change occurred or not were subjected to a fully within-subjects factorial ANOVA. Only reaction time data from accurate responses on change trials were analyzed. Mean reaction times to correctly detect a change are presented in **Figure 4A**.

Analysis of the reaction time data revealed that there was a significant main effect of scene-schema consistency, $F(1, 23) = 5.38, p = .03$ (significant), with reaction times for inconsistent-object changes being smaller (faster) than for consistent ones ($M = 2,341.7$ vs. $2,549.2$ ms, respectively). There was no main effect of visual saliency,
(1, 23) = 1.78, p = .20, and no interaction between the two factors, F(1, 23) = 0.11, p = .75. The analysis of the percentage accuracy data (see Figure 4B) followed the same pattern, uncovering a significant main effect of scene-schema consistency, F(1, 23) = 15.55, p = .001. Participants made significantly more accurate responses in the inconsistent-object change condition than in the consistent one (M = 87.3% vs. 78.3%, respectively). Once again, neither a main effect of visual saliency nor a significant interaction was observed, F(1, 23) = 0.26, p = .62, and F(1, 23) = 0.72, p = .40, respectively. This pattern of results supports the absence of a speed–accuracy trade-off.

Discussion

Saliency models of scene inspection suggest that regions with high visual salience receive early eye fixations, and therefore, objects attracting these fixations should show a detection advantage in a change detection task (see Findlay & Walker, 1999, for a model of eye guidance). This study compared speed and accuracy at detecting changes to objects that were objectively classified as having high and low visual salience and that were also either consistent or inconsistent with the scene-schema. Our findings support previous work...
reporting an inconsistent-object detection advantage (Loftus & Mackworth, 1978; Underwood & Foulsham, 2006; Underwood et al., 2007); however, this advantage was seen to occur irrespective of differences in objectively measured low-level visual salience. The lack of an interaction between visual salience and scene-schema consistency, combined with an absent main effect of visual salience, strongly supports the idea that the congruency/scene-schema consistency effect is not simply a result of attention being allocated to regions of high visual salience, which may correlate with inconsistent objects. Our findings support the conclusions drawn by Kelley et al. (2003) that changes to objects in the scene that are more semantically informative/salient (central vs. marginal interest objects and scene-schema-inconsistent objects) show a detection advantage in the flicker paradigm. Rather than attempting to compete two changes within a scene that were subjectively matched for low-level visual salience (color, size, etc.), as Kelley et al. did, we made changes to a single object in half of the scene presentations and included catch trials. Instead of matching scene-consistent and -inconsistent objects for visual saliency by equating them in terms of color, brightness, and contrast, we objectively rated their visual salience as high or low, based on the model of Itti and Koch (2000). Despite these objectively determined differences in visual saliency, change detection speed and accuracy for high- and low-saliency objects did not differ significantly.

The possible confound of visual saliency has been used as a potential explanation (see Henderson et al., 1999) for the contradictory findings regarding the inconsistent-object effect. However, the results of this study suggest that knowledge about the categories of objects likely to appear within a scene can guide visual attention independently of guidance by low-level visual processes. The finding that low-level visual processing did not affect change detection in this study suggests that a model of eye guidance based on just the visual properties of a scene cannot fully account for adequate scene perception. The Itti and Koch (2000) model is a purely visual salience-driven model that does not account for and cannot explain semantic effects, such as the inconsistent-object advantage described in this article and elsewhere. Models based on a unitary measure of saliency fail to predict change detection, as Wright (2005) reported. In his change detection study, only subjective measures of salience correlated with change detection, whereas a number of saliency measures based on objective low-level visual characteristics of the scene failed to predict change detection behavior. If we consider models of eye guidance that allow for top–down information to bias the allocation of attention within a scene (see Findlay & Walker, 1999, or Torralba, 2003), then Wright’s findings on subjective salience can be better explained. Subjective ratings of object salience are likely to be based on a combination of a number of different saliency measures averaged and derived from both visual and semantic information.

Although attention and, therefore, change detection may be controlled by bottom–up visual processes when scenes are semantically uninteresting (Wright, Green, & Baker, 2001, used simple Gabor stimuli), more naturalistic scenes tap existing semantic knowledge, and this may override the attention-grabbing nature of high visual salience, leading to the inconsistent-object detection advantage. Such semantic information builds up over successive fixations and has been shown to affect the allocation of attention as early as 150 ms into viewing (VanRullen & Thorpe, 2001).

Possible explanations for the inconsistent-object detection advantage within a change blindness paradigm have been suggested by Hollingworth and Henderson (2000) and Hollingworth and Henderson (2003). One candidate explanation is that semantically salient (scene-inconsistent) objects attract both earlier fixations and longer fixation durations than less salient objects do. It is therefore more likely that an inconsistent object is the focus of attention as the change occurs and will therefore be more rapidly detected. However, other explanations that do not rely strictly on attentional factors should be considered. To successfully detect change across the blank interstimulus interval used in the flicker paradigm, representations of the prechange and postchange objects must be maintained in visual short-term memory. The attentional engagement theory suggests that violations to the scene-schema attract attention and that the input of attentional resources to the fixated location supports a stronger perceptual representation of the object. This offers an advantage for change detection compared with less semantically salient objects whose representations may be weaker. Another plausible explanation for the effect is that attention is initially allocated according to a low-level visual saliency map, but that attention may be later captured by conceptual processing difficulty of inconsistent objects and takes time to disengage. This again leads to a better representation in memory, supporting change detection.

How are changes detected? Rensink et al. (1997) point out that they are only detected when attention is allocated to the changed object. But what attracts attention? This study demonstrates that low-level visual saliency is not associated with the early capture of attention in this task, and this might be used as an indication of the role of memory in change detection. Our previous investigations have demonstrated that saliency does attract early attention when viewers are encoding the scene in preparation for a recognition memory test (Underwood & Foulsham, 2006; Underwood et al., 2006). In this task, extensive scanning of the image is observed as each prominent object is inspected. In contrast, when set the task of determining whether a specific target object is present in the same real-world scene, visual saliency does not capture attention.

Extrapolating to the present task and results would lead to the conclusion that, because visual saliency is not associated with change detection, the task does not require
general encoding of the scene and that it can be performed with minimal involvement of visual memory. The detection of the changed object must involve a visual representation of the object that is retained over a period of a few hundred milliseconds, however, but such a very short-term memory of the image would be sufficient to deliver detection in this task.

Conclusions

In summary, this article provides evidence that refutes a model of change detection based solely on the visual properties of a scene and finds that semantic salience can be used to differentiate between speed and accuracy to detect changes to objects in natural scenes, independent of visual salience.

Acknowledgments

We would like to thank Emma Templeman and Laura Lamming for help in preparation of the photographic stimuli. We are also most grateful to Laurent Itti, who kindly provided us with his saliency algorithm software.

Commercial relationships: none.
Corresponding author: Jonathan A. Stirk.
Email: jas@psychology.nottingham.ac.uk.
Address: School of Psychology, University of Nottingham, Nottingham, NG7 2RD, UK.

References


