Quarterly Journal of Experimental Psychology: in press

Visual saliency and semantic incongruency influence eye movements when inspecting pictures

Geoffrey Underwood and Tom Foulsham
University of Nottingham, Nottingham, UK

Abstract

Models of low-level saliency predict that when we first look at a photograph our first few eye movements should be made towards visually conspicuous objects. Two experiments investigated this prediction by recording eye fixations while viewers inspected pictures of room interiors that contained objects with known saliency characteristics. Highly salient objects did attract fixations earlier than less conspicuous objects, but only in a task requiring general encoding of the whole picture. When they were required to detect the presence of a small target, then the visual saliency of non-target objects did not influence fixations. These results support modifications of the model that take the cognitive override of saliency into account by allowing task demands to reduce the saliency weights of task-irrelevant objects.

The pictures sometimes contained incongruent objects that were taken from other rooms. These objects were used to test the hypothesis that previous reports of the early fixation of congruent objects have not been consistent because the effect depends upon the visual conspicuity of the incongruent object. There was an effect of incongruency in both experiments, with earlier fixation of objects that violated the gist of the scene, but the effect was only apparent for inconspicuous objects, arguing against this hypothesis.

Address for correspondence: Geoffrey Underwood, School of Psychology, University of Nottingham, Nottingham NG7 2RD, UK. Email: geoff.underwood@nottingham.ac.uk or lpywtf@psychology.nottingham.ac.uk

We are very grateful to Laurent Itti who generously made his saliency software available to us, and to Editha van Loon who helped install the software. We also thank Mike Burton, John Findlay and two anonymous reviewers for their helpful comments on a previous draft of this paper.

What attracts our attention when we first look at a picture of a scene such as a kitchen, a football match, or a harbour? Conspicuous objects might be expected to gain early inspection, and in two experiments here we investigated the effects of visual saliency and scene incongruency. The effects of visual and semantic conspicuity were observed in a free inspection task in which viewers prepared for a recognition memory test, and in a search task, in which they looked for a specific object. In each case we asked whether their early eye fixations would be taken to objects that were visually prominent by virtue of characteristics such as brightness and colour, and to objects that violated the gist of the scene by virtue of not being in their expected environment.

Itti and Koch (2000) have developed an algorithm that enables the measurement of the visual saliency of an image on the basis of its physical properties, by the identification of peaks in the distribution of intensity and changes in colour and orientation. The algorithm builds an overall "saliency map" of the image that was suggested by Koch and Ullman (1985) to drive attentional selection of regions of displays. By assuming that attention is drawn to changes in the environment, then a composite description of the changes from one area to another in an image will provide the basis for predictions about when attention should be directed. The saliency distribution therefore generates predictions as to where in a scene attention and eye movements should be guided, and forms the basis of the Itti and Koch model of visual attention. This model relies upon the low-level visual characteristics of the image to build the saliency map, which in turn determines in what order the objects in the scene should be inspected. In the case of two-dimensional images such as pictures these characteristics are colour, intensity and orientation, but with dynamic displays such as movies the relative motion of an object would also contribute to its saliency value (Itti, 2005). For each of the characteristics a separate saliency map is first computed by searching for change relative to adjacent regions. The separate maps are then combined to find saliency peaks, with a change in any of the three characteristics resulting in an increase in the saliency value assigned to that region of the image. In the image shown in Figure 1, for example, the most visually salient region is the ashtray on the coffee table. This object is differentiated from its surround by variation of intensity (it is bright), colour (white, on a brown surface) and orientation (circular components), and the Itti and Koch algorithm picks this region as having the greatest saliency value. Once a region has been inspected its saliency weighting is reduced, to initiate an inhibition of return process without which inspection would be restricted to the two most salient peaks.

The analysis of low-level visual information is also central to the Henderson, Weeks and Hollingworth (1999) "saliency map framework" in which the first fixation is attracted to the region with the greatest weighting, and the duration of that fixation is determined by the complexity of processing. Only at this point does the map start to incorporate meaningful information about the gist of the scene, and objects are then identified. The meaning or gist of the scene is extracted only after the saliency map has been built.

Findlay and Walker (1999) have also proposed a model of eye guidance in which a "salience map" influences the decision about the location of the next fixation. This map is a spatiotopic representation of weightings that are troughs and peaks in the distribution of information about a scene. In this model, as in the Itti and Koch (2000) model, the principles of the Koch and Ullman (1985) representations of saliency peaks, determine the decision about which object to inspect next in visual search tasks. The currently dominant peak controls the saccadic trajectory with a "winner-take-all" process that selects the highest peak and then directs attention to the location on the map that is represented by that peak. This model of eye guidance builds the saliency map with low-level visual analyses of the scene, but differs from the Itti and Koch (2000) and Henderson et al. (1999) versions in that it has a role for top-down cognitive factors in the selection of saccadic targets and in modifying the saliency map.

In each of these models, low-level visual processes determine the early fixations during picture inspection, and only after fixation can the saliency map of a scene incorporate semantic information. The distinguishing feature of the Itti and Koch (2000) saliency map model is that it has been implemented in software that can be used to build a representation of the saliency peaks in a picture, and these peaks form the basis for predictions about the early fixation of objects in that picture.

The saliency map provides directions for attention to move around the image, according to Itti and Koch (2000), with the region of greatest saliency attracting attention first in a winner-take-all algorithm. This saliency peak is then suppressed by a process of inhibition of return, to enable attention to be disengaged from this region and attracted by the next most salient peak. In each operation attention moves to the next most salient region. Itti and Koch evaluated the predictions of the saliency model with search tasks using photographs of natural scenes and using geometric shapes in a conjunctive search (Treisman & Gelade, 1980). With a single feature change (a red rectangle in a display of green rectangles, or a rectangle oriented at right angles to a set of background rectangles) the model predicts a pop-out effect, with the first fixation expected to identify the singleton, as it does when human observers see these displays. With a conjunctive search, the time taken to find the target, and the number of fixations required, depends upon the number of distractors, with both the Itti and Koch model and with the observers they tested. Nothdurft (2002) reported a similar result with single-feature variations in displays, again suggesting that target saliency attracts focal attention in pop-out. One of the strengths of the saliency model is its prediction of pop-out in search tasks with simple geometric shapes, and the saliency values of distractors also determine search times (Lamy, Leber & Egeth, 2004).

Many of the studies that support the saliency map model of visual attention rely upon search tasks with simple displays of targets and distractors. One of the few exceptions is a study of eye fixations reported by Parkhurst, Law and Niebur (2002) who confirmed the predictions of the model, and extended it to emphasize the importance of the relationship between visual sensitivity and stimulus eccentricity. In their study a range of images were shown, including photographs of home interiors, buildings and city scenes, and natural environments, as well as computer-generated fractals. Viewers inspected each image for a few seconds while their eye movements were recorded. The saliency values of regions of each image provided a good prediction of the order of fixations, especially for the first few fixations. Saliency strongly predicted fixation probability during first two or three fixations, but the model performed above chance throughout each trial. Parkhurst et al. concluded that a purely bottom-up account of visual attention was sufficient to account for fixation behaviour, although their viewers were instructed only to "look around at the images". These instructions possibly precluded the top-down influences seen in the search experiments reported by Rao, Zelinsky, Hayhoe and Ballard (2002) and by Wolfe, Horowitz, Kenner, Hyle and Vasan (2004), in which expectations were influential. Itti and Koch's (2000) model was also supported in one of the experiments reported by Underwood, Foulsham, van Loon, Humphreys and Bloyce (2006). Viewers looked at photographs of office scenes in preparation for a recognition memory test, and early eye fixations were found to be more likely to be on a high saliency object (e.g., a brightly coloured coffee mug) than upon other objects in the scene (e.g., computer equipment, books, keys, and a piece of fruit). In a second experiment, in which viewers searched for a specific example of a low saliency object (the piece of fruit) saliency had a lesser effect, leading to the conclusion that the saliency weights can be modulated by cognitive influences such as the need to look for a specific object.

When we search scenes we do so purposely, and with cognitive override of visual saliency. In the sentence verification task used by Underwood, Jebbett and Roberts (2004), viewers scanned the whole photograph to encode as much content as possible when the sentence was to appear after the image, because they had to remember the scene before being asked to judge a statement about it. When the sentence was presented first and they knew what they were looking for, fixations were directed towards the objects mentioned in the sentence. The present experiments compare the predictions of the Itti and Koch (2000) saliency model with actual fixation behaviours in two tasks that give varying emphasis to top-down influences.

It is not only visual conspicuity that can produce a pop-out effect in the inspection of an image. It has sometimes been reported that an object that violates the gist of a scene can attract eye fixations earlier than the same object placed in a congruent context. This form of conspicuity might be termed semantic saliency, but for purposes of clarity here we will restrict the use of the term saliency to bottom-up visual features, and refer to violations of the scene schema or gist as an effect of incongruency or scene inconsistency. Loftus and Mackworth (1978) presented line-drawings of scenes while recording eye movements, and found that objects that were incongruous were fixated earlier than others (for example, an octopus in a farmyard scene). More recently a study by Gordon (2004) reported an effect of incongruous objects in a task requiring a decision about the identity of a simple probe stimulus that was located near to an object that was congruous or incongruous relative to the scene. Decision times were influenced by the congruency of the object. These results suggest that information about the gist of a scene can be extracted early and that incongruency can be detected prior to the fixation on an incongruous object. The gist of a scene is the overall meaning of what is being represented, such as a bathroom, or a roadway, or a ski slope (for a recent review see Underwood, 2005). The gist can be identified in less time than it takes to make the first eye movement around a scene (e.g., Biederman, Mezzanotte & Rabinowitz, 1982; Potter, Staub, Rado & O'Connor, 2002), and violations of gist can influence the identification of objects within the scene in this time (e.g., Davenport & Potter, 2004). The early interaction between an object and the overall gist suggests that both can be identified prior to the first eye movement. If an object is identified that violates the gist, it may then attract attention either because the current schema may then need to be revised, or because the identity of the incongruous object needs to be confirmed. The early detection of incongruous objects is plausible, but has not always been found to be associated with early eye fixations.

De Graef, Christiaens and d'Ydewalle (1990) recorded eye fixations while viewers inspected line-drawings to detect non-objects. Also in the drawings were objects that violated the gist of the scene, such as a parking meter in a laboratory or a petrol pump in a playground. There were effects on the duration of the first fixation of an incongruous object, but not on the time taken to first fixate that object. This argues against the possibility of a gist violation attracting an eye movement, and using a similar set of line-drawings Henderson, Weeks and Hollingworth (1999) reported a similar effect. They found an effect of incongruency on fixation durations, but not upon the number of fixations required to first fixate on object. Reports of object incongruency influencing the time to first fixate the object are unsupported, therefore, even though studies of the perception of briefly presented pictures have established that the gist of the scene, and violations of the gist, can both be identified very early.

One possible explanation of these conflicting results is that the successful demonstrations of early fixations on incongruous objects may have used objects that were visually conspicuous. Perhaps incongruous objects only attract early fixations when they are highly distinguishable from their backgrounds. It is possible that the incongruency effect reported by Loftus and Mackworth (1978) was not an effect of the violation of gist, but an effect of low-level visual saliency. In contrast, when De Graef, Christiaens and d'Ydewalle (1990) and Henderson, Weeks and Hollingworth (1999) failed to find effects it was perhaps because their incongruous objects had low saliency. High visual saliency may be responsible for the appearance of an incongruency effect, or it may simply be a confounding factor. The present experiments investigate the incongruency effect with visual saliency controlled, using photographs of real scenes. Establishing the saliency values of line drawings is possible, but Itti and Koch's (2000) three dimensions of colour, intensity and orientation would then generate a map using mainly orientation, and with a contribution from the density of lines that reduced the intensity of some regions. To isolate the effects of visual saliency and semantic congruency the following experiments used photographs of indoor room scenes with specific objects selected and placed on the basis of their visual conspicuity and on the basis of being in an expected or unexpected indoor location.

Rooms were photographed as being readily recognisable and distinctive. As such, they each had an identifiable gist or scene schema. Two objects were placed in each picture, one with high visual saliency and the other with low saliency. The saliency values were determined using the Itti and Koch (2000) program. In addition, objects were taken from another room in the house, and therefore violated the gist of the scene. For example, a stapler appeared in the bathroom scene, and a bathplug appeared on an office desk. The incongruency of these violating objects was established with a separate screening experiment in which participants judged the consistency of each object within the scene. Either, both, or neither of the two critical objects could be congruent or incongruent, but one of them was a highly salient object, while the other was much less conspicuous. While viewers inspected these pictures, their eye movements were recorded, and the experiments investigated the possibility of early fixations on objects being associated with high saliency or with semantic incongruency. In the first experiment viewers encoded the pictures in preparation for a recognition memory test, and in the second experiment they searched for a small target object that had also been introduced into some of the pictures. By comparing the effects of saliency and congruency in the two experiments, it is possible to observe top-down task influences on picture perception. In one experiment the task was to encode as much of the scene as possible, and in the other a focused search was required and much of the scene could be neglected. The search experiment was used to ask whether the visual conspicuity of objects would be influential in attracting eye fixations when those objects were irrelevant to the task being performed.

EXPERIMENT 1

General encoding of a picture

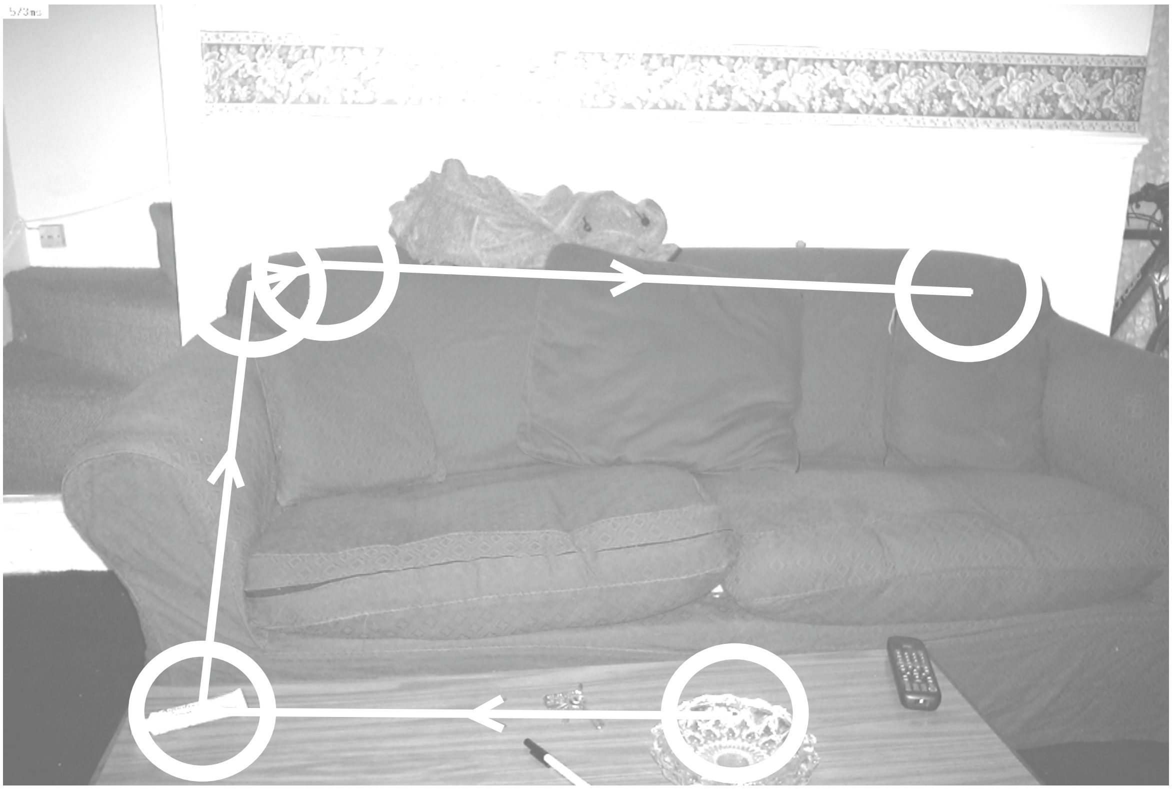
The task in this experiment was to encode each picture in preparation for a recognition memory test. This task was designed to match that used by Henderson et al. (1999) in their first experiment, and to indicate the pattern of inspection when the whole scene was of relevance and when no particular object was of special importance to the viewer. The memory test was only administered during a practice session, as our interest was with the distribution of visual attention during inspection, and specifically whether the Itti and Koch (2000) model of visual saliency provided a good prediction of the early eye fixations upon the scene.

Method

Participants

Sixteen students (aged 18-25 years) with normal or corrected-to-normal vision participated in this experiment. Two subjects were replaced as data were missing from over half of trials due to not fixating centrally at the beginning of a trial or not fixating the objects of interest.

Figure 1. One of the photographs used, with graphical output from the saliency software, showing the five most salient areas, with circles. Pictures were displayed in the experiments in colour. In this example, the most salient object is the ashtray on the coffee table, and the tube of toothpaste (near the bottom left in the picture) is visually conspicuous and out of place, whereas the TV remote control (also on the table, near the bottom right in the picture) is inconspicuous and consistent with the scene. Note that the saliency algorithm identifies the toothpaste tube as the second most salient point, while the TV remote control does not feature in the first five peaks.



Stimuli and Design

The stimuli were digital photographs of rooms in a house, displayed on a colour computer monitor at a resolution of 1024 by 768 pixels. Viewing distance was fixed at 60 cm from the participant, giving an image that subtended 31 by 25 degrees from this seating position. There were 32 experimental stimuli, all of which contained two objects of interest alongside other objects and items of furniture found in a house environment. Four types of room were used equally often (kitchen, living room, bathroom and an office desk), with several instances of each type included. The scenes and the objects used are listed in the Appendix. The two principal objects were of a similar size. Each object could be located anywhere in the scene, although they were always on different sides of the picture, and equidistant from the centre. They were manipulated on two dimensions: visual saliency and scene congruency.

A saliency map of each picture was computed using the software developed and described by Itti and Koch (2000). This map identifies the visual saliency of each part of the image according to variations in orientation, intensity and colour. Salient objects are thus objects that stand out from their background. The criteria for all the photographs here was that one object should have high visual salience and therefore be one of the three most salient points ("peaks") in the image, whilst the other object had low visual salience and did not feature in the first five peaks. These objects will be described as having high or low visual saliency. Figure 1 shows an example of the graphical output from the saliency algorithm with the most salient objects linked in a series of circles. The ordering of saliency peaks provides the basis of the model's predictions about the ordering of eye fixations when first inspecting the scene.

The congruency of the two objects of interest was also manipulated by altering the semantic consistency of each object within the scene. Each object used in the experiment was highly associated with one of the rooms (for example, a food whisk in a kitchen scene) and was inconsistent with the others (the same whisk in a bathroom scene, in this example). In the picture shown in Figure 1, the incongruent object was the tube of toothpaste on the living room coffee table, and the congruent object was the TV remote control. Each object featured in its congruent and incongruent contexts, providing a control for any spurious differences between objects. This matching of stimuli was checked in a pilot investigation by showing a set of modified stimuli to a separate group of ten participants drawn from the same population as those in the main experiments. Each participant saw the set of experimental photographs with one of the two objects highlighted and was asked to rate the consistency of the object with the scene on a scale from 1 (highly inconsistent) to 9 (highly consistent). A mean consistency rating was calculated for each object/room combination. Some instances (for example a remote control on a desk) were rated as neither wholly consistent nor inconsistent, and so were discarded. An analysis of variance (ANOVA) conducted on the mean consistency ratings for consistent and inconsistent objects used in the experimental stimuli found the effect of the manipulation to be reliable. Objects described as congruent here were rated as being more consistent with their scene (mean rating, 7.77, sd=0.58) than objects that we described as incongruent (mean rating, 2.67, sd=0.62), $F(1,9) = 415.6$, $p<0.001$.

The combinations of visual saliency and scene congruency gave four conditions, each containing eight pictures. These conditions were:

(i) congruous high saliency object plus congruous low saliency object;

(ii) congruous high saliency object plus incongruous low saliency object;

(iii) incongruous high saliency object plus congruous low saliency object;

(iv) incongruous high saliency object plus incongruous low saliency object.

Eight additional pictures were composed to give practice for the memory task that showed none of the rooms or objects included in the experimental stimuli.

Apparatus

Eye movements were recorded using a SensoMotoric Instruments (SMI) EyeLink system that was head mounted and recorded pupil position from the right eye every 4 msec, and that was spatially accurate to within 0.5°. Fixations were described as having terminated when a movement of at least 35 deg/sec was detected by the tracker. Head position was recorded remotely and a chin rest was used to maintain a constant viewing distance and to minimise head movements.

Procedure

Following calibration with the SMI eye-tracker, participants were given written instructions. They were told to view the scenes “in preparation for a memory test”. In the practice session, after viewing four scenes, a two-alternative forced choice recognition test was administered with one previously seen picture and one that differed slightly (for example in the position of an object or a piece of furniture). In the main experiment a recognition test was never actually given, although verbal report suggested participants expected it.

In the experimental session, the 32 pictures were presented to each participant in a unique randomised order. Each picture was preceded by a central fixation cross and a drift correction marker that confirmed that fixation was in the centre of the screen. Each picture was displayed until the participant pressed a computer key.

Figure 2. A visual representation of the first seven fixation locations made by one subject overlaid onto the same photograph as Figure 1. The first fixation was located in the centre of the picture, and attention moved to the highly salient object (the toothpaste tube near the bottom left of the picture) on the third fixation. Circle diameter represents fixation duration, and the circles and movements between fixations are drawn in black here to distinguish observed fixations from the saliency peaks that are drawn in white in Figure 1.

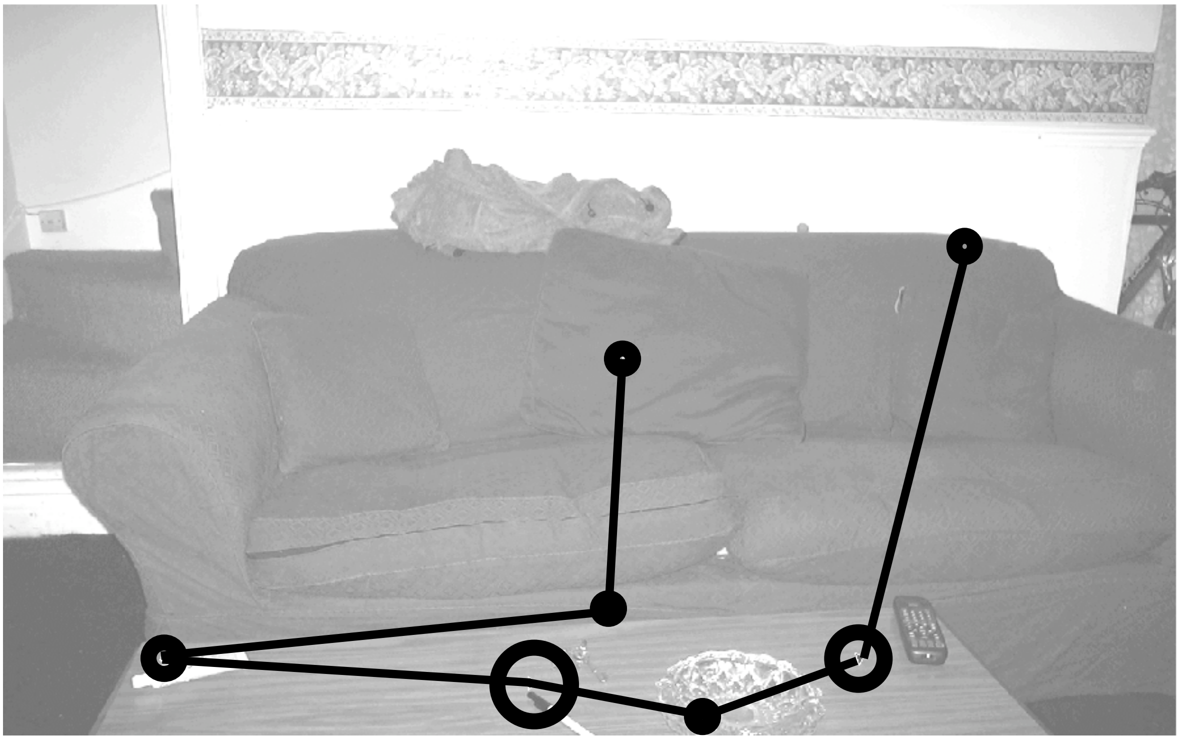


Table 1. Means of the measures taken in Experiment 1, in which the picture was encoded in preparation for an anticipated memory test. (Standard deviations are in parentheses).

High Visual Saliency Object:	**Congruent**		**Incongruent**	
Low Visual Saliency Object:	**Congruent**	**Incongruent**	**Congruent**	**Incongruent**
Overall Inspection				
Total inspection time (sec)	6.26 (3.10)	6.00 (3.23)	5.64 (2.58)	5.27 (2.47)
High Saliency Object				
Time prior to fixation (sec)	1.86 (1.04)	1.70 (1.51)	1.44 (0.58)	1.69 (0.65)
No of fixations prior to fixation	6.24 (2.86)	5.32 (3.60)	4.89 (1.58)	5.46 (1.53)
1st gaze duration (msec)	360 (111)	372 (157)	398 (140)	498 (258)
Low Saliency Object				
Time prior to fixation (sec)	3.00 (1.55)	2.11 (1.10)	2.34 (1.27)	1.81 (1.69)
No of fixations prior to fixation	10.12 (4.85)	7.17 (2.67)	7.73 (3.74)	5.87 (3.57)
1st gaze duration (msec)	374 (132)	550 (303)	348 (108)	526 (208)

Results and Discussion

The eye tracking data for each participant consisted of position co-ordinates for each time sample. Fixations longer than 100msec were included and compared to the known pixel co-ordinates of the two principal objects in each picture: an object was considered to be fixated when gaze was within a rectangle that enclosed the object and that had a standard size for all objects. The initial fixation for each picture had to lie within one degree of the central co-ordinates of the picture for the trial to be included. This was encouraged by the presentation of a central fixation cross at the start of each trial. Figure 2 shows a schematic representation of the eye-movements of one participant.

Several measures were taken to address two main questions. Firstly, the time before an object was fixated, and the number of fixations elsewhere in the scene prior to its fixation, were calculated to see how early each object was fixated. Secondly, to investigate how much attention was paid to each object, the first gaze duration was recorded. Total inspection time was also calculated for each picture, as an indication of picture processing difficulty.

The means of the measures used are presented in Table 1. Participant means were used to perform a number of within-groups analysis of variance (ANOVA) on each measure, first to test the influence of visual saliency (high/low) and then to inspect the fixations on each of the two objects for scene congruency (high/low).

Total inspection duration

The task required inspection of each picture, in preparation for a memory test, and inspection was self-paced. The total inspection time is interval between appearance of the picture and the participant pressing the computer key to indicate that they were ready for the next trial. As the task was to prepare for a memory test this time might be indicative of the perceived complexity of the image and therefore the amount of information that needed to be memorised.

A two-factor ANOVA was carried out with the consistency of each object as factors. The semantic consistency of the salient object had a significant effect, $F(1,15) = 5.00$, $p<0.05$, with longer inspection of pictures containing a high saliency object that was congruent (6.13 sec) rather than one that was incongruent (5.46 sec). There was no effect of the congruency of the low saliency object, $F(1,15) = 2.14$, and no interaction, $F<1$. The overall encoding of a picture for a subsequent recognition test was extended only by the congruency of the most salient object, with another incongruent object having minimal effect.

Time prior to the first fixation on an object

This measure indicates the potency of an object in attracting early attention using non-foveal vision, and is the time elapsed between onset of the picture and the first fixation of one of the two critical objects. Objects that are fixated sooner are assumed to be more potent in attracting attention than are other parts of the scene.

A one-way analysis of variance (ANOVA) was first used to determine the effect of saliency on the time elapsed prior to object fixation. Visual saliency was a reliable main effect, *F*(1,15) = 15.3, *p*<0.01, with the highly salient object being fixated earlier (after 1.67 sec) than the less salient object (2.32 sec).

Two further ANOVAs were used to determine the effects of the semantic congruency of the two objects. Each of these ANOVAs had two factors, the congruency of the object itself, and the congruency of the other object. For the highly salient object, there was no effect of congruency, *F*<1, no effect of the congruency of the other object, and no interaction, *F*(1,15) = 2.92. The second ANOVA indicated that the object that had low visual saliency was fixated earlier when it was incongruous (1.96 sec) than when it was congruous (2.67 sec), *F*(1,15) = 7.63, *p*<0.05. The congruency of the more salient object also had an effect on the time taken to fixate the inconspicuous object, *F*(1,15) = 6.73, *p*<0.05, with earlier fixation when the more salient object was incongruous (2.07 sec) rather than congruous (2.56 sec). There was no interaction, *F*<1.

High visual saliency was associated with early fixation of the object, and this attraction was resistant to any influences the semantic incongruency. In contrast, inconspicuous objects that were incongruent with the scene, such as a stapler in a kitchen, attracted their first fixation earlier than when they appeared in a congruent setting such as on a desktop.

Number of fixations prior to the first fixation on an object

The number of fixations between the onset of the display and the fixation of an object is a second indicator of how effective that object is in attracting attention. As the fixation position at picture onset, before the first saccade, was necessarily in the centre, the earliest an object could be fixated was on the second fixation, after one saccadic movement.

The same three ANOVA designs were used here as in the previously. The first analysis was a significant effect of visual saliency, *F*(1,15) = 23.24, *p*<0.001, with fewer fixations before inspection of the visually salient object (5.48 vs. 7.73 fixations). The analysis of the fixations leading up to fixation of the most salient object showed no effect of its congruency, *F*<1, no effect of the congruency of the inconspicuous object, *F*<1, and no interaction, *F*(1,15) = 3.61. The inconspicuous object itself was inspected earlier when it was incongruous (6.5 fixations) rather than congruous (8.9), *F*(1,15) = 8.21, *p*<0.05. Inspection of this object was made following fewer fixations when the more salient object was congruous (6.8 vs. 8.6 fixations), *F*(1,15) = 9.30, *p*<0.05. There was no interaction, *F*<1.

The pattern of results when using this measure of the number of fixations prior to an object's first fixation is identical to that obtained when time to first fixation is used. Visually salient objects attracted attention early and were resistant to the effects of semantic congruency. When an inconspicuous object was incongruent it was fixated earlier than when it was congruent with the scene.

First gaze duration

The duration of the first gaze on an object provides a measure of the difficulty of processing. Gaze was defined here as the total duration of all consecutive fixations on an object before fixating elsewhere, and indicates the total visual attention given to an object on its first inspection.

The same three ANOVAs were used here as previously. There was no effect of visual saliency on the duration of the first gaze, $F(1,15) = 1.66$. The second analysis, of gazes on the more salient object, indicated that there was an effect of the semantic congruency, $F(1,15) = 9.13$, $p<0.01$, with longer gazes on incongruent objects (448 msec vs. 366 msec). The congruency of the other object had no effect on the gaze on the conspicuous object, $F(1,15) = 3.95$, and there was no interaction, $F(1,15) = 4.39$. The analysis of gazes on the less salient object also found an effect of congruency, $F(1,15) = 13.42$, $p<0.01$, again with longer gazes on incongruous objects (538 msec vs. 361 msec). The congruency of the other object had no effect, $F<1$, and there was no interaction, $F<1$.

Gaze durations were influenced more by semantic congruency than by visual saliency. There was no difference in the duration of the first gazes on the two objects. Incongruous objects attracted longer gazes than their congruous equivalents, whether they were conspicuous or not.

Summary of the results of Experiment 1

When pictures were inspected for a memory test, the visually salient object was fixated earlier, and after fewer previous fixations, than the less salient object. This provides evidence in support of the Itti and Koch (2000) saliency model, in which visual attention is predicted to move round a picture in response to the low-level prominence of regions. Once acquired, the visually salient object was fixated for no longer than the less visually salient object.

The time and number of fixations before fixating the more salient object was not influenced by its semantic consistency. However, the less salient object was fixated earlier when it was incongruous than when it was consistent with its setting. The first gaze on an object was longer when it was incongruous, regardless of visual conspicuity. These results, particularly the time and fixations prior to inspection of the object, support the notion that a violation of the gist of a scene is detected early during inspection, but only for objects that do not stand out visually. Conspicuous objects were resistant to the effects of congruency until they were fixated, and at this point inspection was prolonged. This result argues against the hypothesis that the incongruency effect is apparent only when the object is visually conspicuous. The opposite result was seen here, with the congruency of a conspicuous object not influencing its early fixation.

The implications of these results will be discussed after first describing the use of the same pictures in a task requiring the viewers to determine whether or not a small target object was present in the scene. The same measures were taken in the two experiments.

EXPERIMENT 2

Focused search for a target

In this experiment the same photographs were used, but the effect of a change of task was explored. Experiment 1 had viewers look at pictures in preparation for a memory test, and in doing so they needed to look at the whole picture without looking at any specific object at the neglect of any other, and in Experiment 2 the viewers looked for the presence of a small target object in a directed search task. Henderson et al. (1999) used both of these tasks in their investigation of congruency in picture perception, and Underwood et al. (2006) found a variation in the influence of saliency between a memory task and a search task. Experiment 2 also used a search task, to allow comparison of the effects of saliency and congruency with the memory task of Experiment 1. The search task allowed us to ask whether viewers would ignore salient objects in order to find a target efficiently. When searching for occluded keys on an untidy desktop, are eyes are not necessarily drawn to the brightest or highest contrast object in view - a brightly coloured mouse mat, perhaps, or a black stapler resting on a blank sheet of white paper. Saliency can be overridden by cognitive need, and we are able to search selectively using the characteristics of the target object. Accordingly, we would expect visual saliency to have less effect in this search task than in the memory task of Experiment 1.

Method

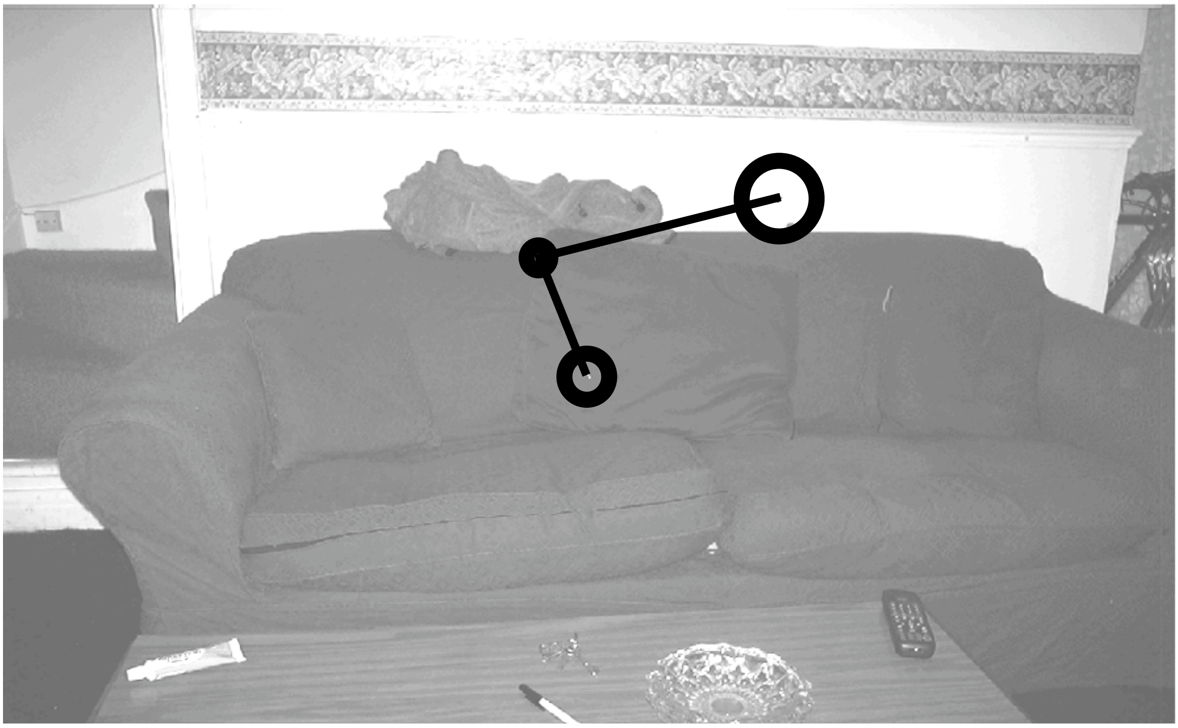
Participants

Eighteen students (with one replacement) with normal or corrected-to-normal vision who had not taken part in Experiment 1 volunteered and gave their informed consent. Their ages ranged between 18 and 25 years.

Stimuli and Design

These were exactly the same as in Experiment 1. In half of the photographs a small grey rubber ball was placed somewhere in the scene. This was the target stimulus and was typically placed somewhere near the edges of the photograph, away from both of the critical objects. It was never occluded but it had lower luminance than its surroundings, and it was of very low visual saliency. The target can be seen in Figure 1 (resting on the back of the sofa, to the right of centre), and the predictions of the saliency algorithm indicate that it is not a significant area of complexity in the image. The target was equally likely to appear in each of the four saliency/congruency conditions. Six practice pictures were also prepared and used to confirm that the participants knew what the task entailed.

Figure 3. Fixations made by one participant in Experiment 2 whilst viewing the same stimulus depicted in Figures 1 and 2. The first fixation was in the centre of the picture. In this example the target was fixated on the third fixation, and the response indicating that the target had been found was made during this fixation. This response acted to terminate the display.



Apparatus and Procedure

The equipment was the same as that in Experiment 1. After calibration, an instruction screen showed a picture of the target stimulus and instructed participants to search for the target as quickly as possible. A practice session familiarised the participants with what the target looked like in a scene and confirmed they had understood the task. Participants were instructed to press a computer keyboard keys marked "*Yes*" and "*No*" key to indicate the presence or absence of the target and to respond as quickly and accurately as possible. The experimental stimuli were then presented in an order randomised for each participant, with each picture displayed until a response was made.

Table 2. The means (and standard deviations) of the overall inspection time (sec) from Experiment 2, in which the picture was searched for a target object that was present or absent.

High Visual Saliency Object:	**Congruent**		**Incongruent**	
Low Visual Saliency Object:	**Congruent**	**Incongruent**	**Congruent**	**Incongruent**
Target present	1.35 (0.71)	1.37 (0.81)	1.28 (1.16)	1.71 (0.86)
Target absent	3.85 (3.35)	3.51 (3.05)	3.17 (3.06)	3.46 (3.32)

Results and Discussion

Accuracy on this task was very high, with many participants responding to all pictures correctly. The data from those rare trials that were responded to incorrectly were not included in the following analyses. As in Experiment 1, any trials that did not begin with an initial fixation within one degree of the centre were also excluded. Figure 3 shows a typical search path from a trial in which a target was present.

Total inspection duration

Table 2 presents the inspection times for pictures with and without a target. This is the period between picture onset and the response to indicate the search decision, at which point the display was terminated.

A three-factor ANOVA was first performed, using data from all trials with target (present/absent), congruency of the high saliency object (high/low congruency), congruency of the high saliency object (high/low congruency), and congruency of the low saliency object (high/low congruency) as the factors. Responses were faster when the target was present (1.43 sec) rather than absent (3.50 sec), $F(1,17) = 12.76$, $p<0.01$, and responses were faster when the high saliency object was incongruent (2.34 sec) rather than congruent (2.59 sec), $F(1,17) = 13.11$, $p<0.01$. No other effects were reliable.

The remaining comparisons involved the principal objects of interest, and these objects were fixated on less than 20% of trials when a target was present. Participants often moved their eyes to the target within the first two or three fixations, without fixating either of the objects of interest, and then responded to terminate the display. The remaining analyses use the data from those trials where a target was not present in the picture. On those trials an exhaustive search of potential targets was necessary, and that is indicated by the longer search times. A consequence of the exhaustive search is that the principal objects were fixated on most of the trials. These data are presented in Table 3.

Time prior to the first fixation on an object

As in the analyses of the measures taken in Experiment 1, three ANOVAs were performed on the duration of the interval between onset of the display and the first fixation on an object. The first analysis inspected the factor of saliency, finding no difference in the time taken to first fixate the conspicuous and inconspicuous objects, $F(1,17) = 1.18$, in contrast to the result from Experiment 1.

ANOVAs were conducted on the time taken to first fixate each object, as a function of the congruency of the two objects. For the more salient object, there was no effect of its semantic congruency, $F<1$. There was an effect of the congruency of the other object, $F(1,17) = 6.99$, $p<0.05$, with earlier fixation of the conspicuous object when the less salient object was incongruous (0.92 sec) than when it was congruous (1.23 sec). The incongruency of the two objects interacted, $F(1,17) = 5.64$, $p<0.05$, and an analysis of simple main effects indicated an effect when the conspicuous object was congruous, $F(1,17) = 15.27$, $p<0.01$. For congruous conspicuous objects only, the time prior to its fixation was shortened when the other object was incongruent. There was no effect of the less salient object on an incongruous conspicuous object, $F<1$.

For fixations on the less salient object, there were no effects of congruency. There was no effect of the congruency of the less salient object itself, $F<1$, no effect of the congruency of the conspicuous object, $F(1,17) = 1.08$, and no interaction, $F(1,17) = 1.97$.

The saliency of an object did not influence the time elapsed before the first fixation on that object, and the semantic congruency of the object also failed to influence its time to first fixation. The congruency of the other object has an effect, however, but only upon the time to fixate the conspicuous object, when a scene containing an inconsistent object resulted in quicker fixation.

Number of fixations prior to the first fixation on an object

Similar ANOVAs were performed on the number of fixations prior to inspection of each specific object as were performed on the measure of time to first fixation. There was no effect of visual saliency, $F(1,17) = 3.43$, as was the case with time to first fixation.

A two-factor ANOVA of the number of fixations prior to fixation of the conspicuous object revealed no effect of its congruency, $F<1$, but the congruency of the less salient object did have an effect, $F(1,17) = 4.71$, $p<0.05$. The presence of an incongruous inconspicuous object in the scene resulted in fewer fixations prior to fixation of the conspicuous object (3.67 vs. 4.64 fixations). There was also an interaction, $F(1,17) = 6.52$, $p<0.05$, and simple main effects indicated that the congruency of the less salient object influenced fixation of the more salient object only when that object was itself congruent, $F(1,17) = 12.14$, $p<0.01$, and not when it was incongruous, $F<1$. For congruous conspicuous objects, there was earlier fixation when it was accompanied by an incongruous than by a congruous less salient object. This is the same effect as was reported for the measure of time to first fixation.

There were no effects on the number of fixations prior to the first fixation of the less salient object. The semantic consistency of this object had no effect, $F<1$, the consistence of the more salient object had no effect, $F<1$, and there was no interaction, $F(1,17) = 2.42$. This is the same pattern as was reported when time was used as the measure.

First gaze duration

An object's visual saliency had no effect on the duration of the first gaze on the object, $F<1$. The duration of gaze on a visually salient object was influenced by the consistency of that object (congruent: 191 msec; incongruent: 220 msec), $F(1,17) = 6.37$, $p<0.05$, but not by the consistency of the less salient object, $F<1$. An interaction between the congruencies of the two objects, $F(1,17) = 18.93$, $p<0.001$ was inspected with an analysis of simple main effects. When the low saliency object was incongruous, gazes on the high saliency object varied according to its own congruency. There were longer gazes on conspicuous objects that were incongruous than upon those that were consistent with the scene, but only when the other object was congruous, $F(1,17) = 15.94$, $p<0.001$. When the low saliency object was incongruous, there was no effect of the congruency of the more salient object, $F<1$. For the less salient object, there was no effect of its congruency, $F<1$, or of the congruency of the more salient object, $F(1,17) = 1.45$, and no interaction, $F<1$.

Table 3. The means (and standard deviations) of the measures from Experiment 2, in which the picture was searched for a target object that was present or absent. These measures are from trials in which the target was absent.

High Visual Saliency Object:	**Congruent**		**Incongruent**	
Low Visual Saliency Object:	**Congruent**	**Incongruent**	**Congruent**	**Incongruent**
High Saliency Object				
Time prior to fixation (sec)	1.39 (0.65)	0.73 (0.38)	1.08 (0.61)	1.10 (0.45)
No of fixations prior to fixation	5.19 (2.4)	2.99 (1.3)	4.09 (1.6)	4.35 (1.9)
1st gaze duration (msec)	179 (41)	203 (46)	244 (49)	196 (35)
Low Saliency Object				
Time prior to fixation (sec)	1.12 (0.56)	0.76 (0.54)	0.99 (0.58)	1.24 (0.99)
No of fixations prior to fixation	4.28 (2.2)	2.78 (1.9)	3.56 (2.2)	4.54 (3.7)
1st gaze duration (msec)	203 (60.4)	222 (37.5)	227 (75.1)	225 (40.1)

Summary of the results of Experiment 2

When searching for a target, participants were able to guide attention efficiently and respond correctly, rarely fixating either of the two principal objects of interest to the questions about visual saliency and semantic congruency, except when no target was present. This meant that total inspection times were generally less than in Experiment 1, and when they did look at these objects it was with shorter gazes than in Experiment 1. Their potential as targets could be dismissed rapidly, and the search continued to other locations.

Trials where there was no target were analysed separately to further explore the influence of a strategy of searching for a well-defined target. In contrast to Experiment 1, where there was an effect of the more visually salient object to be fixated earlier, the saliency of the objects had no influence in the search task. The time elapsed prior to fixation of an object, and the number of fixation taken to fixate an object, did not differ as a function of visual conspicuity. Viewers were able to ignore visually prominent objects in their search for the target.

There was an unexpected interaction between saliency and incongruency in the search task, with both of the measures showing the same pattern. The congruency of the conspicuous object did not influence the delay in its fixation, but it was fixated earlier when the other object in the scene was incongruent. If the scene contained only objects that did not violate the gist, then it took the longest time to fixate the conspicuous object, and when it was fixated, it received the shortest gaze. This pattern is consistent with a highly salient object being of little interest to the task of searching for a small dark object, and so it was inspected late and only briefly. If the scene contained an inconspicuous object that was incongruent, then there was a shorter delay before fixating the salient object. This is an indication that the viewers had recognised that the scene contained a gist violation, and their response was to look at the most conspicuous object.

GENERAL DISCUSSION

In each experiment viewers inspected pictures of rooms and their eye movements were recorded. Objects that varied in visual saliency and in semantic congruency were placed in the scenes, and the measures allowed us to ask whether early eye fixations are attracted to objects that are visually conspicuous or that violate the gist of the picture by being out of place in that particular room. The two experiments varied the cognitive demands involved in picture perception, with a number of differences in the effects of conspicuity and congruency. The saliency map model of early visual attention gained good support with the recognition memory task, but not in the search experiment. Semantically incongruous objects did attract attention early in sequence of inspections, but this depended upon their visual conspicuity. Again, this effect of the early fixation of objects that violated the gist of the scene was evident only in the memory experiment. As in the picture-sentence verification task we have used previously, general encoding resulted in a wide distribution and increased number of fixations relative to a task in which it was possible to engage in a focused search for a specific target (Underwood et al., 2004). The shorter gazes in Experiment 2 confirm the focused nature of the search for a well-defined target object. The average gaze in the search task was about half the duration of the average gaze in the memory task.

Attention was attracted to visually salient and to semantically incongruent objects when pictures were first inspected, but only during the general encoding of the scene. When viewers inspected the pictures in preparation for a recognition memory test, their eyes moved earlier to a highly salient object than to a less salient object. This confirms the predictions of models of visual attention that suggest that viewers first determine the regions of low-level variation and build a saliency map that is used to direct the initial eye movements around the scene (Findlay and Walker, 1999; Henderson et al., 1999; Itti and Koch, 2000). It also confirms the results from the memory experiment reported by Underwood et al. (2006) in finding a relationship between conspicuity and the early fixation of an object. In Experiment 2 the viewers searched for a small target that was present in half the pictures, and here they were unaffected by conspicuity or congruency. The demands of the task allowed them to focus upon the search for a target, with cognitive override of low-level features either through avoidance of the process of building the saliency map or through disregard of the saliency peaks. This also confirms the result from the search task used by Underwood et al. (2006), in which conspicuity was also ineffective in guiding eye movements. The cognitive override of visual saliency is also a feature of Torralba's (2003) model of contextual cueing, in which the visual context becomes available sufficiently early to allow modulation of the saliency map. The bottom-up saliency map models do not give a good account of scene inspection when inspection can focus upon the detection of a specific and well-defined object.

The Findlay and Walker (1999) version of the model does acknowledge top-down influences on saccadic control, with viewers able to suppress saccades to prolong fixations, or to move their eyes voluntarily. This version of the saliency map model can account for the variations in the efficacy of visual conspicuity in the two tasks. In addition to recognising the influences of low-level visual features in the inspection of scenes that can account for the general encoding of the pictures in Experiment 1, their model also has a role for cognitive override and can account for the disregard of the saliency distributions when the same pictures were used in a search task in Experiment 2. Cognitive control can operate by three processes in this account. A process of "spatial selection" can modify the saliency weightings to allow the fixation of visually inconspicuous regions or disregard a high saliency region, but if this process operated alone then viewers would move their eye over the scene without any guidance, and a random pattern of fixations might be recorded. The second cognitive intervention involves a process of "search selection" that promotes saccadic movements to visual features possessed by the target object, and this process can guide the search to candidate targets. The third process is described as “intrinsic saliency” and acknowledges the influence of the viewer’s knowledge of the scene. We will return to this third process when we consider the effects of gist violation. Each of Findlay and Walker’s cognitive override processes may have operated in Experiment 2, with “search selection” being particularly potent. Their account would rely upon the notion of search being actively driven by the detection and inspection of target-related features that plausibly account for the absence of saliency effects when viewers are instructed to search for a target with well-defined visual characteristics. In Experiment 2 viewers searched for a small dark ball, and as a large, bright and colourful object could not be a target it would be pointless to inspect it. The search would be more effectively directed to small perturbations on supporting surfaces, as these are the candidate locations for the target object. Knowledge of where the target ball might have been placed – the intrinsic saliency of supporting surfaces – perhaps also helped to guide the search process.

Although Findlay and Walker (1999) define intrinsic saliency as “visual contours and high-contrast areas of the visual field” (p. 664) – akin to the Itti and Koch (2000) definition of saliency – they also suggest that “long- and medium-term learning and adaptive processes may also modify the salience of visual information”. They cite the way in which learned orthographic sequences can modify saccadic movements during reading as an example. Readers learn these sequences, and when they detect an unusual combination of letters their eyes move towards it. In a similar way an object that is unusual in a particular context might attract fixations, by virtue of a violation of a learned association between the object and the other objects that together contribute to the gist of the scene.

The initial low-level saliency map model is too simple, and a subsequent modification has been proposed to take task demands into account (Navalpakkam and Itti, 2005). This version of the model can search for specific objects on the basis of low-level features that have been learned, biasing attention towards objects that share these features. The saliency map that is built is then able to represent task-relevant features and guide eye movements towards potential targets. Motion is also used to modify the saliency peaks in this version of the model, with highly conspicuous but task-irrelevant features having reduced saliency. These representations are re-described as “task-relevance maps”, rather than saliency maps, and this version of the model gives an account for why attention is not captured by a bright sky as we drive around a corner, as it would do if a simple saliency peak were able to override task concerns. The motion of other objects is highly pertinent to a driver, of course, and by giving the model the ability to learn the features of task-relevant objects it is able to account for why a driver’s attention would be captured by another road user whose motion indicates an intersecting trajectory (e.g., a cyclist emerging from a side road into the driver’s path), but not by on-coming road users who are passing without incident (Chapman & Underwood, 1998; Underwood, Chapman, Berger & Crundall, 2003). This development of the saliency map model into a task-relevance map model would account for the focused pattern of fixations and neglect of visual saliency peaks observed in the search experiment.

The Itti and Koch (2000) algorithm classified the two objects of principal interest as being of high or low visual saliency, and the objects were also classified as being congruent or incongruent with the gist of the scene. Independent judges viewed the pictures prior to the experiments, and classified the consistency of each object within its setting. Previous experiments have reported an unreliable effect of object congruency, and one explanation of this failure to find a robust effect rests with the visual conspicuity of the objects. Loftus and Mackworth (1978) found earlier fixation of an object that was incongruous in a scene, suggesting that the early recognition of gist, and the detection of an object that violates the gist, can influence the early direction of saccadic eye movements. They reported that an incongruous object was fixated as early as the second fixation on a line-drawing. In contrast, the studies reported by De Graef et al. (1990) and Henderson et al. (1999) found no effect of incongruency upon the time taken to first fixate an incongruous object. All of these experiments used line-drawings rather than photographs of actual scenes, and there is a striking difference in the detail in the drawings used in their experiments. The incongruous object in the example presented by Loftus and Mackworth was drawn to be isolated from the other objects that comprised the scene – there was no overlap of features between the object and its background – whereas the drawings used by De Graef et al. and by Henderson et al. had much richer detail and extensive overlap and occlusion of objects. It is possible that Loftus and Mackworth found an effect of incongruency because the object was conspicuous, whereas in the other experiments with line-drawings the object did not have any “pop-out” quality. In the experiments where there was no effect of incongruency the violating object was identifiable, but its detection required a focused search around its location in the scene. The possibility investigated by combinations of salient and incongruent objects in our experiments here is that the incongruency effect only appears when the object is highlighted by being visual conspicuous.

There were a number of effects of object congruency in Experiment 1, and these were mainly moderated by the visual saliency of the object being inspected, but not in the direction that was expected. The inspection of a highly salient object was unaffected by whether or not it was congruent with the scene, while the less salient object was fixated earlier when it was incongruent. In neither experiment was there any evidence of a congruency effect – the early fixation of objects that violated the gist of the scene – that relied upon an incongruent object being visually conspicuous. When an object was conspicuous it did attract early fixations in the memory experiment, but this was independent of its congruency. Conspicuous objects attracted attention early, and this effect was robust to any effect of scene violation.

If the appearance of an congruency effect had depended upon the object being visually conspicuous, then it should have been the visually salient object that gained from violating the gist of the scene here, whereas earlier fixation was seen for the less salient object. Both the time elapsed and the number of fixations made prior to fixating an object were reduced if the object was incongruent, but this held as a robust effect only for less salient objects. It is important to note that the congruency effect did not influence object fixation immediately. The picture had been scanned for two or three seconds by the time an incongruent object was fixated, providing an opportunity for the scene gist to be developed and for a violating object to be detected with parafoveal vision. A version of Findlay and Walker's (1999) principle of intrinsic saliency, by which inspection of a scene can be guided by the viewer's knowledge, could be the process that detects gist violations. They presented the idea of intrinsic saliency as a predominantly visual form of control in which contours and contrast act to influence saccadic programming, but the moderation by learning could provide the basis for an explanation of incongruency effects. The viewer's expectations about what objects could appear in a scene depicting a specific room would be disrupted by the appearance of an object from a different room, and this disruption may attract attention, and also prolong the gaze on the object when the it is finally fixated. We suggest that intrinsic saliency could include a process in which the early understanding gist generates expectations about what component objects might be present (Underwood, 2005). When an object is detected that violates the gist, then attention is drawn to it to allow confirmation by close inspection, and integration of the object in its unfamiliar context.

Congruency also had an effect in Experiment 2, but only as an interaction between two objects. When both of the two objects were consistent, fixation of the more salient of them was slow, presumably because inspection of a bright object had low priority when the task required the detection of a small dark target. When the inconspicuous object violated the gist of the scene then fixation of the salient object was sooner than when it was consistent with the scene. The incongruency of the less salient object interfered with the search path, prompting earlier fixation of an object, and the object that was then inspected was the more conspicuous of them. When it was fixated, the duration of the gaze was no different whether the other object was congruous or incongruous.

The two experiments reported here offer qualified support for models of eye guidance that suggest that low-level visual features determine the first few fixations on a picture. The high visual saliency of an object was associated with its earlier fixation than an object of lower saliency, but this relationship held only in a task requiring the general encoding of the whole photograph. When viewers searched the photograph for the presence of a specific target object, the saliency of an object did not predict its fixation. This cognitive override of visual saliency requires modification of the simple conspicuity model, and supports guidance models that suggest that the saliency map is modified according to task demands. The incongruency of the objects in the scene influenced fixation behaviour in both tasks. The experiments were designed to test the hypothesis that inconsistent reports of the incongruency effect have resulted from previous studies not controlling the conspicuity of the incongruous object. The earlier fixation of incongruous objects may have been associated with their greater conspicuity, but this hypothesis was not supported. Indeed, a conspicuous object attracted earlier fixation independently of its congruency.

REFERENCES

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

Chapman, P. & Underwood, G. (1998). Visual search of driving situations. *Perception, 27,* 951-964.

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 effect of scene context on object identification. *Psychological Research, 52,* 317-329.

Findlay, J. M. & Walker, R. (1999). A model of saccade generation base on parallel processing and competitive inhibition. *Behavioral and Brain Sciences, 4*, 661-721.

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.

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

Itti, L. (2005). Quantifying the contribution of low-level saliency to human movements in dynamic scenes. *Visual Cognition,* in press.

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

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219-227.

Lamy, D., Leber, A., & Egeth, H. E. (2004). Effects of task relevance and stimulus-driven salience in feature-search mode. . *Journal of Experimental Psychology: Human Perception and Performance, 30,* 1019-1031.

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.

Navalpakkam, V. & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45,* 205-231.

Nothdurft, H.-C. (2002). Attention shifts to salient targets. *Vision Research, 42,* 1287-1306.

Parkhurst, D., Law, K. & Niebur, E. (2002). Modelling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107-123.

Potter, M. C., Staub, A., Rado J., & O'Connor D. H. (2002). Recognition memory for briefly presented pictures: The time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 1163-1175.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research, 42,* 1447-1463.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America, 20,* 1407-1418.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Underwood, G. (2005). Eye fixations on pictures of natural scenes: Getting the gist and identifying the components. In G. Underwood (ed.), *Cognitive Processes in Eye Guidance* (Oxford; Oxford University Press), pp. 163-187.

Underwood, G., Chapman, P., Berger, Z., & Crundall, D. (2003). Driving experience, attentional focusing, and the recall of recently inspected events. *Transportation Research Part F: Traffic Psychology and Behaviour, 6,* 289-304.

Underwood, G., Foulsham, T., van Loon, E., Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology,* in press.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology, 57A,* 165-182.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research, 44,* 1411-1426.

Appendix 1

The scenes used in the four conditions of the experiment, with the two objects of interest in each scene (high/low saliency). Different views were used within each scene.

High saliency object congruent, low saliency object incongruent:
Bathroom (razor/toothpaste), kitchen (spoon/spatula), kitchen (whisk/fork), kitchen (spoon/fork), lounge (TV controller/videotape cassette), lounge (video cassette/TV controller), office desk (stapler/pen), office desk (pen/stapler).

High saliency object congruent, low saliency object incongruent:
Bathroom (razor/spoon), bathroom (razor/fork), bathroom (toothpaste/video cassette), kitchen (spoon/bathplug), kitchen (spatula/ruler), lounge (videotape cassette/spatula), office desk (stapler/toothbrush), office desk (scissors/razor).

High saliency object incongruent, low saliency object congruent:
Bathroom (TV controller/bathplug), bathroom (scissors/toothpaste), kitchen (videotape cassette/spoon), kitchen (TV controller/spatula), kitchen (scissors/toothpaste), lounge (spoon/TV controller), lounge (toothpaste/TV controller), lounge (spatula/TV controller).

High saliency object incongruent, low saliency object incongruent:
Bathroom (pen/fork), bathroom (spoon/whisk), bathroom (stapler/whisk), kitchen (pen/stapler), lounge (spatula/fork), lounge (bathplug/spoon), office desk (bathplug/fork), office desk (bathplug/spoon).