Combining top-down processes to guide eye movements during real-world scene search

George L. Malcolm School of Informatics, University of Edinburgh, UK, & Psychology Department, University of Edinburgh, UK

John M. Henderson Psychology Department, University of Edinburgh, UK

Eye movements can be guided by various types of information in real-world scenes. Here we investigated how the visual system combines multiple types of top-down information to facilitate search. We manipulated independently the specificity of the search target template and the usefulness of contextual constraint in an object search task. An eye tracker was used to segment search time into three behaviorally defined epochs so that influences on specific search processes could be identified. The results support previous studies indicating that the availability of either a specific target template or scene context facilitates search. The results also show that target template and contextual constraints combine additively in facilitating search. The results extend recent eye guidance models by suggesting the manner in which our visual system utilizes multiple types of top-down information.

Keywords: visual cognition, search, eye movements, scene recognition

Citation: Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision, 10*(2):4, 1–11, http://journalofvision.org/10/2/4/, doi:10.1167/10.2.4.

Introduction

Successful search for a target object in a real-world scene involves selecting and extracting relevant information from a noisy environment. This task invokes eye movements that direct the region of highest resolution on the retina, the fovea, toward an informative part of the external environment. The fovea makes up only a tiny portion of the visual field, so in order to direct eye movements efficiently our visual system must integrate low-resolution information in the visual periphery with our knowledge about the current task and environment. A fundamental goal in the study of visual search in scene perception, therefore, is to understand how the visual system integrates available information to guide eye movements optimally.

There are three main sources of guidance information available in the percept of a novel, real-world image that the visual system can utilize: low-level saliency, target template information, and scene context. When the visual system uses low-level saliency, it selects regions of contrast in the image—usually on the basis of color, luminance, and intensity (Itti & Koch, 2000, 2001; Koch & Ullman, 1985). This is a purely bottom-up form of guidance, independent of the current task. When the visual system utilizes a target template, it matches a representation of the target stored in visual working memory against the scene. Regions in the scene containing features correlating highly with the template are then selected for fixation (Malcolm & Henderson, 2009; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Zelinsky, 2008). When the visual system uses scene context to guide eye movements, high-level knowledge is used to identify global regions in the scene that have a high probability of containing the target object (Castelhano & Henderson, 2007; Eckstein, Drescher, & Shimozaki, 2006; Neider & Zelinsky, 2006; Torralba, Oliva, Castelhano, & Henderson, 2006).

In the real world, when some or all of these information types are simultaneously available, an optimal system would combine them to guide visual search. Recent computational models suggest that model-generated fixation positions bear a closer resemblance to those of humans when they include more than one type of guidance information (Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Kanan, Tong, Zhang, & Cottrell, 2009; Torralba et al., 2006). For example, Torralba et al.'s (2006) *contextual guidance model* combines low-level salience and scene context when guiding search. Areas of high salience within a selected global region are given higher weights on an activation map than those that fall outside of the selected global region. These highly activated regions are then selected for fixation. The contextual guidance model outperformed a purely salience-driven model in predicting human fixation locations in a search task. In particular, the contextual guidance model predicted the location of the first few human fixations with a high degree of accuracy. This increased accuracy in predicting human fixation positions when two information types were combined suggests that humans similarly benefit from using more than one information source to guide search. This research has since been updated (Ehinger et al., 2009). Separate models were generated using each of the above three information types and were again assessed by comparing their predicted fixation locations with those of humans. Ehinger et al. found a high degree of agreement in fixation locations among participants, and while models using just one information type

doi: 10.1167/10.2.4 Received August 12, 2009; published February 10, 2010 ISSN 1534-7362 © ARVO

outperformed a control, models that combined all three information types accounted for the greatest percentage of human fixation locations.

Similar to the contextual guidance model, Kanan et al.'s (2009) Saliency Using Natural statistics (SUN) model combines top-down and bottom-up information to guide eye movements during real-world image search tasks. However, unlike the contextual guidance model, SUN implements target features as the top-down component. SUN outperformed a salience-driven model in predicting human fixation positions during real-world image search. The SUN model also slightly outperformed the contextual guidance model, though performance was similar overall. The important point is that both models found that combining two sources of guidance significantly improved their abilities to predict human fixation locations, suggesting that humans similarly combine information types to guide search.

Collectively, these results indicate that the visual system does not restrict itself to a single source of information to guide search. However, there are some limitations to what these studies can tell us. First, it is not known if combining different types of information actually facilitates search. When the above models combine multiple information types they produce fixation locations that are more highly correlated with human fixation locations. However, correlation with human data does not establish causality. The previous findings suggest that search should be affected by manipulations of the availability of different information types. In the present study, we directly manipulated the specificity of the search template and the usefulness of scene context and see if human search patterns are affected.

Second, if the availability of multiple types of information does facilitate search, it is not known how the visual system utilizes the extra information. Do these sources of information affect the same search behaviors or different ones? For instance, both scene context (Castelhano & Henderson, 2007; Eckstein et al., 2006; Neider & Zelinsky, 2006) and target template information (Malcolm & Henderson, 2009; Rao et al., 2002; Zelinsky, 2008) facilitate eye guidance. When both types of information are available in a search task is the benefit on eye guidance additive or superadditive? Does the system use each information type to facilitate the same or a different search process? For example, if scene context improves eye guidance to potentially relevant scene regions, might the system then use the target template to facilitate a different search process, such as determining if the target is present at each fixation? In the present study we use eye tracking to divide search time into three behaviorally defined epochs. These epochs allowed us to investigate which behaviors are affected by the presence of extra guidance information.

In the present study we manipulated two top-down information types: target template and scene context. Previous studies have focused on the relationship between salience and scene context (Torralba et al., 2006), or salience and target template information (Kanan et al., 2009), or all three information types at once (Ehinger et al., 2009), but have not focused specifically on the target template–scene context relationship. Understanding this relationship is important: there is growing evidence that top-down information dominates real-world image search processes, such that the influence of low-level salience information on search guidance is minimal (Ehinger et al., 2009; Einhäuser, Rutishauser, & Koch, 2008; Einhäuser et al., 2007; Foulsham & Underwood, 2007; Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009; Tatler, Baddeley, & Gilchrist, 2005; Tatler & Vincent, 2009; Turano, Geruschat, & Baker, 2003; Zelinsky, 2008; Zelinsky, Zhang, Yu, Chen, & Samaras, 2006). Even the high intersubject agreeability of the first saccade—which had been thought to be a result of the visual system selecting low-level salient areas to fixate (Carmi & Itti, 2006; Parkhurst, Law, & Niebur, 2002)—has since been suggested to be derived from common high-level knowledge strategies (Tatler et al., 2005). Low-level salience's minimal influence on real-world image search is most likely due its independence from a task goal while the placement of eye movements is heavily task dependent (Castelhano, Mack, & Henderson, 2009; Foulsham & Underwood, 2007; Land & Hayhoe, 2001; Land, Mennie, & Rusted, 1999; Rothkopf, Ballard, & Hayhoe, 2007; Yarbus, 1967). For these reasons, we concentrate on top-down information types in the current study as they will reveal the most about the guidance of attention and eye movements in real-world search.

Present study

To begin answering the above questions, we ran a visual search experiment in which the target could either be cued by an abstract cue (a word) or a specific cue (an exact matching picture of the target), and the target appeared in either a high-probability region of a scene (scene regions where the target object was likely to be found: e.g., staplers on desks, ceiling fans on the ceiling, etc.) or a low-probability region (scene regions less likely to contain the target object: e.g., staplers on the ceiling, ceiling fans on desks, etc.). Picture cues supported creation of a specific target template that could guide the search process, whereas word cues supported a less specific and more abstract template that would provide less constraint on search guidance. If the target was positioned in a high-probability region, then scene context would provide information about where the target was located; if the target was positioned in a low-probability region, then scene context would not provide information about the target's location.

If integrating target features with scene context provides for a more efficient search strategy than using either source of information alone, then searches with a specific cue and a target located in a high-probability region should result in the fastest search times.

Search is also a process that can be broken down into sub-processes that change over the course of time, and analyses that only measure overall search time may miss comparatively more fine-grained behaviors. There is precedent in using eye tracking to divide overall search time into sub-epochs, each reflecting different underlying processes (Castelhano, Pollatsek, & Cave, 2008; Malcolm & Henderson, 2009). In the present study, we divided total trial duration (overall search time) into those epochs used in Malcolm and Henderson (2009): search initiation time, scanning time, and verification time. Each of these epochs reflects a separate hypothesized underlying search process.

Search initiation time is measured as the time from appearance of the search scene until the first saccade away from the initial fixation point (the initial saccade latency) and measures the time needed to begin search. Scanning time is defined as the time from the first saccade (the end of the search initiation epoch) to first fixation on the target object and represents the actual search process. Verification time measures the participant's gaze duration on the target object, reflecting the time needed to decide that the fixated object is actually the target. Total trial duration, the response time measure reported in most previous visual search studies, is equal to the sum of these three epochs (Figure 1).

Previous results suggest that cue type should not affect search initiation time but should affect scanning time, with more specific cues leading to shorter scanning epochs (Malcolm & Henderson, 2009). Similarly, previous research has shown that the first saccade in real-world scene search tasks tends to land nearer regions that are more likely to contain the target (Castelhano & Henderson, 2007; Eckstein et al., 2006; Neider & Zelinsky, 2006). If scene context benefits search by providing the participant with a probable region within which to search, we should see shorter scanning times when the target is in a high-probability region. In terms of verification time, past research suggests that a more specific target template should facilitate the verification process (Burgess, 1985; Castelhano et al., 2008; Greenhouse & Cohn, 1978; Judy, Kijewski, Fu, & Swensson, 1995; Malcolm & Henderson, 2009) and that objects located in a high-probability region of a scene should take less time to verify (Biederman, Mezzanotte, & Rabinowitz, 1982).

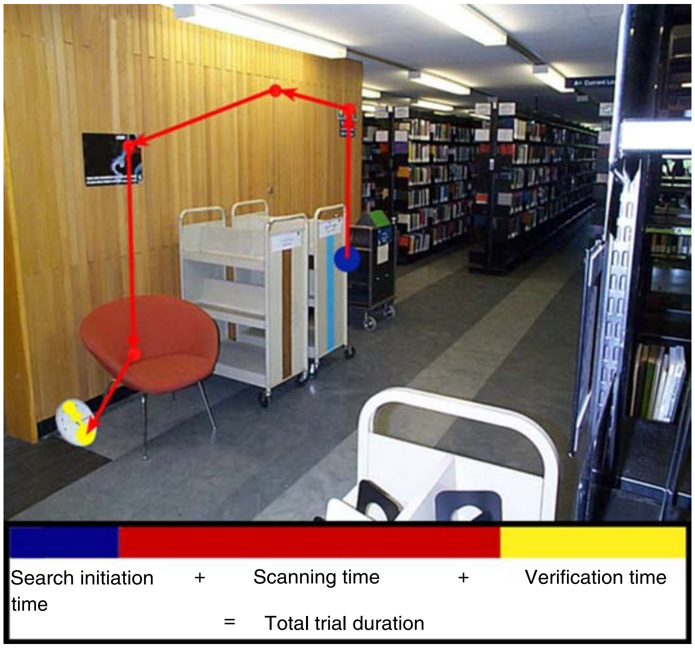

Figure 1. The total trial duration was segmented into three functionally defined epochs: Search initiation time (blue), which measured time to begin search; scanning time (red), measuring the time to locate the target, and verification time (yellow), measuring the time to accept the target. This trial used a target object (the clock) placed in a low-probability region of the scene (on the floor). The participant fixated a high-probability region of the scene first (the wall) before saccading toward the target in the low-probability region.

Methods

Participants

Twenty-four participants (sixteen females, ages 19–32, mean age 22.3) gave informed consent in accordance with the institutional review board of the University of Edinburgh. All participants were naive about the purpose of the study.

Apparatus

Eye movements were recorded using an EyeLink 1000. Experiments were programmed in Experiment Builder and analyzed in DataViewer (SR-Research, Mississauga, ON). Stimuli were shown on a 21″ ViewSonic G225f monitor positioned 90 cm away from the participant, taking up an 18.72° × 24.28° field of view, with a refresh rate of 140 Hz (ViewSonic, London, UK).

Stimulus materials

Fifty-two photographs of real-world scenes from a variety of categories (indoor and outdoor, natural and man-made) were used as stimuli. All images were scaled to 800 × 600 pixel resolutions and shown in full color. In each scene, two objects were taken from Hemera Images database (Hemera Technologies, Gatineau, Canada) and then modified and placed into the scene with Adobe Photoshop CS (Adobe, San Jose, CA). Each object appeared in only one scene.

In order to manipulate scene context, each scene had two targets assigned to it, only one of which appeared for a given participant during testing (see Figure 2). Both

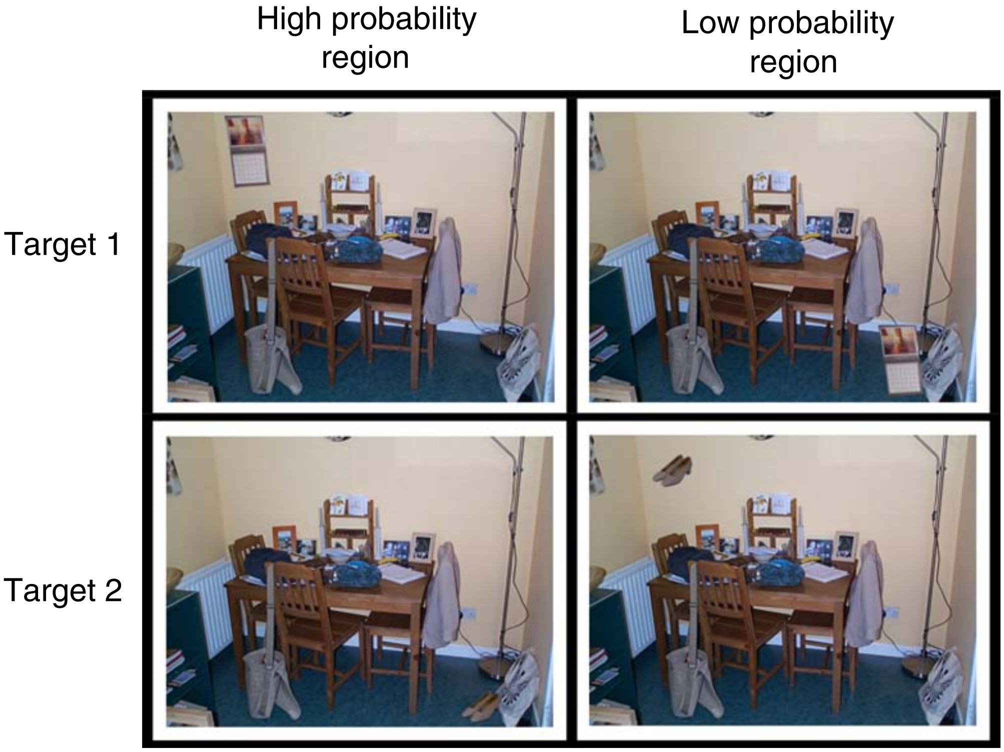

Figure 2. Example of how each scene had two possible targets (here, the calendar and the shoes) in two possible locations (here, on the wall or on the floor) to create four possibilities for each scene. Only one target ever appeared in a scene at a time. During the experiment, participants were shown either a word or picture cue of the target, followed by the scene with the target in either a high-probability or low-probability region. The cued target was always present.

objects were positioned so as to appear in high-probability regions (creating two possible scenes). Targets were then swapped so that they appeared in the reciprocal target's location so that they were now in low-probability regions (creating two more scenes).

Eight participants ranked the scenes to determine whether the targets were placed in high and low-probability regions as intended. None of these participants took part in the search experiment, and none of them had seen the scene images before. Participants were given a 7-point Likert-like scale and asked to evaluate whether the target was positioned where they would expect to find it in the given scene; 7 for yes, definitely; 1 for no, not at all. Participants were divided into two groups. Each group saw half of all the possible scenes, viewing each scene twice but with a different target each time. A participant never saw the same object at two different locations within the scene. Targets positioned in high-probability regions were judged to be in more expected regions of the scene than targets positioned in low-probability regions, $t(7) = 18.5$, $p < 0.001$.

To create the picture cues for the experiment, each target was pasted into a gray background, appearing exactly as it would in the scene. A further 52 corresponding word cues were created that contained only the name of the target object in 72 point font subtending 2.14 degrees in height, centered on a gray background.

Fifty-two further scenes were added to the experiment as fillers. Filler scenes used an existing object in the scene as the target. Targets in filler scenes were positioned in high-probability locations, meaning that 75% of all the scenes viewed by participants had target objects in high-probability regions. This percentage ensured that participants would recognize scene context as a potential source of guidance throughout the experiment. Half of the filler scenes were cued with words and half with pictures. The eye movements from the filler trials were not analyzed.

Procedure

Prior to the experiment each participant underwent the EyeLink calibration procedure: Eye positions were recorded as participants fixated a series of 9 dots arranged in a square grid extending to 19.25° eccentricity. Calibration was then validated against a second set of 9 dots.

For the experiment each participant began a trial by fixating a drift correction dot in the middle of the screen (for the purpose of eliminating any tracking error). The experimenter then initiated the trial. A central fixation cross appeared for 400 ms followed by a cue identifying the search target for 800 ms. The cue was either a word identifying the target or an exactly matching picture of the

	Word cue, HPR		Word cue, LPR		Picture cue, HPR		Picture cue, LPR	
	Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*
Accuracy	0.93	0.01	0.92	0.02	0.89	0.02	0.88	0.02
Total trial duration	1375.34	61.08	1765.19	75.41	1092.87	53.92	1351.87	80.12
Search initiation time	250.22	10.30	257.86	8.89	241.99	8.94	248.96	8.73
Scanning time	609.45	41.93	950.35	58.61	445.86	40.36	664.96	56.69
Verification time	515.53	39.02	556.89	33.30	405.22	26.36	437.95	32.82

Table 1. Results. All means are in milliseconds. HPR = target positioned in a high-probability region; LPR = target positioned in a low-probability region.

target. This was followed by a central fixation point lasting another 200 ms, making a stimulus onset asynchrony of 1000 ms. The search scene then appeared and participants searched for the target object, responding with a button press as soon as it was located.

There were four types of test scene: Target 1 in a high-probability region, Target 2 in a high-probability region, Target 1 in a low-probability region, Target 2 in a low-probability region. Participants only saw one of these four types of scene layout for each scene during an experiment, so each of the 52 scenes was seen only once. The four manipulations were rotated through scenes across participants in a Latin Square design. Test scenes and filler scenes were intermixed and occurred in a random order for each participant.

Results

Trials with errors were removed from analysis. Error trials were defined as those in which participants incorrectly identified the target; participants fixated the target, moved away, and returned before correctly identifying it; or the total trial duration exceeded 5500 ms. Under these criteria, 90.3% of the trials were scored as correct. If a participant fixated the target once, moved off the target for one fixation, and then returned to the target on the next fixation, this was accepted as a correct trial. Here, a single fixation deviating away from the target was considered to be the result of a pre-programmed oculomotor command and not due to a decision to attend to a different possible target. This scenario occurred on 4.7% of the correct trials.

Accuracy

A 2 × 2 repeated measure ANOVA showed a trend for a significant main effect of cue type on accuracy, with word cued trials producing marginally higher accuracy than picture cued trials, $F(1, 23) = 3.702$, $\text{MS}_{\text{E}} = 1.489$, $p = 0.067$. There was no main effect of scene context, nor was there a significant interaction between cue type and scene context, Fs < 1 (Table 1).

Analysis of total trial duration and the three search epochs

Additional 2 × 2 repeated-measures ANOVAs with cue type (word vs. picture) and target location (high-probability region vs. low-probability region) as factors were conducted on total trial duration, search initiation time, scanning time, and verification time (Table 1).

Both variables produced a main effect on total trial duration, with shorter total trial durations for picture cues and targets located in high-probability regions ($F(1, 23) = 45.925$, $\text{MS}_{\text{E}} = 63248.538$, $p < 0.001$, and $F(1, 23) = 36.587$, $\text{MS}_{\text{E}} = 69044.168$, $p < 0.001$, respectively). There was no significant interaction between cue type and target location, $F(1, 23) = 1.413$, $\text{MS}_{\text{E}} = 72678.914$, $p = 0.247$.

There was no main effect of either cue type or target location on search initiation time ($F(1, 23) = 1.559$, $\text{MS}_{\text{E}} = 1130.253$, $p = 0.224$, and $F(1, 23) = 2.031$, $\text{MS}_{\text{E}} = 630.416$, $p = 0.168$, respectively). There was also no significant interaction between cue type and scene context, $F(1, 23) < 1$.

The results of the scanning epoch mirrored those of the total trial duration: both cue type and target location produced a main effect on scanning time with picture cues and targets located in high-probability regions generating faster scanning times ($F(1, 23) = 24.576$, $\text{MS}_{\text{E}} = 49213.488$, $p < 0.001$, and $F(1, 23) = 28.844$, $\text{MS}_{\text{E}} = 65235.903$, $p < 0.001$, respectively). Again, there was no significant interaction between the two factors, $F(1, 23) = 1.509$, $\text{MS}_{\text{E}} = 58968.255$, $p = 0.232$.

The verification epoch showed the same pattern of results as the total trial duration and scanning epoch. There were main effects of both cue type and target location on verification time, with picture cues and targets located in high-probability regions producing faster verification times ($F(1, 23) = 52.641$, $\text{MS}_{\text{E}} = 5990.367$, $p < 0.001$, and $F(1, 23) = 8.212$, $\text{MS}_{\text{E}} = 4011.385$, $p = 0.009$, respectively). Again, there was no significant interaction between cue type and scene context, $F(1, 23) < 1$.

As stated above, we only analyzed correct trials; that is, trials in which the participant fixated the target in the allotted time and pressed the response button. Cases in which participants fixated the target and then saccaded away from it were considered errors and the trial was not analyzed, except when the participant saccaded away for exactly one fixation and then immediately returned with the next fixation. In this case the saccade away from the target was treated as a result of a pre-programmed oculomotor command and not due to a decision to reject the target and attend elsewhere. If this scenario occurred during one condition more than another, it may have biased the verification time analysis. However, when we reanalyzed the data we found that this was not the case.

The results indicate that both information types facilitate locating and verifying the target object but not initiating the search process. When scene context and a specific target template are both available, they facilitate search in an additive manner.

The underlying behavior affecting scanning time

The results suggest that both cue type and target location have similar effects on the scanning process during search. However, within the scanning epoch we can identify further sub-processes that the two variables may affect differently. Specifically, there are two decisions occurring during each fixation in the scanning epoch: the visual system must process the fixated object (making a decision about whether it is the target) and, when the object is rejected, must decide where to fixate next (Fischer, 1999; Henderson & Ferreira, 1990; van Diepen, Wampers, & d'Ydewalle, 1998; Zelinsky, 2008). In order to gauge whether target template and scene context affect these scanning sub-processes in similar or different ways, we compared fixation durations and the number of scene regions fixated as a function of the two variables.

Fixation durations reflect processing time in scene viewing and reading tasks (Henderson & Ferreira, 1990; Henderson & Pierce, 2008; Henderson & Smith, 2009; Rayner, 1998). In previous real-world scene search tasks, a more specific template led to shorter scanning fixations, meaning that fixated distracters were compared with the target template and rejected faster (Malcolm & Henderson, 2009). This was most likely due to a specific target template having more features to compare against the fixated region: if any one of those features mismatched the fixated region, a reject decision could be made quickly. In the current study, if having a more specific template allowed the visual system to process and make a quicker rejection decision at each fixation, then picture cue trials should produce shorter scanning fixations.

The number of scene regions fixated during scanning reflects how spatially selective or distributed search was across conditions. In order to measure spatial selectivity each scene was divided into 48 square regions of 100 × 100 pixels. The number of regions fixated in the scanning epoch of each trial was then counted. Given that we were examining how many regions were visited, but not how often, regions fixated more than once were still scored as one. If the visual system capitalizes on scene context about likely target locations, scanning eye movements should be more selective when the target is in an expected location, and fewer regions should be fixated. If the visual system uses a target template to select where to fixate in real-world images, then fewer regions should be visited in the picture cue condition (see Malcolm & Henderson, 2009).

It is also worth noting that scanning time showed an additive effect when both information types were available to the participant. There are three possible ways that both information types could affect the scanning sub-processes to produce this effect. First, target template and scene context information could affect both scanning sub-processes; we would therefore find that the availability of both information types produces an additive effect on both the regions visited and the fixation duration measures. Second, target template and scene context information could affect only one scanning sub-process; we would therefore find that the availability of both information types produces an additive effect on either the region visited or fixation duration measures. Third, the visual system may utilize one type of information to facilitate one scanning behavior, and the other type to affect the other scanning behavior; we would therefore find a main effect of one information type on one process, and the other information type on the other process (e.g., scene context affects the number of regions visited but has no affect on the fixation durations, whereas target template information has the opposite effect).

We found that picture cues produced shorter fixation durations than word cues during scanning, $F(1, 23) = 46.106$, $MS_E = 262.793$, $p < 0.001$, replicating Malcolm and Henderson (2009). Targets located in a high-probability region also produced shorter fixation durations during scanning, $F(1, 23) = 22.785$, $MS_E = 335.030$, $p < 0.001$. Finally, cue type and target location did not interact, $F(1, 23) < 1$ (Table 2). The visual system can process fixated regions faster when both information types are available, with both processes facilitating search in an additive manner.

Picture cues led to a more selective distribution of attention over the scene during scanning, with fewer scene regions fixated, $F(1, 23) = 8.015$, $MS_E = 0.280$, $p < 0.01$, replicating previous results (Malcolm & Henderson, 2009). Similarly, attention was more narrowly distributed when the target appeared in the high-probability compared to the low-probability region, $F(1, 23) = 27.616$, $MS_E = 0.533$, $p < 0.001$. Again, there was no interaction between cue type and location context, $F(1, 23) < 1$. The visual system is more selective in deciding where to fixate when both information types are available, with both processes facilitating this decision in an additive manner.

The results indicate that the additive effect found in scanning time when both information types were available

	Word cue, HPR		Word cue, LPR		Picture cue, HPR		Picture cue, LPR	
	Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*
Scanning time fixation duration	188.71	6.15	208.28	6.97	167.98	6.65	184.07	6.11
Scanning time regions visited	2.35	0.14	3.03	0.16	1.94	0.13	2.83	0.17
Verification time fixation duration	247.47	9.58	261.05	11.23	224.12	11.62	241.19	13.53
Verification time fixation count	2.06	0.11	1.96	0.07	1.89	0.11	1.92	0.14

Table 2. Results. All means are in milliseconds except for Regions Visited, which is measured in the number of regions visited on the display screen (maximum 48), and Fixation Count. HPR = target positioned in a high-probability region; LPR = target positioned in a low-probability region.

was a result of each scanning sub-process being affected in an additive manner.

The main effect of target location on fixation duration is surprising as it implies the speed with which distracters are rejected is dependent on the target's location, even before the target is located. A potential explanation might come from participants finding the target faster when it is in a high-probability region. Trials in this condition have fewer scanning fixations and average fixation durations are known to increase during a trial (Antes, 1974; Buswell, 1935; Castelhano et al., 2009; Tatler & Vincent, 2008; Unema, Pannasch, Joos, & Velichkovsky, 2005). When the target is located in a high-probability region, there will be fewer of these later, longer fixations (since the target is found faster). This could explain why the mean fixation duration was shorter.

If the effect of target location on fixation duration is solely a result of these later fixations, then there should be no difference between conditions if we compare only early fixations. The mean duration of these early fixations would not be influenced by the increased duration of later fixations. However, if another factor is causing the effect, then we should find a main effect of target location on scanning fixation durations even during these early fixations. In order to investigate this possibility, we focused our analysis on the first two scanning fixations. This corresponded to the mean fixation count in the shortest condition (trials with a picture cue and the target in high-probability location had a mean of 2.2 scanning fixations). We found that there was a main effect of cue with picture cues producing shorter fixations, $F(1, 23) = 38.291$, $\mathrm{MS_E} = 374.098$, $p < 0.001$, and a main effect of target location with targets in high-probability regions producing shorter fixations, $F(1, 23) = 17.774$, $\mathrm{MS_E} = 203.572$, $p < 0.001$. There was no interaction between the variables, $F(1, 23) = 1.001$, $\mathrm{MS_E} = 248.644$, $p = 0.327$. The results do not rule out the possibility that the number of fixations affects the mean fixation duration, but they do rule out the possibility that the number of fixations was the only cause of the main effect of target location. This suggests that there is another factor stemming from the target's location that influences fixation durations.

Another possible explanation for the influence of target location on scanning fixation duration relates to the processes occurring during scanning fixations: The fixated region is evaluated and the subsequent saccade is planned (van Diepen et al., 1998). Given that the mean scanning fixation count was 2.2 in the picture cue/high-probability region condition, it is possible that in some of these trials the target was identified in the periphery. This would happen more often in conditions where the target was positioned in highly probable region since, in these cases, initial saccades would probably have been directed toward the appropriate location. In these trials, the verification of the target object may have actually begun prior to the target being fixated. Thus, the final scanning fixation in this particular condition may not reflect typical scanning processes. In order to ensure that our analysis of the mean scanning fixation duration did not include any of the verification process, we reanalyzed the mean durations across trial types, but now without including the final scanning fixation.

When we analyzed the mean scanning fixation duration excluding the final scanning fixation, we found a main effect of cue with picture cues producing shorter fixations, $F(1, 23) = 11.702$, $\mathrm{MS_E} = 521.677$, $p = 0.002$, and a main effect of target location with targets in high-probability regions producing shorter fixations, $F(1, 23) = 5.352$, $\mathrm{MS_E} = 815.871$, $p = 0.030$. There was no interaction between the variables, $F(1, 23) < 1$. These results indicate that the shorter scanning fixations found in trials with the target positioned in a high-probability region were not a result of participants identifying the target in the visual periphery just prior to landing on it. Of course, it could be that in some cases the target was identified two fixations or more prior to its direct fixation, but given how quickly the targets were often found in the present study, it is not possible to test this hypothesis with the current data.

The underlying behavior affecting verification time

Similar to the scanning epoch, the verification epoch was shorter following more specific cues and with targets located in high-probability regions. As with the scanning epoch, the verification epoch can be analyzed with more detailed measures, including fixation durations and fixation count. Both fixation duration and fixation count reflect how easily the participant can process the target object once fixated. We therefore examined verification fixation durations and verification fixation counts. For both analyses, the data included only those fixations that were on the target. Though we accepted trials in which the participant fixated the target, saccaded away for one fixation and then immediately returned as correct, the fixations occurring outside the target were not included in these analyses.

Verification fixation durations were shorter following picture cues, $F(1, 23) = 12.014$, $\text{MS}_\text{E} = 932.595$, $p = 0.002$, and with targets located in high-probability regions, $F(1, 23) = 8.103$, $\text{MS}_\text{E} = 695.498$, $p = 0.009$. There was no interaction between cue type and location context, $F(1, 23) < 1$ (Table 2).

Fixation count was not significantly affected by cue type, $F(1, 23) = 2.406$, $\text{MS}_\text{E} = 0.113$, $p = 0.135$, or target location, $F(1, 23) < 1$. There was no significant interaction between cue type and scene context, $F(1, 23) < 1$.

Discussion

In the present study, we asked whether the visual system can combine multiple sources of top-down information—scene context and target template information—within a single real-world scene search task and, if so, how these information sources are combined and which underlying processes they affect.

The results indicated that either information type, by itself, reduces total trial duration; when both forms of top-down information are available, total trial duration is reduced additively. Our results thus indicate that the visual system will actively use multiple sources of top-down information to facilitate a search process. The results also indicate that the visual system treats scene context and target template information independently.

The facilitation of scene search by contextual constraint and template specificity was the result of facilitation during two specific search epochs: scanning and verification. Search initiation time, unlike the other two epochs, was unaffected by either variable. When we further analyzed the scanning epoch by examining the durations and spatial distributions of fixations during scanning, we found that target template and scene context produced main effects on each measure. When both information types were available, they influenced both scanning measures additively. This result suggests that scene context and target template information facilitate similar processes during the scanning epoch. Finally, scene context and target template information also affect the same verification behaviors. Each information type affected verification fixation durations additively when both were available. Neither affected verification fixation count.

The similarity between the search behaviors affected by the two information types is surprising in one instance. Scanning fixation durations reflect the processing time needed to compare a fixated region to a target template and make a reject decision. This process is facilitated when a specific target template is available, presumably because the specific target template has visual features that can be quickly compared to the fixated region (Malcolm & Henderson, 2009). However, we found that this process was also facilitated by scene context. This is surprising as it implies the speed with which distracters are rejected is dependent on the target's location, *prior* to the target being fixated. This result seems at odds with the type of information scene context provides: scene context information only indicates which region of a scene a target should be located in but not the properties of the target. Since scene context does not generate any internal visual properties of the target, what then causes fixated distracters to be rejected faster when the target is located in a high-probability region? One possibility is that this effect is a consequence of participants finding targets faster when they are located in high-probability regions. Since early fixations are short and later fixations are longer (Antes, 1974; Buswell, 1935; Castelhano et al., 2009; Tatler & Vincent, 2008; Unema et al., 2005), and since there would be fewer of these later, longer fixations in the high-probability condition (since the target is found quicker), the average fixation duration would be shorter. However, when we focused our analysis on just the mean duration of the first two scanning fixations—which were not influenced by the increased duration of later fixations—we still found an effect of target location. A second possibility is that participants may sometimes identify the target prior to fixating it. This is particularly likely when the target was in a high-probability region since the initial saccade would be directed toward the appropriate region. In this condition, the last scanning fixation may not actually reflect the typical scanning processes. However, when we analyzed the mean scanning fixation durations again, but this time without including the final scanning fixation (which may have included some of the verification process), we still found an effect of target location. These two analyses do not entirely rule out the possibility that the effect of target location was due to later fixations or initiation of the verification process, but they do reduce their plausibility.

Combining information types in real-world visual search models and future questions

The results suggest that any real-world search model will not only have to identify potential sources of top-down information available in a scene, but also how the visual system combines these sources of information to guide fixation placement and duration.

Focusing on the fixation placement analysis, it is clear that both information types are utilized during the scanning epoch. Scene context provides global information, directing the eyes toward high-probability regions of the scene, while the target template distinguishes probable targets from distracters. A future question when examining fixation placement is whether multiple sources of information are used simultaneously during scanning (the visual system selects local features that match the target within a probable global region) or whether the visual system alternates between the two (the visual system uses scene context to select a probable region and then, once there, target template information to select a target-similar object). Scene context and target template information provide different sets of coordinates in suggesting where the target is likely to occur, implying functionally disparate goals in saccadic targeting. There is also some evidence to suggest that scene context may direct search initially without the use of a template (Castelhano & Henderson, 2007; Eckstein et al., 2006; Ehinger et al., 2009; but see Kanan et al., 2009). It could be the case that instead of the visual system using both information types concurrently to locate a target, the visual system utilizes scene context and target template information for different purposes. If the visual system does not rely on both information types equally throughout a search task, then identifying when and why the visual system relies on each information type during a task will help reveal the nature of the visual system's search strategy.

Previous analysis of saccade amplitudes, both in image viewing (Tatler & Vincent, 2008; Velichkovsky, Joos, Helmert, & Pannasch, 2005) and real-world viewing (Hayhoe, 2000, Hayhoe & Ballard, 2005; Land & Hayhoe, 2001; Land et al., 1999), suggests that saccades can be divided into those directed toward nearby and far-away targets. Considering that scene context directs the eyes toward selected global regions, it could be that context is used intermittently by the visual system for the purpose of selecting new scene regions to search. Between these saccades, the visual system would search for a target within a region: here scene context would be of minimal use while a target template would be very useful. In this way, the visual system can guide search optimally while alternating between the two information types. Whether the visual system utilizes scene context and target template information simultaneously or sequentially to guide search is, unfortunately, not evident in the current data since we used complex scenes with high-probability and low-probability regions that were not easily divisible by an obvious boundary. A saccade could move in the direction of a high-probability region, but land on a low-probability region, masking the goal of the saccade (e.g., if the participant were searching for a plate in a restaurant, they might saccade downward toward a table in the foreground, but land on the floor in a gap between two tables. Does that mean that the participant was directing their eyes toward the table and came up short, or that the participant was looking for the plate on the floor?). Issues like this make it difficult to determine which particular region a participant was intending to fixate in the current study. Regions must be easily defined if we are to separate fixations by region and begin gauging the visual system's underlying goals.

A second issue to be resolved in a future real-world search model is fixation durations. Most current attention models for real-world scenes concentrate solely on where the participant will fixate next. However, as demonstrated previously and supported here, fixation durations in scene perception tasks are not constant: weighing fixated regions by their relative durations changes the attention map of a participant dramatically (Henderson, 2003, 2007). The current data not only confirm that fixations differ over the course of a scene search task, but that they differ depending on what top-down information is available, with more information leading to shorter fixations. Future models will have to ascertain how these sources are added together to facilitate quick accept/reject decisions. For the reasons listed above, it is difficult in the current study to determine whether multiple sources of top-down information are combined at each fixation in the scanning epoch to reduce processing time, or if the visual system relies on only one source at a time to process a region, alternating which source to use throughout the epoch. In the verification epoch, however, evidence suggests that multiple information sources can be used simultaneously to reduce processing time. There were very few verification fixations (ranging from means of 1.9–2.1 across conditions), and target template and scene context had an additive effect on fixation durations. This suggests that the two sources simultaneously facilitate the process of evaluating the target, reducing verification time accordingly.

Conclusion

When available, the visual system actively uses multiple types of top-down information to facilitate search. This facilitation occurs during the scanning epoch, particularly in the processing of the fixated region and selecting the next fixation position. Both information sources similarly reduce verification time by reducing fixation durations. Each behavior was facilitated additively when both information types were available, suggesting that the two information types were treated independently by the visual system. Future studies will be needed to probe whether the visual system uses scene context and target template information simultaneously to direct saccades during the scanning epoch, or alternates between the two.

Acknowledgments

We thank members of University of Edinburgh's Visual Cognition Lab for their feedback on this research. We also thank two anonymous reviewers for their critical comments. This research was funded by ESRC grant RES-062-23-1092 to JMH.

Commercial relationships: none.
Corresponding author: George L. Malcolm.
Email: g.l.malcolm@sms.ed.ac.uk.
Address: Psychology Department, 7 George Square, Edinburgh, EH8 9JZ, Scotland.

References

Antes, J. R. (1974). Time course of picture viewing. *Journal of Experimental Psychology, 103,* 62–70. [PubMed]

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

Burgess, A. (1985). Visual signal detection: III. On Bayesian use of prior knowledge and cross correlation. *Journal of the Optical Society of America A, Optics and Image Science, 2,* 1498–1507. [PubMed]

Buswell, G. T. (1935). *How people look at pictures.* Chicago: University of Chicago Press.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46,* 4333–4345. [PubMed]

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

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences on eye movements during scene perception. *Journal of Vision, 9*(3):6, 1–15, http://journalofvision.org/9/3/6/, doi:10.1167/9.3.6. [PubMed] [Article]

Castelhano, M. S., Pollatsek, A., & Cave, K. R. (2008). Typicality aids search for an unspecified target, but only in identification and not in attentional guidance. *Psychonomic Bulletin & Review, 15,* 795–801. [PubMed] [Article]

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science, 17,* 973–980. [PubMed]

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17,* 945–978. [PubMed] [Article]

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1–19, http://journalofvision.org/8/2/2/, doi:10.1167/8.2.2. [PubMed] [Article]

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., et al. (2007). Human eye–head coordination in natural exploration. *Network: Computations in Neural Systems, 18,* 267–297. [PubMed]

Fischer, M. H. (1999). An Investigation of attention allocation during sequential eye movement tasks. *Quarterly Journal of Experimental Psychology, 52,* 649–677. [PubMed]

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual saliency in scene perception? *Perception, 36,* 1123–1138. [PubMed]

Greenhouse, D. S., & Cohn, T. E. (1978). Effect of chromatic uncertainty on detectability of a visual stimulus. *Journal of the Optical Society of America, 68,* 266–267. [PubMed]

Hayhoe, M. M. (2000). Vision using routines: A functional account of vision. *Visual Cognition, 7,* 43–64.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9,* 188–194. [PubMed]

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7,* 498–504. [PubMed]

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16,* 219–222.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements:* A window on mind and brain (pp. 537–562). Oxford, UK: Elsevier.

Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 417–429. [PubMed]

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance versus visual salience during search for non-salient objects in real-world scenes. *Psychonomic Bulletin & Review, 16,* 850–856. [PubMed]

Henderson, J. M., & Pierce, G. L. (2008). Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin & Review, 15,* 566–573. [PubMed] [Article]

Henderson, J. M., & Smith, T. J. (2009). How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition, 17,* 1055–1082.

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

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2,* 194–203. [PubMed]

Judy, P. F., Kijewski, M. F., Fu, X., & Swensson, R. G. (1995). Observer detection efficiency with target size uncertainty. *SPIE, 2436,* 10–17.

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition, 17,* 979–1003.

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

Land, M. F., & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41,* 3559–3565. [PubMed]

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28,* 1311–1328. [PubMed]

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision, 9*(11):8, 1–13, http://journalofvision.org/9/11/8/, doi:10.1167/9.11.8. [PubMed] [Article]

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research, 46,* 614–621. [PubMed]

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123. [PubMed]

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422. [PubMed]

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14):16, 1–20, http://journalofvision.org/7/14/16/, doi:10.1167/7.14.16. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659. [PubMed]

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research, 2,* 1–18.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition, 17,* 1029–1054.

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

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research, 43,* 333–346. [PubMed]

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition, 12,* 473–494.

van Diepen, P. M. J., Wampers, M., & d'Ydewalle, G. (1998). Functional division of the visual field: Moving masks and moving windows. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 337–355). Oxford, UK: Elsevier.

Velichkovsky, B. M., Joos, M., Helmert, J. R., & Pannasch, S. (2005). *Two visual systems and their eye movements: Evidence from static and dynamic scene perception*. Paper presented at the Cognitive Science Society, Stresa, Italy.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115,* 787–835. [PubMed] [Article]

Zelinsky, G. J., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (vol. 18, pp. 1569–1576). Cambridge, MA: MIT Press.