Scrambled eyes? Disrupting scene structure impedes focal processing and increases bottom-up guidance

Tom Foulsham · Rana Alan · Alan Kingstone

Published online: 7 June 2011
© Psychonomic Society, Inc. 2011

Abstract Previous research has demonstrated that search and memory for items within natural scenes can be disrupted by “scrambling” the images. In the present study, we asked how disrupting the structure of a scene through scrambling might affect the control of eye fixations in either a search task (Experiment 1) or a memory task (Experiment 2). We found that the search decrement in scrambled scenes was associated with poorer guidance of the eyes to the target. Across both tasks, scrambling led to shorter fixations and longer saccades, and more distributed, less selective overt attention, perhaps corresponding to an ambient mode of processing. These results confirm that scene structure has widespread effects on the guidance of eye movements in scenes. Furthermore, the results demonstrate the trade-off between scene structure and visual saliency, with saliency having more of an effect on eye guidance in scrambled scenes.

Keywords Scene perception · Eye movements · Visual search · Memory · Object recognition · Saliency maps

Our perception and understanding of what we see does not take place in a vacuum. Instead, the objects most important for our behavior are embedded in the context of a scene. These scenes are made up of a large amount of visual

information: backgrounds, surfaces, and large and small objects, defined by colors, textures and other features. Importantly, unlike the randomly arranged items sometimes used in laboratory investigations of, for example, visual search, the spatial configuration of these objects is predictable. Before stepping inside a room, one can be fairly sure that tables, chairs, and other furniture will be on the ground and in the lower half of the visual field, whereas paintings will be on the wall and light fittings in the ceiling. In the present article, we consider the effects of this context on how people scan natural scenes.

There have been several previous demonstrations that scene context affects the speed at which people locate or recognize objects. Of particular relevance for the present research, in a series of experiments, Biederman and colleagues (Biederman, 1972; Biederman, Glass, & Stacy, 1973; Biederman, Rabinowitz, Glass, & Stacy, 1974) manipulated the structural layout of scenes and recorded how well observers could perceive objects in a briefly flashed display. Biederman divided photographs of natural scenes into six sections and presented these sections either in their original spatial configuration or randomly rearranged (a manipulation that has been called “jumbling” or “scrambling”). Participants were probed with an arrow pointing at a spatial location and were required to identify which of four objects had been at this location. Across display durations of 300, 500, and 700 ms, observers were more accurate at recognizing objects in coherent than in scrambled scenes. Because the target object was in the same location in both normal and scrambled scenes, and because the local features of this object were not modified, the authors attributed the recognition benefit for coherent scenes to an effect of the holistic gist of the scene on object perception. Furthermore, this context effect remained even when participants knew the cued location and the

T. Foulsham (✉)
Department of Psychology, University of Essex,
Wivenhoe Park,
Colchester CO4 3SQ England, UK
e-mail: foulsham@essex.ac.uk

R. Alan · A. Kingstone
University of British Columbia,
Vancouver, BC, Canada

alternative objects in advance. This evidence led Biederman to conclude that scrambling affected the perception of the object, rather than just the allocation of attention to the object or short-term memory. In subsequent experiments, Biederman et al. (1974) found that the scrambling effect emerged at scene viewing durations of just 100 ms, which is consistent with a very early perception of the gist (Potter, 1976). Biederman et al. (1973), meanwhile, reported that the scrambling effect interacted with the likelihood of a target appearing in the scene: The advantage for recognizing objects in coherent versus scrambled scenes was reduced when the target was incongruent with the scene gist (e.g., a sofa on a beach).

Other researchers have used eye movements to explore the influence of scene context. In particular, several studies have measured eye movements to key objects that are congruent or incongruent with the scene. Incongruent objects tend to be fixated for longer, but there is some debate as to whether they are fixated earlier, with some studies reporting that oddball objects are looked at sooner than congruent objects (Loftus & Mackworth, 1978; Underwood & Foulsham, 2006) and others finding the opposite (De Graef, Christiaens, & d'Ydewalle, 1990; Gareze & Findlay, 2007; Henderson, Weeks, & Hollingworth, 1999; Vo & Henderson, 2009). This discrepancy is likely due to differences in the stimuli and task between studies (e.g., because oddballs that pop out may also be more visually conspicuous). Either way, the fact that incongruent objects are inspected differently confirms that viewers can perceive the (global) context of a scene, and that this context can influence the way that (local) elements within this scene are inspected with eye movements.

In the case of searching for an object within a natural scene, rather than just freely viewing an image, researchers can be more precise about the prior information that is available for viewers. For example, in Henderson et al. (1999; Experiment 2), participants searched for a named target object in line drawings of scenes. Objects consistent with the scene context were saccaded to more quickly than were those that were inconsistent, presumably because congruent objects are associated with a typical or predictable position within a scene, and attention can be moved immediately to such locations. Furthermore, a preview of a scene is enough to speed up search for an object, even when peripheral vision is restricted by a gaze contingent window and the preview is flashed for just 50 ms (Castelhano & Henderson, 2007; Vo & Henderson, 2010). This preview likely provides a viewer with information about scene identity, which in turn activates stored knowledge about where an object is likely to occur. It may also provide global structural information about the scene, such as the location of ground, walls, surfaces, and so on.

Despite the use of scrambled images in experiments on rapid scene recognition, object identification and visual short-term memory (Velisavljevic & Elder, 2008), it is not known how participants scan such images, which leave local information intact while disrupting configural relations. In addition, although the information summarized thus far suggests strong effects of top-down scene recognition and cognition on eye movements, others propose that bottom-up visual saliency can explain some or all of where people fixate. Specifically, following a model proposed by Itti and Koch (2000), it has been suggested that eye movements in scenes can be predicted by feature contrast, with regions of a scene that stand out from their background being preferentially fixated, regardless of their meaning (Parkhurst, Law, & Niebur, 2002). Although a majority of recent research has shown that saliency alone does not predict fixation (Foulsham & Underwood, 2008; see the recent special issue, Tatler, 2009, for many more examples), especially when viewers are engaged in an active task such as visual search, the combination of top-down and bottom-up guidance in scenes remains an important topic for models of scene viewing. Our aims in the present study are thus (a) to describe the effect on eye movements of disrupting scene context through scrambling, and (b) to investigate the relationship between this context and eye guidance to salient image regions.

The present research

In the present research, we conducted two experiments to investigate the effect of scrambling on eye movements. On the basis of previous research, we anticipated that scrambling would alter the configuration of the scene, leading to disruptions in accessing scene gist. Thus, differences in the eye movements made to unmodified versus scrambled scenes will inform us about the effects of this configuration in normal scene viewing. Additionally, measuring eye movements will allow us to explore which aspects of active visual search (Experiment 1) and scene encoding (Experiment 2) are disrupted by scrambling. Previous research suggests that scrambling slows search and makes memory encoding and change detection less efficient, which might manifest itself in eye movements that were less well guided—more fixations and scanpaths, which take longer to reach the target. If scrambling affects the initial recognition of scene gist, which occurs very quickly, then it should also have an effect on the first eye movement (e.g., by prolonging the latency of the initial saccade). The duration of individual fixations is often assumed to mainly reflect the processing of information at that location, and so this measure might not be affected by scrambling, which changes the global context while leaving local elements intact.

Moreover, in combination with variations in saccade amplitude, fixation duration may indicate a change in the overall “processing mode” of observers, with short fixations and large saccades thought to reflect “ambient” or less focused exploration (Pannasch, Helmert, Roth, Herbold, & Walter, 2008; Velichkovsky, Joos, Helmert, & Pannach, 2005). For a deeper understanding of the potential factors involved in eye guidance, we also used a computational model to examine the bottom-up saliency of target regions. This analysis provided a test of the relative role of top-down and bottom-up guidance in scene viewing, as well as gave potential explanations for any differences caused by scrambling. We also manipulated the degree of scrambling by dividing images into either 16 or 48 regions and randomly rearranging the pieces, with the prediction that smaller pieces would leave less scene structure intact and would therefore disrupt performance and eye movements to a greater extent.

Experiment 1

In Experiment 1, participants saw a patch containing a target object, which they subsequently searched for in either normal or scrambled natural scenes. We analyzed the eye movement behavior during viewing of the search display, as well as the manual reaction time (RT). In addition, we computed the visual saliency of target patches to see whether regions that were more salient were found more quickly, as predicted by Itti and Koch (2000). If, within normal scenes, eye guidance by saliency is overridden or moderated by top-down guidance (such as that gleaned from the global context), then we hypothesized that saliency should have more of an effect on search in scrambled scenes.

Method

Participants Twenty-four undergraduate students (17 females) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision.

Stimuli and design A set of 80 high-resolution color photographs of natural scenes were collected from a Google image search and were used as stimuli. Images were resized to 1024×768 pixels and, during the experiment, participants maintained a fixed viewing distance of 60 cm, resulting in an image that subtended $31^\circ \times 25^\circ$ of visual angle.

Four versions of each scene were prepared, reflecting the orthogonal combination of two factors: the size (or “grain”) of the grid used to divide up the scene, and whether the image was scrambled or not. In “coarse” conditions, the

scene was divided into 16 large regions using a regular 4×4 grid, whereas in “fine” conditions, it was divided into 48 (8×6) regions. In “scrambled” conditions, the grid regions were randomly repositioned, whereas in “normal” conditions, they stayed in their original positions. Figure 1 shows an image in all four conditions. Following Velisavljevic and Elder (2008), we overlaid a two-pixel thick black grid over all images. This grid functioned to compensate for some of the color and intensity discontinuities that were introduced into the scrambled images, in all pictures.

The 80 pictures that were presented to each participant consisted of 20 images in each of the four conditions: coarse normal, coarse scrambled, fine normal, and fine scrambled. Each participant saw each scene once, but stimuli were rotated across participants so that images were seen in all four conditions across the entire experiment. In each image, one grid region was selected as the target, and in order to equate target position, this region remained in its original position in all conditions. Regions were chosen so as to encompass an object in the scene (see Fig. 1 for an example). Half of the pictures from each condition were paired with a target from that image, whereas the other half were presented with incorrect targets that did not originate from the same image. The incorrect targets were randomly selected regions from other natural scenes and were cropped to the same size as the correct targets.

Apparatus An SR Research EyeLink II system was used to record eye movements. The system consisted of a head-mounted camera that sampled pupil location monocularly from the pupil image at 250 Hz. A nine-point calibration and validation procedure was repeated several times to ensure a high spatial resolution (less than 0.5° error). The system parsed the eye movements into fixations and saccades using criteria of displacement (saccades $> 0.15^\circ$), velocity ($> 30^\circ/\text{s}$) and acceleration ($> 4,000^\circ/\text{s}^2$). Head movement was minimized by means of a chin rest.

Procedure Participants were given written instructions prior to the experiment. The eye tracker was then set up and calibrated. The experiment consisted of 80 trials. Each trial started with a fixation point in the center of the screen, at which point the experimenter confirmed gaze position, allowing the tracker to compensate for any drift in the calibration. An image of the target region was then presented in the center of a gray screen for 3 s. This was followed by the scene, in which the participant was instructed to search for the target and make a keyboard response to indicate target presence or absence, as quickly and as accurately as possible. In pilot experiments, we found that asking participants to specify the location of the target encouraged them to search more thoroughly, rather than trying to guess on the basis of the gist of the image.

Fig. 1 One of the scenes used in Experiments 1 and 2, appearing in all four conditions. Scenes were divided into either a coarse grid (top row) or a fine grid (bottom row), and the pieces were either coherent

(left column) or randomly scrambled (right column). Note that the target patch (here containing the clock and highlighted for illustration only) remained in the same location

Following a “present” keyboard response, we therefore presented an empty black grid on a white background (of the same size as the image divisions) and asked participants to click with the mouse where the target had been located. This mouse response terminated the trial and triggered the next fixation point.

Results

We considered accuracy and manual reaction time, to ascertain the effect of grain and scrambling on search performance, followed by several measures of eye movement behavior. In each case, participant means were analyzed using a 2×2 repeated measures ANOVA with factors of grain (coarse vs. fine) and scrambling, and interactions were probed with paired comparisons where necessary.

Task performance Participants were very accurate at searching, resulting in a mean proportion of hits (where present targets were correctly found) of 91% ($SD = 2\%$). Participants made very few false alarms (where absent targets were

incorrectly endorsed as being present; $M = 2\%$, $SD = 1\%$). Hits and false alarms did not differ reliably between conditions [all effects and interactions, $F(1, 23) < 2.3$, $ps > .14$]. We therefore confined all subsequent analyses to only those trials that culminated in a correct response.

There were effects of grain, $F(1, 23) = 14.0$, $p = .001$, and scrambling, $F(1, 23) = 3.9$, marginal at $p = .06$, on the correct, target-present reaction time (Fig. 2). Participants were slower to respond to small targets in a fine display than to the coarse displays, and scrambled scenes were also responded to more slowly than were normal scenes. However, these differences were qualified with a significant interaction, $F(1, 23) = 5.0$, $p = .03$. In fact, scrambling had an effect only in fine displays, $t(23) = 2.3$, $p = .03$, whereas in coarse displays, normal and scrambled scenes elicited similar reaction times, $t(23) < 1$. Conversely, grain only had an effect in scrambled scenes, $t(23) = 3.4$, $p = .002$, whereas in normal scenes, fine and coarse images were responded to equally quickly, $t(23) < 1$.

When participants identified a target as being present, they were asked to localize the target in a blank grid, and we recorded the proportion of these trials in which

Fig. 2 Mean reaction time (with standard error bars) for the different search displays

participants chose the correct region. Participants were generally very accurate at this task, and there were no effects of grain, $F(1, 23) = 1.3$, $p = .3$, or scrambling, $F(1, 23) < 1$. However, there was a reliable interaction, $F(1, 23) = 7.8$, $p = .01$. Participants were equally accurate in coarse normal, coarse scrambled, and fine normal trials (all $M_s = 97\%$), but they were less accurate in fine scrambled trials ($M = 93\%$).

Paired t test comparisons showed that fine scrambled trials were associated with significantly poorer performance than were fine normal trials, $t(23) = 2.1$, $p = .047$, but no other comparisons were reliable. Thus, scrambling made search slower and localization less accurate, but only when a fine grid size was used.

Eye movements Figure 3 summarizes the eye movement measures taken. As reviewed in the introduction, eye tracking allows us to look at some of the different events that may have led to the observed differences in manual reaction time. First, we considered the time between the onset of the search display and the first eye movement, which can be termed the *initial saccade latency*, and which others have suggested is the “initiation time” during which the overall gist of the scene is perceived and search planning begins (Malcolm & Henderson, 2009; Vo & Henderson, 2010).

The initial saccade latency was between 200 and 300 ms, on average, in all cases. There was a significant effect of grain on this latency, $F(1, 23) = 4.5$, $p = .045$, a marginal effect of scrambling, $F(1, 23) = 3.2$, $p = .09$, and a reliable

Fig. 3 Oculomotor measures of search in the different conditions: the initial saccade latency (a), the time to first fixate the target (b), the number of fixations per trial (c), and the average fixation

duration (d). In each case, bars show the mean (with standard error bars) across participants

interaction between the two factors, $F(1, 23) = 5.6, p = .03$. The saccade latency with normal, coarse displays was faster than it was in all other conditions, which did not differ from each other. Thus, the main effect of grain was reliable only in normal scenes, $t(23) = 2.7, p = .01$, and scrambling prolonged the initial saccade latency, but only in coarse images, $t(23) = 2.7, p = .01$. Therefore disruptions in scene context introduced by scrambling had an effect on the planning of the first saccade, but only in the relatively easier coarse displays. It is also interesting that, relative to that in the coarse normal condition, the fine normal condition had an increased latency.

The oculomotor components of the overall search time, during which participants were scanning the scene looking for the target, was investigated by measuring the time taken to first fixate the target region (the “time to target”), the number of fixations, their average duration, and the average amplitude of the saccades. The time to target supplements the manual reaction time as a measure of search speed, and ensures that longer RTs are not due only to differences in the time spent processing the target or preparing the response once the target had been fixated. Because rejection of target absent trials tends to take longer, and because it is not always clear what participants are doing in these trials, we calculated these measures for correct, target present trials only.

Target regions were fixated on almost every trial, and the time to target was calculated from only those trials in which it was fixated at least once. The time-to-target results were qualitatively similar to the RT results: Targets in coarse displays were fixated earlier than those in fine displays, $F(1, 23) = 16.2, p = .001$, and targets in normal scenes were fixated before targets in scrambled scenes, $F(1, 23) = 8.2, p = .008$. In this case, the interaction did not reach significance, $F(1, 23) = 2.2, p = .15$, although as was previously observed, the difference between normal and scrambled was more pronounced in fine displays. Thus, it took longer to guide the eyes to targets in scrambled (and fine) scenes.

The differences in search time might be explained by either differing numbers of fixations (e.g., if more locations had to be inspected before finding the target) or fixations of a different duration (e.g., if more processing time was needed on each fixation), or a combination of the two. Consistent with the increased search time, participants made more fixations per trial, on average, in fine displays than in coarse displays, $F(1, 23) = 5.9, p = .02$, and they made more fixations in scrambled scenes than in normal scenes, $F(1, 23) = 6.6, p = .02$. However, these effects were qualified by a Grain $\times$ Scrambling interaction that came about because scrambling made a difference to the fixation count only in finely divided scenes, $t(23) = 2.6, p = .02$.

Interestingly, scrambling also made a difference to the average fixation duration in a scene. Although search took longer in scrambled scenes, the mean duration of all fixations in a scene was shorter in scrambled than in normal images, $F(1, 23) = 14.2, p = .001$. There was also an effect of grain on average fixation duration, with fixations being longer, in general, in finely divided scenes than in coarsely divided scenes, $F(1, 23) = 8.4, p < .005$. There was no interaction, $F(1, 23) < 1$. If the differences in mean duration had come about because of different subpopulations of fixations (e.g., because scrambled images led to an increase in very short fixations, as well as to an increase in much longer ones), one would expect to see a difference in the frequency distribution of these durations (in this case, a bimodal distribution). However, visual inspection of the distributions for all conditions confirmed that all were unimodal and approached a normal distribution. It is also important to note that a fine grid, as compared with a coarse grid, also lengthened fixations in normal scenes. As with the effect of grain on initial saccade latency, the likely explanation is that the increased number of grid lines competed for attention, thus prolonging saccade planning.

Velichkovsky et al. (2005) suggested that shorter fixations were associated with a particular “ambient” mode of scene processing, a mode that was also distinguishable by larger saccades. In the present study, we also found that scrambled scenes elicited saccades with greater mean amplitude than did normal scenes, $F(1, 23) = 70.1, p < .001$. There was a marginal effect of grain, $F(1, 23) = 3.9, p = .06$, and a reliable interaction, $F(1, 23) = 4.8, p = .04$. Scrambling produced larger saccades in both coarsely divided scenes (scrambled $M = 5.6^\circ$; normal $M = 4.9^\circ$) and fine displays (5.8° vs. 4.3°). Intact scenes with a fine grid showed the smallest saccades. Although the distribution of saccades in all conditions showed a characteristic, long-tailed positive skew (see, e.g., Foulsham, Tetzka, & Kingstone, 2011), none were bimodal, and the same effects were seen when the median amplitude was analyzed.

Visual saliency

The second aim of the present study was to investigate whether the scrambling manipulation interacted with the visual saliency of the target. On the basis of previous research demonstrating that top-down factors such as the knowledge of the task and the gist of the scene override saliency, we hypothesized that salient targets would be quicker to find, but that this effect would be larger in scrambled conditions because of the absence of scene structure.

To investigate this hypothesis, we used Walther and Koch’s (2006) saliency toolbox, an implementation of the Itti and Koch (2000) model of visual saliency, to compute

visual saliency maps for all of the search displays (see Fig. 4 for an example). These maps indicate the feature contrast within color, intensity, and orientation channels for each pixel in the image, with high values (bright areas or “peaks” in the saliency map) indicating regions that stand out from their surround. The model simulates overt and covert attention with a winner-take-all network, predicting that people should move to the most salient point in the image, followed by the next most salient point, and so on. Saliency maps were scaled to a fixed range between 0 and 1, with 1 indicating the most salient location in each map.

Inspection of the saliency maps revealed that the grid lines on the images were occasionally selected as salient by

the model, but we made a conservative assumption that, because these lines were present in all images and regions, and were irrelevant for the task, they would not be fixated often in our experiment. We tested this assumption by defining a relatively liberal region of interest of 1° around the grid lines. On average, only 21% of fixations in coarse images and 33% of fixations in fine images were on these regions (and this was not affected by scrambling, both $ps > .15$). This proportion of fixations is less than expected given the area covered by these regions (26% in coarse and 41% in fine displays), and the higher likelihood of fixating near the grid in fine displays can be attributed to this area being greater (i.e., because there were more grid lines). When the

Fig. 4 Search displays (e.g., **a**, from the coarse scrambled condition) were processed to create saliency maps (e.g., **b**) showing the regions that were most visually salient. Bright parts of the saliency map correspond to regions with high contrast in color, intensity, or orientation, and empty regions with low saliency are black. Areas

around the grid lines were excluded and set to zero. The saliency of target regions was then correlated with the average time to target for coarse (**c**) and fine (**d**) search displays. Data points show the mean per image, with least-squares regression lines, for normal and scrambled displays separately

number of gridlines was taken into account, there was also no difference in the model-predicted saliency of coarse and fine gridlines. Nevertheless, to reduce any influence of the grid lines, we excluded a border of approximately 0.5° around the edge of each region in the saliency map prior to calculating target saliency.

We took the maximum saliency value of all of the pixels in each target region of the search displays and used this as an estimate of how salient the target was, and therefore how likely it was to draw attention and fixation in a bottom-up fashion. Target saliency was calculated separately for normal and scrambled versions of each image. Scrambling sometimes changed the model-predicted saliency of target regions by introducing discontinuities at region edges, but these differences were very minor and were largely eliminated by the grid lines and by excluding the boundary as previously described. We took time to target as our measure of search, and performed an analysis by items, calculating the mean time to target across participants for each image–target combination. We then correlated target saliency with time to target, and the results are shown in Fig. 4 (right column). In both coarse and fine displays, the slope relating saliency and search time were more steeply negative for scrambled trials, whereas for normal trials, the slope was closer to zero, indicating no relationship. In trials with a coarse grid, the correlation between saliency and time to target was not significant for normal displays, $r = -.20$, $p = .21$, but was reliable for scrambled displays, $r = -.40$, $p = .009$. In fine displays, the relationship was almost zero in normal scenes, $r = -.07$, $p = .67$, but marginally significant in scrambled scenes, $r = -.29$, $p = .07$. In sum, all conditions showed a negative correlation, such that more salient targets were fixated more quickly, but this relationship was far stronger, and significant only in scrambled scenes.

Discussion

Disrupting the scene by scrambling had widespread effects on performance, both in terms of manual RT and in the eye movements made. We confirmed the effect of scrambling on performance demonstrated by Biederman (1972) and others, and we asked how this change in search time was related to a change in oculomotor behavior. Scrambling disrupted the global structure of the image, violating the semantically appropriate arrangement of objects and the relationship between different items in the scene while leaving the local visual information intact. Furthermore, because targets remained in the same location, we can be sure that they were not found more easily in normal scenes because of their absolute spatial position. The reason that the scrambling effect is interesting is that it cannot be readily explained by a simple image-based account of

search, whereby each candidate region is compared to the known features of the target until the target is found. Such an account is implied by several classic theories of visual attention in search (Treisman & Gelade, 1980; Wolfe, 1994). More recent versions of these theories are designed to model search for coherent objects, and so it is possible that they could accommodate our findings with reference to the object structures that are disrupted in scrambling the image. That said, it is interesting that Zelinsky's (2008) target acquisition model is explicit about the operations required to compare a target template to regions in a to-be-searched scene, but these target-local feature comparisons should *not* be affected by scrambling. Instead, we interpret the effect as being due to a disruption of gist recognition and the ability to use this gist to guide search. This top-down effect is reminiscent of re-entrant, feedback pathways in attention (Hochstein & Ahissar, 2002), and is supported by neural evidence of the role of context on object recognition (Bar et al., 2006).

It is worth noting that the RT decrement for scrambled scenes was found only in fine conditions, in which there were more candidate regions, and the region containing the target was smaller. This is likely due to both the number of distractor regions (i.e., a set size effect) and the size of the region (because coarse regions would have been identified more easily in peripheral vision). This may have made the coarse trials too easy to be disrupted by scrambling. Alternatively, it may have been easier to identify the gist and layout of the scene with large scrambled pieces than with fine displays where less structure was left intact. The eye movement results uncovered several subtleties about the effect of the scrambling manipulation on search.

First, scrambling affected the latency of the first eye movement in the scene, at least some of the time. This is consistent with scrambling disrupting the initial processing of gist, which can be accomplished quickly and thus occurs before the first saccade (Biederman et al., 1974; Potter, 1976). On the other hand, the latency effect was found only in the relatively easy coarsely subdivided scenes. Because this pattern is the opposite of the RT result, the initial saccade latency cannot have determined the difference in search time. The grain of the grid also affected the planning of the first saccade in normal images. One potential explanation is that the additional grid lines introduced in the fine condition caused simultaneous masking of the scene, making potential saccade targets such as objects and the target harder to identify in peripheral vision. Another explanation is that the additional grid lines competed for attention, producing salient potential targets that were meaningless for the task. By this account, increased saccade latencies reflect an inhibition of these additional targets. Indeed, the grid lines in fine displays were fixated more frequently, suggesting that increased interference from these

lines underlies the difference in initial saccade latency. Moreover, assuming that part of the duration of individual fixations comprises the planning of the next saccade, this explanation would also account for the longer average fixation duration in fine normal displays.

A second insight into the effects of scrambling was that the RT pattern was mirrored by the time it took to move the eyes to the target region. This is important because it confirms that the RT effect is due to poorer search guidance, and not just, for example, a difficulty in processing or recognizing the target after it had been fixated. The prolonged search time was mainly due to a higher frequency of fixations in the fine scrambled condition. This presumably reflects the fact that more regions had to be fixated before the target could be found, and thus that eye guidance was impaired in this condition.

Third, we found a novel effect of scrambling on average fixation duration. The increased number of fixations and RT in scrambled images was partially offset by the fact that these fixations were shorter on average than those in normal scenes. This is particularly interesting because scrambling preserved the local image information within each region, and it is normally assumed that fixation durations reflect processing of this *local* information. This result is consistent with research from experiments that show effects of peripheral information on fixation durations. For example, Loschky and McConkie (2002); van Diepen and d'Ydewalle (2003); Foulsham and Underwood (2011); and Foulsham et al. (2011) used a gaze contingent display to mask or blur the information outside a window at fixation, and found that this lengthened average fixation duration. Why did disrupting the global organization of a scene by scrambling *shorten* fixation durations, when regions tend to be more easily recognized if context is intact? There are at least two possible explanations for this counterintuitive result. First, it might be that because scrambling disrupted the guidance of fixations, the locations selected were, on average, less semantically informative. Thus, rather than dwelling for longer on each location and trying to recognize scene features without the aid of complete contextual information, participants chose to reposition fixation after a relatively brief fixation. Second, because the fixation duration presumably includes some time associated with the planning of the next saccade, saccades in scrambled scenes may have been planned differently. If saccades in the scrambled conditions were more random, or were less driven by top-down information or the recognition of objects in the periphery, then they might be planned and initiated more quickly. Indeed, this could be the case if saccades in scrambled scenes were driven by bottom-up saliency to a greater extent than in other conditions.

To test this possibility, we were able to investigate the guidance of saccades to targets more directly by looking at

the relative role of visual saliency. Previous work has demonstrated that the effects of saliency on search are minimal (Chen & Zelinsky, 2006; Foulsham & Underwood, 2007), and the implication of this research is that when we search, what we know about the target and the scene overrides any bias to look at conspicuous items. Where saliency is predictive of search, it is likely because saliency coincides with other properties of the target that make it easier to find, such as ease with which its location can be predicted in a scene (Foulsham & Underwood, 2011). In the present research, we found that there was almost no correlation between saliency and the time taken to move the eyes to a target in normal scenes. However, there was a much stronger, and statistically significant, relationship between target saliency and search time in scrambled scenes. This new finding reveals that, when contextual information is disrupted, eye guidance is accomplished more on the basis of bottom-up feature contrast. Saliency is a better predictor of where we look when the contextual organization of where things should be is disrupted.

A final point of interest from this experiment is that the pattern seen in scrambled images—that of shorter fixations and larger saccades—is similar to that seen in Velichkovsky et al. (2005), when participants were thought to be in an ambient processing mode. This mode, which Velichkovsky et al. identified with the dorsal visual system, was associated with reduced cognitive analysis, which would explain the poorer performance in the search task used here. The scrambling manipulation may have prolonged a transition to a closer, focal mode of scanning. This is also consistent with the poorer performance associated with ambient processing that Velichkovsky et al. found in a memory experiment. In Experiment 2, we used a memory test with scrambled stimuli and we asked whether we would find the same effects of scrambling on performance, eye movements, and processing mode.

Experiment 2

The manipulation of context by scrambling has also been shown to affect the memory of a scene. In a memory version of his scene coherence paradigm, Biederman (1972) found that participants were better at recalling an object that had previously appeared at a certain spatial location when the global context was coherent, than in scenes divided into six blocks and randomly rearranged. The contextual configuration also affects visual short-term memory for scene regions: Velisavljevic and Elder (2008) demonstrated that when scrambled scenes were flashed briefly, recognition of local elements was impaired relative to coherent scenes. Zimmermann, Schnier, and Lappe (2010) also found persistent RT benefits when normal, rather than

scrambled scenes were viewed during a change detection task. Thus, objects and scene regions seem to be recognized and remembered better when in a coherent than in a scrambled context. In this experiment, we asked whether the eye movement results from Experiment 1 would transfer into a memory task where scenes were encoded without the prior knowledge of a specific target region.

Participants were shown normal and scrambled scenes for a few seconds, and were then probed with the presentation of a region from the same, or from another, scene. Because participants did not know the target's identity before the image presentation, and viewing duration was fixed, we expected to see fewer effects on eye movements, and these effects cannot be due to guidance toward target features. We have previously demonstrated that saliency is more closely related to fixation in an encoding task in which participants have no specific target than in a search task (Foulsham & Underwood, 2007). We therefore predicted that saliency would correlate better with fixation than in Experiment 1, and the extent to which scrambling enhances this correlation will signify the relative contribution of top-down and bottom-up guidance in a different task.

Method

Participants Twenty-four undergraduate students (11 females), who had not taken part in Experiment 1, participated in the experiment for course credit. All participants had normal or corrected-to-normal vision.

Stimuli, apparatus, and design The stimuli and design for this experiment were exactly the same as in Experiment 1. The same targets and foils were used, and we will refer to these as targets to remain consistent with Experiment 1, even though the participants did not know their significance until after viewing the scene. Eye movements were tracked during scene encoding using the EyeLink 1000 system, which is a desk-mounted eye-tracker with very good spatial (accurate to within 0.5°) and temporal (1000 Hz) resolution. Participants kept their head rested on a chin rest, and all other aspects of the set up were the same as those in Experiment 1.

Procedure The procedure was essentially the inverse of the trial sequence used in the previous experiment. Following calibration and a drift correct marker, each image was presented for 3 s, and participants were instructed to look at the scene and to try to remember it. After 3 s had elapsed, the scene was replaced with a target region that was either from the same image or from another scene. The viewing duration was chosen on the basis of memory performance in previous studies (e.g., Foulsham & Underwood, 2008)

and led to trials of a similar length to Experiment 1. Participants were instructed to respond, as quickly and accurately as possible, with one of two keys on the keyboard to indicate whether the target had been in the previous scene or not. An affirmative response was followed by a spatial probe in which participants had to click with a mouse cursor to indicate the grid region in which the target had appeared.

Results

Task performance We assessed task performance by measuring the participants' accuracy at recognizing the target and their RT to do so (taken as the time between the onset of the target region and the participant's keyboard response).

Participants were more accurate at recognizing regions that they had seen in a coherent context than in a scrambled context [Fig. 5a; $F(1, 23) = 15.4$, $p < .005$]. Grid size also had a reliable effect, $F(1, 23) = 23.8$, $p < .001$. There was no interaction, $F(1, 23) < 1$, and very few false alarms were made in any of the conditions ($M = 2\%$, $SD = 3\%$). The effects on accuracy were quite striking, with participants making around 30% more hits in coarse normal scenes than in fine scrambled images.

Targets were also responded to more slowly in scrambled scenes than in intact scenes, although, as in Experiment 1, this was only the case in finely divided images (Fig. 5b). Neither of the main effects were reliable [grain, $F(1, 23) = 1.4$, $p = .25$; scrambling, $F(1, 23) = 1.2$, $p = .28$]. The interaction did not reach conventional significance, $F(1, 23) = 3.1$, $p = .09$, although as in Experiment 1, the scrambled fine condition elicited the longest RT. Thus, even though the response in this case was made to targets that were identical between conditions, and were around 5 s after the onset of the scene, scrambling had a large effect on accuracy and a marginal effect on reaction time.

Eye movements Analysis of eye movements in this experiment again served to investigate how scene perception was affected by scrambling. However, the measures of interest differed somewhat from the search task in Experiment 1 because target identity was unknown for the observer at the time of inspection. We therefore focused on differences in the general scanning of the scenes rather than eye guidance to a specific target. In particular, we asked whether the poorer accuracy in scrambled scenes could be attributed to a difference in the way that scenes were looked at in preparation for the memory probe. The measures taken are summarized in Table 1.

We began by looking at the latency of the first saccade. In contrast to Experiment 1, there was no reliable effect of

Fig. 5 Accuracy and reaction time in the region recognition task. Charts show the mean across participants, with standard error bars

grid size, $F(1, 23) = 2.1$, $p = .16$, or scrambling, $F(1, 23) = 3.0$, $p = .1$, and no interaction, $F(1, 23) = 1.8$, $p = .19$. The latencies in this experiment were also in general longer than those in Experiment 1 [independent samples t test, $t(46) = 4.0$, $p < .001$], a finding probably attributable to time pressure in the search task.

Unlike in the search task, participants had a fixed time of 3 s during which to look at the scenes, so differences in the number or duration of fixations, their distribution, or the length of the saccades made would indicate a change in the way that different parts of the scene were sampled within this trial time. There was no effect of grain on either the number of fixations per trial, $F(1, 23) < 1$, or the average fixation duration, $F(1, 23) = 1.5$, $p = .2$. However, there were marginal effects of scrambling on both these measures, indicating a change in the way the images were sampled during encoding. Trials with scrambled displays elicited more fixations [although this was a small effect: scrambled, $M = 11.7$; normal, $M = 11.45$; $F(1, 23) = 3.7$, $p = .07$], which were slightly shorter in duration, $F(1, 23) = 3.1$, $p = .09$. Although these trends did not reach significance, they do confirm patterns seen in Experiment 1 and therefore suggest that, even when viewing time is fixed, scrambling affects the way that eye movements are distributed over this time. There were no interactions in either measure [fixation number, $F(1, 23) < 1$; fixation

duration, $F(1, 23) = 1.4$, $p = .24$]. There was no evidence of a bimodal distribution of fixation durations in this experiment. However, as shown in Table 1, there was an effect of scrambling on saccade amplitude, $F(1, 23) = 76.1$, $p < .001$. As in Experiment 1, larger saccades were made in scrambled scenes than in normal scenes. This effect was qualified by an effect of grain, $F(1, 23) = 5.1$, $p = .03$ and an interaction, $F(1, 23) = 4.6$, $p = .04$. A finer grid elicited shorter saccades, but this was only the case in a scrambled image.

Why were participants less likely to recognize the target in scrambled scenes? One possibility is that they failed to encode it, perhaps because they were less likely to look at the to-be-probed region. We defined the probability of target fixation as the proportion of target-present trials where the target patch was fixated at least once. If all fixations were distributed uniformly, and regions were not refixated, the chance of fixating the target region was much lower in fine trials ($12/48 = 25\%$, as compared with $12/16 = 75\%$ in coarse trials), by virtue of its smaller size and the larger number of nontarget regions. The probability of target fixation was greater than chance in fine trials and was close to chance in coarse trials (see Table 1). Across all target-present trials, this probability was affected by grid size, $F(1, 23) = 44.1$, $p < .001$, but not by scrambling, $F(1, 23) = 1.1$, $p = .3$. Participants were less likely to fixate the target

Table 1 Summary statistics for the eye movement measures taken in Experiment 2

	Coarse				Fine			
	Normal		Scrambled		Normal		Scrambled	
	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>
Initial saccade latency (ms)	281	8.0	277	12.2	303	15.4	286	13.1
Number of fixations per trial	11.5	0.25	11.7	0.29	11.4	0.27	11.7	0.26
Mean fixation duration (ms)	233	6.6	226	7.5	258	21.2	228	8.2
Mean saccade amplitude (°)	6.6	0.17	7.6	0.20	6.6	0.15	7.2	0.22
Probability of target fixation	69%	3.6	68%	3.6	45%	3.7	39%	4.4

region in fine images than in coarse images, presumably because there were many more nontarget regions, but scrambling had no reliable effect. There was no interaction, $F(1, 23) < 1$. This advantage likely reflects the fact that target regions were not random but were predictable from their contents (e.g., because they contained an object on which they were more likely to be probed). Across trials, there was a significant association between fixating a target and subsequently recognizing it, with correct recognition probability being higher when the target had been fixated (89% hits) than when it had not (67%; $X^2 = 77$, $df = 1$, $p < .001$). However, since the target fixation probability was not significantly affected by scrambling, this cannot, in itself, explain the poorer recognition in scrambled trials.

A remaining possibility is that fixations were distributed differently in scrambled scenes. To analyze this in a way that was independent of the size of the grid, we performed an additional test of scene coverage. In each trial, we defined a circular region of 2° radius around each fixation location, and we calculated the proportion of the total scene area that was covered by the accumulation of these regions. The size of region was chosen to reflect a conservative estimate of the useful field of view, but other sizes produced the same outcome. As we have seen, observers made around 12 fixations per trial, but this analysis measured whether these fixations were spread out over different areas or concentrated on a few parts of the image. Scene coverage was reduced in fine (mean coverage = 15.5%, $SD = 2.0$) relative to coarse displays [mean coverage = 16.4%, $SD = 2.0$; $F(1, 23) = 19.1$, $p < .001$], which is further evidence that the grid size affected scanning. More interestingly, there was an effect of scrambling, with fixations being more distributed in scrambled scenes (mean coverage = 16.6%, $SD = 2.1$) than in normal scenes [mean coverage = 15.2%, $SD = 1.9$; $F(1, 23) = 69.6$, $p < .001$]. The interaction was not reliable, $F(1, 23) < 1$. This analysis therefore confirms that fixations in scrambled scenes were more distributed than those in normal, undisrupted images.

Visual saliency

In Experiment 1, we found that saliency was a better predictor of target fixation in scrambled scenes. We repeated this analysis for fixations in the memory task. However, because target regions were fixated relatively less often (since participants didn't know what to look for until after the viewing period), we used the probability of target fixation as the outcome measure. Doing this allowed us to include all trials in the analysis, and we asked whether targets that were more salient relative to the other parts of the image were more likely to be fixated. We calculated the probability of the target region being fixated for each scene, and then correlated this value with the saliency of the target,

computed in exactly the same way as in Experiment 1. The results are shown in Fig. 6 for the different types of scene.

All of the correlations showed a positive slope, confirming the predicted direction in that targets that were more salient were more likely to be fixated during the memory task. However, the relationships were slight, particularly in coarse images in which the correlation was not significant for either normal, $r = .16$, $p = .33$, or scrambled, $r = .23$, $p = .16$, images. In the fine condition, saliency was not a significant predictor in normal scenes, $r = .12$, $p = .47$, but in scrambled scenes, the correlation was reliable, $r = .32$, $p = .04$. Some targets were always fixated (i.e., fixation probability = 1), and some were never fixated (fixation probability = 0), but saliency did not provide a good explanation for why. On the other hand, the effect of scrambling on this correlation is consistent with that seen in

Fig. 6 The correlation between the saliency of a target and the likelihood of it being fixated during the memory task. Data show the mean across items, with least-squares regression lines, for coarse (a) and fine (b) trials

Experiment 1. In both coarse and fine displays, scrambling led to a stronger positive correlation between saliency and fixation, and the only condition in which saliency could reliably predict target fixation was in (fine) scrambled scenes.

Discussion

In Experiment 2, we examined the extent to which the effects of scrambling on eye movements in search would generalize to a different task. Participants were free to explore the scenes for a few seconds before being probed by a region from the image. Differences in eye movements, therefore, could not be due to target guidance. Instead, particularly where consistent with Experiment 1, differences would indicate more general effects of scene structure on scanning.

As reported by Biederman (1972), and more recently by Velisavljevic and Elder (2008), participants were impaired at recognizing objects encountered in scrambled scenes. These studies found scene coherence effects on visual short-term memory, and they used brief scene exposures (e.g. 100–300 ms). Our experiment extended this research by giving participants several seconds in order to scan the scene, thus allowing us to see whether differences in the way that overt visual attention is allocated in normal and scrambled scenes can explain performance differences. This would not exclude the possibility that memory for the global context has an effect on processes at recall, or that this global context is available very quickly. Instead, the experiment asked whether scanning of scrambled scenes was different, and thus provides an opportunity to extend the results from the search task in Experiment 1 to a different viewing task.

There were several interesting effects of scrambling on eye movements. First, there was no effect on the latency of the first saccade, which is contrary to what was found (for coarse trials only) in Experiment 1. Across tasks, therefore, although scene structure may be perceived quickly, it does not have a clear effect on the time required to plan the first eye movement. It is worth noting, however, that initial saccades in this experiment were made more slowly than in search, and it may be that the differences on the first saccade emerge only when there is pressure on the participant to respond quickly.

Second, participants made more fixations in scrambled than in intact scenes, these fixations had a shorter average duration, and the saccades in between had a larger average amplitude. In this experiment, viewing duration was fixed, so these changes in scanning reflect an altered sampling strategy within this time limit. This was confirmed by the finding that observers in scrambled scenes showed greater coverage of the image; thus, their fixations were more

distributed. Together with the decreased fixation duration, which was also seen in scrambled scenes being searched in Experiment 1, these findings paint a picture of more erratic, and less focused, attention in scrambled scenes, which we suggest underlies the effect of context on both tasks.

Finally, analysis in this experiment confirmed a stronger relationship between saliency and fixation in scrambled scenes. When freely encoding an image prior to a memory probe, participants were more likely to look at high saliency regions than low saliency regions, at least in finely scrambled scenes. This confirms the pattern seen in Foulsham and Underwood (2007, 2008) that saliency is predictive of fixation in a memory-encoding task. However, this finding should not be overstated. The correlations between saliency and fixation probability were weaker than those with time to target in Experiment 1, and were nonsignificant in three out of four conditions. Therefore, although saliency does not appear to be a strong predictor of what gets looked at, this relationship is more pronounced when global scene structure is disrupted.

General discussion

In our present experiments, we replicated previous reports that disrupting scene structure through scrambling affects scene perception. The present work also provides a considerable advance over previous research by using eye tracking during two different tasks to look at how scene structure affects the guidance of attention in natural scenes.

Before discussing the results, it is important to consider the validity of the stimuli and tasks used here for exploring visual attention in scenes. Participants in Experiment 1 searched for image patches, which contained an identifiable object alongside background and other visual information, and they did so in a display with high contrast grid lines superimposed. It is therefore important that the search behavior was entirely consistent with that shown in previous studies. For example, the manual RT (grand mean = 2.1 s) was roughly between the time participants took to search for categorically defined objects (pieces of fruit; 1.3 s) in Foulsham and Underwood (2007) and the time it took for people to find meaningless image patches in Pomplun's (2006) saccadic selectivity task (3.3 s). The fixation durations and saccade amplitudes were also within the range predicted by previous search tasks in scenes. Foulsham and Underwood (2011) used a search task for regions, in intact scenes without a grid, and found similar RTs and eye movement statistics. Probing memory for a patch has also been used previously (Velisavljevic & Elder, 2008), and our results from Experiment 2 add to what is known about this task. Therefore the results from these tasks are likely to generalize to search and encoding with normal scenes.

This is not to say that the grid had no effect on the deployment of attention. In fact, an unexpected result was that, within unscrambled images, a fine grid was more disruptive. In Experiment 1, a fine grid resulted in longer initial saccade latencies and fixations with longer average duration. In Experiment 2, where viewing was not guided by the prior knowledge of a target, grid size made no significant difference to these measures, but scene coverage was reduced in fine displays. The grid lines were not informative for the task, and were fixated only rarely. We consider their possible interaction with scrambling in subsequent sections, but why would fine versus coarse grid lines have an effect on normal scene viewing? By using two levels of grid, we demonstrated an increasing influence of dividing up the intact scene, and it is likely that this was caused by the low-level effect of introducing visual noise or clutter into the image. Specifically, the presence of additional lines may have masked the perception of features in the periphery. Alternatively, these additional lines would have competed as potential saccade targets. This is supported by the fact that the overall model-predicted saliency of these lines was greater, and they were fixated more often. In the time-pressured search task, masking or inhibition of the extra saccade targets would have prolonged the initiation of saccades, leading to the observed results on latency and fixation duration. A tendency to get distracted by these lines would also have reduced scene coverage, as was found in Experiment 2. Further research is necessary to confirm these explanations, but it should be noted that, because grid lines were present in all conditions and scrambling changed behavior within levels of grain, the presence of the grid cannot explain the effects of scrambling.

Our results imply four ways in which scene structure affects visual attention, to which we now turn.

Coherent scene structure guides visual search

Our results from Experiment 1 provide further evidence that when looking for a known target, a coherent visual scene provides global cues as to where an object is likely to be positioned. For example, when looking for the wall clock in the classroom scene in Fig. 1, once they have identified the scene gist (i.e., a classroom), observers can make some spatial predictions about where the target will occur (i.e., high on the wall, so probably in the upper half of the image). In addition, knowing the relative positions of different objects in the scene (e.g., chairs on the floor, a blackboard on the wall) will help finding the clock (e.g., because it is likely to be on the wall and therefore not on the floor but near the blackboard). When scrambling the elements disrupts the configuration of these objects, it is more difficult to use such cues. It may not be possible, or

will at least take longer, to recognize the gist or layout of a scrambled scene (Biederman et al., 1974). Thus, scrambling made people slower to move their eyes to targets and to make a manual response. This evidence is concordant with the finding that objects placed in unlikely positions (e.g., those that violate their likely location in a scene; Biederman, Mezzanotte, & Rabinowitz, 1982) or in out-of-place contexts (e.g., a microscope in a cocktail bar; Henderson et al., 1999) are harder to find.

By using eye tracking, we were able to confirm that scrambling affects the time taken to move the eyes to the target region, and not only the time to recognize it and make a manual response. This longer search time also meant that there were more fixations in scrambled trials, and together the implication of these measures is that removing scene structure impaired target guidance. Because the local visual properties of the target were the same across normal and scrambled conditions, as was its location in space, this guidance must entail more than just moving toward target features or likely locations in the visual field. Instead, we propose that scrambling interfered with both the configural relationships between objects in the scenes, and with the early perception of scene gist (and thus with expectations about where a target will be).

Why did the scrambling effect interact with grain? RTs and localization performance was impeded for scrambled scenes in finely divided scenes only. Moreover, the number of fixations and the time taken to first fixate the target showed the same pattern (although there remained a nonsignificant numerical advantage for coarse normal scenes). To put this another way, scene guidance was relatively preserved in coarse scrambled scenes, despite the large change in scene structure. This is most likely because coarse scrambling left more of the scene structure intact, so high levels of performance could be maintained. Support for this comes from the observation that the time to fixate the target was still impaired in coarse scrambled trials, though to a lesser extent that did not translate to slower RTs. A second unexpected interaction emerged with the initial saccade latency in Experiment 1, which scrambling delayed the initiation of the first saccade in coarsely divided images, but not in finely divided ones. Scrambling and grain had no significant effects on the first saccade in Experiment 2, so we interpret this result with caution. In search, there was significant time pressure to begin exploring the scene, and we have already discussed the possible influences of the additional grid lines in the fine displays on the first saccade. Both scrambling and grain thus affected the initial saccade latency, but the effects were not additive, perhaps because of an upper bound on this measure (i.e., a ceiling effect). By this account, saccade dynamics, or more likely the time pressure in the search task, meant that although saccades in the fine scrambled condition should have been the slowest

to be initiated, they were curtailed by a “deadline” of around 250 ms. One limitation of this study is that the grain manipulation changed the size of the target and the number of distractor regions, as well as the degree of spatial coherence. Future experiments could tease these factors apart by keeping the size of the grid constant while varying the coherence (e.g., by perturbing the structure to a lesser or greater extent).

Our results add to what has already been learned about the sorts of contextual cues that configural or gist information might provide for search in natural scenes (Foulsham & Underwood, 2011; Henderson, 2003; Vo & Henderson, 2010). Building this contextual information (e.g. the knowledge that people tend to be at street level) into a model of active visual search in natural scenes boosts model performance, providing a better simulation of what humans do when searching (Torralba, Oliva, Castelano, & Henderson, 2006).

Coherent scene structure promotes the allocation of fixation to meaningful regions

In both of our experiments, scrambling resulted in a change in the way that fixations were distributed in space and time. As we have seen, search in scrambled scenes was prolonged and required more fixations. However, the more extensive search afforded by these multiple fixations was offset by the fact that their durations were shorter, on average, than those in normal images. Interestingly, we also found this pattern in Experiment 2, in which participants had a fixed trial time during which to scan the scene. In this experiment, participants viewing scrambled images made more frequent, shorter fixations than did those viewing intact scenes.

In many information-processing tasks, longer fixation durations are associated with increased task difficulty and more extensive processing at fixation. For example, in reading, low-frequency words are fixated for longer than high-frequency words (see Rayner, 2009, for a review of this and other effects on fixation duration). In scene perception, objects that are inconsistent with their context (or more informative) are fixated for longer (Henderson et al., 1999; Underwood & Foulsham, 2006; Vo & Henderson, 2009), as are scene regions that have been distorted (e.g., those that are low-pass filtered to remove fine detail; Mannan, Ruddock, & Wooding, 1995). A recent computational model of fixation durations in scenes has also begun to elucidate the implications of such results for eye movement control (Nuthmann, Smith, Engbert, & Henderson, 2010).

Given that longer fixation durations are associated with increased visual or cognitive processing at the fixated location, why did scrambled scenes (which were more

difficult to search and memorise) elicit *shorter* fixations? We hypothesize that this is because scrambling interferes with the locations that were selected by fixation. Because of the disruption of scene context, the locations were more spread out, and this was confirmed by our finding that fixations covered more of the display area in scrambled scenes. As a consequence, some fixations would have targeted regions that were not prioritized in normal viewing (such as the high contrast but meaningless edge in Fig. 4), and were therefore not as informative or meaningful for the task at hand. Once at this location, participants may have quickly moved away to a different location, thus leading to short fixation durations. Due to the disruption of scene structure, it may also have been harder to process nearby regions with parafoveal or peripheral vision. Scrambling thus led to more brief fixations in less concentrated locations. Conversely, coherent scene structure provides gist and layout cues, which allow people to spend more time fixating meaningful locations. A question for further research is whether this difference in scanning is strategic and under conscious control, or whether it emerges implicitly from properties of the display.

Coherent scene structure overrides the influence of visual saliency

Much of the recent research into eye movements in complex scenes has made the distinction between bottom-up guidance (which depends only on the visual features of the stimulus) and top-down guidance (which depends on cognitive representations of the situation or task). This distinction has been brought into the foreground by demonstrations that a computational model of visual saliency can predict where people look on the basis of bottom-up feature contrast, at least in some circumstances (Itti & Koch, 2000; Parkhurst et al., 2002). Since these early reports, however, a host of other studies have shown that fixation locations are in fact selected top-down, according to their meaning or relevance. For example, Foulsham and Underwood (2007) demonstrated that search targets are fixated early and often, seemingly regardless of their saliency. Henderson, Brockmole, Castelano, and Mack (2007) found that fixated locations tended to be semantically informative, and that although this sometimes correlates with visual saliency, cognitive control is the dominant factor in eye guidance. There also seem to be some objects that are extremely likely to be fixated, despite being low saliency, presumably because of their social or situational relevance (e.g., the faces and eyes of others; Birmingham, Bischof, & Kingstone, 2009).

If there is a balance, then, between bottom-up and top-down factors in scene viewing, the latter seem to dominate. But does visual saliency contribute anything to where we

look, or do its effects arise only because it is correlated with scene meaning? We have previously presented evidence that, in fact, when top-down information is disrupted or unavailable, saliency provides a better description of fixation position. In Foulsham, Barton, Kingstone, Dewhurst, and Underwood (2009), we described the eye movements of a neuropsychological patient with visual agnosia—a profound deficit in recognizing visual objects. When looking for an object in a real world scene, she was unable to find the target. Moreover, rather than just scanning randomly, she was selectively biased to look at salient objects. Thus, when this patient could not ascertain the meaning of the different parts of the scene she reverted to saliency-based guidance.

In the present study, we found that, in normal scenes, saliency did not correlate well with the speed of fixating a target (Experiment 1) or the likelihood of a region being looked at (Experiment 2). However, this correlation was stronger, and only statistically significant, in scrambled scenes, in both tasks. A major, and novel, implication of the present research, therefore, is that the effect of saliency can actually get larger when top-down guidance is disrupted. In Torralba et al.'s (2006) contextual guidance model, visual saliency is modulated by two top-down pathways. In one, target features are compared to local image features to determine which image regions are likely to contain the target, a process that has been modeled computationally by Rao, Zelinsky, Hayhoe, and Ballard (2002). In the other top-down pathway, analysis of global image features generates a description of the scene gist. Through learning, the gist is associated with a spatial prior representing the likely location of target objects, and this prior modulates visual saliency. Salient items that are unlikely to be the target are vetoed in favor of likely target locations. Our results can be explained by assuming that scrambling interferes with this global pathway, meaning that prior scene expectations are not available and that saliency is not overridden to the same degree as in normal scenes. The local image information, and therefore the local pathway of the model, remains intact in scrambling, and so it is possible that observers are still guiding search top down on the basis of target features. This may explain why the correlation between fixation and saliency is weak, even when significant. In terms of the comparison between viewing of scrambled scenes and the behavior of a patient with visual agnosia from Foulsham et al. (2009), although top-down guidance may be impaired in both cases, the eye movement results indicate a rather different mode of processing in each study. Specifically, the agnostic patient made longer fixations and short saccades, consistent with focal processing and an effortful interpretation of the objects that she had difficulty recognizing. Scrambling, on the other hand, led to *shorter* fixations and longer

saccades. We discuss the implications of this in the final section.

Coherent scene structure promotes a focal mode of processing

One way to interpret the effect of scrambling on scanning eye movements is in terms of the previously proposed distinction between ambient and focal processing (Pannasch et al., 2008; Velichkovsky et al., 2005). Velichkovsky et al. suggest that a systematic combination of fixation durations and saccade amplitudes is indicative of one of two types of processing. Ambient processing, consisting of relatively brief fixations and saccades of more than about 5° (and thus targeting items beyond the parafovea) is proposed as an exploratory process linked to the dorsal visual stream. Focal processing, meanwhile, is characterized by longer fixations and shorter saccades, and by cognitive processing of objects associated with the ventral visual stream. In the context of scene perception, Unema, Pannasch, Joos, and Velichkovsky (2005) observed a shift from ambient to focal processing after 2–4 s of viewing. Furthermore, oculomotor behavior associated with focal processing seems to be related to improved encoding of the items being looked at (Velichkovsky et al., 2005).

In the present study, we found a pattern of fixations and saccades consistent with the idea that viewing of scrambled scenes is comprised more of ambient processing than is that of intact scenes. In search, fixations in scrambled scenes were shorter, a trend that was also seen in Experiment 2. Scrambled saccades were paradoxically larger than those in intact images, in both experiments and despite the increased task difficulty in these stimuli. The distribution of saccades in scrambled scenes showed that the majority moved beyond the parafovea, and it is interesting that in search the division in amplitude between normal and scrambled images was almost exactly 5°, which Velichkovsky et al. (2005) proposed as the cut-off between focal and ambient processing. The poorer task performance in scrambled scenes could also be attributed to a relative absence of focal processing. In Experiment 2, we found poorer recognition for targets from scrambled scenes, as did Velichkovsky et al., in recognition of normal scenes that had been subject to ambient processing.

These results suggest that the effect of scrambling is linked to a change in the mode of processing. If viewing shifts gradually from ambient to focal processing, scrambling may delay this shift. Conversely, we can conclude that an intact scene promotes focal processing by guiding the eyes and reducing the need for exploration. Further research is necessary to establish whether the distinction between processing modes is merely descriptive of a continuum of attention, or evidence for two separate processes and neural

substrates. Our results identify the scrambling effect as a useful paradigm for this investigation.

In sum, we have used the scrambling effect in both search and memorization tasks to explore the impact of scene structure on active visual attention. Our results provide clear evidence that this structure influences visual attention by guiding the eyes to search targets and meaningful regions, by overriding bottom-up saliency, and by promoting focal processing of the scene.

Author Note We are grateful for the very helpful suggestions of Lester Loschky, Melissa Vo, and an anonymous reviewer on a previous version of this manuscript.

References

- Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M.... (2006). Top-down facilitation of visual recognition. *Proceedings Of The National Academy Of Sciences*, 103, 449–454.
- Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177, 77–79.
- Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 97, 22–27.
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception—Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143–177.
- Biederman, I., Rabinowitz, J. C., Glass, A., & Stacy, E. (1974). On the information extracted from a glance at a scene. *Journal Of Experimental Psychology-General*, 103, 597–560.
- Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, 49, 2992–3000.
- Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal Of Experimental Psychology: Human Perception & Performance*, 33, 753–763.
- Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, 46, 4118–4133.
- De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52, 317–329.
- Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual saliency in scene perception? *Perception*, 36, 1123–1138.
- Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal Of Vision*, 8, 1–17.
- Foulsham, T., & Underwood, G. (2011). If visual saliency affects search then why? Evidence from normal and gaze-contingent search tasks in natural scenes. *Cognitive Computation*, 3, 48–63.
- Foulsham, T., Barton, J. J. S., Kingstone, A., Dewhurst, R., & Underwood, G. (2009). Fixation and saliency during search of natural scenes: The case of visual agnosia. *Neuropsychologia*, 47, 1994–2003.
- Foulsham, T., Teszka, R., & Kingstone, A. (2011). Saccade control in natural images is shaped by the information visible at fixation: Evidence from asymmetric gaze-contingent windows. *Attention, Perception, & Psychophysics*, 73, 266–283.
- Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In R. van Gompel, M. Fischer, W. Murray, & R. W. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam, the Netherlands: Elsevier.
- Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends In Cognitive Sciences*, 7, 498–504.
- Henderson, J. M., Brockmole, J. R., Castelano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. W. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam, the Netherlands: Elsevier.
- 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 & Performance*, 25, 210–228.
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36, 791–804.
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489–1506.
- Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal Of Experimental Psychology: Human Perception & Performance*, 4, 565–572.
- Loschky, L. C., & McConkie, G. W. (2002). Investigating spatial vision and dynamic attentional selection using a gaze-contingent multiresolutional display. *Journal Of Experimental Psychology: Applied*, 8, 99–117.
- 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), 1–13.
- Mannan, S., Ruddock, K., & Wooding, D. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision*, 9, 363–386.
- Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, 117, 382–405.
- Pannasch, S., Helmert, J. R., Roth, K., Herbold, A., & Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, 2, 1–19.
- Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42, 107–123.
- Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, 46, 1886–1900.
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2, 509–522.
- Rao, R. P. N., Zelinsky, G. J., Hayhoe, M., & Ballard, D. (2002). Eye movements in iconic visual search. *Vision Research*, 42, 1447–1463.
- Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal Of Experimental Psychology*, 62, 1457–1506.
- Tatler, B. W. (2009). Eye guidance in natural scenes [Special issue]. *Visual Cognition*, 17(6/7).
- Torralba, A., Oliva, A., Castelano, 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.
- Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97–136.
- Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruity influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, 59, 1931–1949.

- 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., & d'Ydewalle, G. (2003). Early peripheral and foveal processing in fixations during scene perception. *Visual Cognition*, 10, 79–100.
- 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. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the XXVII Conference of the Cognitive Science Society* (pp. 2283–2288). Mahwah, NJ: Lawrence Erlbaum.
- Velisavljevic, L., & Elder, J. H. (2008). Visual short-term memory for natural scenes: Effects of eccentricity. *Journal Of Vision*, 8(16), 1–17.
- Vo, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal Of Vision*, 9(3), 1–15.
- Vo, M. L. H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal Of Vision*, 10(3), 1–13.
- Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19, 1395–1407.
- Wolfe, J. M. (1994). Guided Search 2.0—A revised model of visual-search. *Psychonomic Bulletin & Review*, 1, 202–238.
- Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, 115(4), 787–835.
- Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research*, 50, 2062–2068.