What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition

Tom Foulsham
School of Psychology, University of Nottingham, Nottingham, UK

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

Saliency map models account for a small but significant amount of the variance in where people fixate, but evaluating these models with natural stimuli has led to mixed results. In the present study, the eye movements of participants were recorded while they viewed color photographs of natural scenes in preparation for a memory test (encoding) and when recognizing them later. These eye movements were then compared to the predictions of a well defined saliency map model (L. Itti & C. Koch, 2000), in terms of both individual fixation locations and fixation sequences (scanpaths). The saliency model is a significantly better predictor of fixation location than random models that take into account bias toward central fixations, and this is the case at both encoding and recognition. However, similarity between scanpaths made at multiple viewings of the same stimulus suggests that repetitive scanpaths also contribute to where people look. Top-down recapitulation of scanpaths is a key prediction of scanpath theory (D. Noton & L. Stark, 1971), but it might also be explained by bottom-up guidance. The present data suggest that saliency cannot account for scanpaths and that incorporating these sequences could improve model predictions.

Keywords: saliency, eye movements, scene perception, scanpaths


Introduction

Central to many modern theories that seek to explain guidance of eye movements is the concept of a saliency map (Findlay & Walker, 1999; Koch & Ullman, 1985). This is posited to be a pre-attentive representation that explicitly codes the importance of each part of an image or a scene. Shifts of attention and saccadic eye movements are initiated toward the point with the highest saliency, which is then inhibited so that attention can be disengaged and be moved to the next most salient location. In this way, the saliency map provides a control mechanism for dynamically targeting eye movements.

Bottom-up saliency models specify early visual filters that quantify the visual conspicuity of each part of the scene in terms of feature contrast with its surroundings. These models suggest that low-level feature discontinuities represented in the saliency map can explain a significant proportion of where people look. This is supported by studies which show that measures such as local contrast correlate with fixation location (Reinagel & Zador, 1999).

An alternative to this “scene statistics” approach is to test how well model-predicted fixations correspond to human fixation locations. The most widely used model is the computational model of saliency developed by Itti and Koch (2000, 2001), which analyses natural images in terms of intensity, color, and orientation. Within these three channels, filters extract the feature strength at several scales and combine them in a centre-surround scheme that highlights parts that stand out from their background. When summed into a single saliency map, the values from the map, in combination with a winner-take-all network and inhibition-of-return, produce a sequence of predicted fixations that scan the scene in order of decreasing saliency.

Several studies have shown this model to be a useful predictor of where people fixate in natural scenes. Itti and Koch (2000) themselves found a close match between the number of fixations the model predicted in a complex search before reaching the target and the time taken by real observers. Parkhurst, Law, and Neibur (2002) confirmed that saliency at fixation was significantly greater than chance. Adding more realistic features (such as a simulated retina with declining contrast sensitivity at greater eccentricity) makes the saliency map model more
predictive, even in dynamic scenes, than the standard model and greater than chance (Itti, 2006). More recently, several authors have argued that saliency can explain little or none of the variance in fixation locations in natural tasks such as walking (Jovancevic, Sullivan, & Hayhoe, 2006; Turano, Geruschat, & Baker, 2003). In realistic visual search, top-down guidance (for example, knowledge of a target’s appearance) tends to dominate, and participants are rarely distracted by visually salient but irrelevant items in these tasks (Chen & Zelinsky, 2006). Recent updates of the model include top-down information (Navalpakkam & Itti, 2005). It is true that even when free-viewing static scenes, the correlation between saliency and fixation locations tend to be small; Parkhurst et al., for example, estimate mean correlations of 0.55 and 0.45 for fractals and natural photographs respectively.

An important point when considering these results is that they are based on correlations. It is unclear to what degree fixations are actually caused by visual saliency. In a landscape photograph, for example, people may fixate the horizon due to saliency capturing their attention or due to top-down biases and “gist” which include the assumption that informative objects such as people and buildings tend to be located near the horizon. One way to unravel these factors is to take a more experimental approach. For example, in Foulsham and Underwood (2007) and Underwood and Foulsham (2006), we measured attention to critical objects in photographs that were manipulated to have differential saliency. In these experiments, the task being performed was important: Saliency was only a systematic bias in eye movements. In this study, we investigate such biases in more detail.

The Itti and Koch (2000) model treats all parts of the visual scene equally when computing saliency (although see Itti, 2006). The exception to this is that the previously selected location is inhibited, making it much less likely to be fixated again. This “inhibition of return” (IOR) also slightly enhances regions near to this location (due to the excitatory lobes on the IOR kernel), making small shifts more likely than large ones. Human eye movements may contain other biases, such as a left-to-right pattern of scanning, which might produce artefactual correlations. On the other hand, perhaps by building in relatively simple spatial or sequential biases, saliency map models might be able to account for much more of the variance in attentional allocation. For these reasons, the current paper looks at both individual fixation locations and fixation sequences (scanpaths) and compares them to those generated by the model.

The comparison of visual scanpaths is an important part of “scanpath theory” (Noton & Stark, 1971; Spitz, Stark, & Noton, 1971). This theory argues that eye movements are generated top-down, particularly in response to a previously seen image. Demonstrations that the scanpaths made by a viewer while encoding and recognizing the same image are similar have been used to argue that the scanpath made is encoded along with the visual features it explores (Stark & Ellis, 1981). Additional support for scanpath theory comes from two studies reporting similarity between scanpaths made while viewing stimuli and those made when imagining them later (Brandt & Stark, 1997; Laeng & Teodorescu, 2002). Scanpath theory suggests that scanpaths are generated almost completely top-down, a product of the mental model of the observer. In contrast, the saliency map model argues that scanpaths can be explained in terms of bottom-up discontinuities and some simple selection processes that move between these salient points. The specifics of scanpath theory have rarely been replicated, and this has led to a decline in interest in studying fixation sequences (Henderson, 2003). One of the problems with scanpath theory was that it had difficulty accounting for the variability in scanpaths across viewings and observers. A strong form of scanpath theory would predict identical eye movements on the second exposure to an image, but it is interesting that a completely bottom-up saliency model would predict the same thing. If we assume that there is some variance between people and viewings (that is that neither model can fully account for eye movement sequences), then it becomes important to ask how much of the scanpaths people make are explained by previous viewings or saliency. This study will investigate the relationship between scanpaths at encoding and recognition. It is an interesting question whether, if saliency models can explain a sequence of fixations on a scene, they might also be able to explain repetitive sequences in response to the same image. In this case,
similar scanpaths are expected, not because they are encoded in any way, but because the image and the saliency map are constant. The saliency map model is sophisticated enough to predict actual scanpath sequences, and our novel contribution is to use these predictions to try and tease apart any effects of saliency or scanpath repetition.

The effect of memory or prior experience on eye movements has also been discussed by Althoff and Cohen (1999). Their “eye-movement-based memory effect” was used as evidence of a global change in the sampling of visual information when scanning faces. This showed up as changes in the regions fixated when viewing famous versus non-famous faces and also in measures of the degree to which fixations were sequentially constrained. In particular, the scanpaths made when viewing famous faces were less constrained or systematic (as assessed by the first- and second-order transitions between regions) than those made when viewing non-famous faces. Ryan, Althoff, Whitlow, and Cohen (2000) subsequently investigated normal subjects and patients with amnesia viewing photographs of scenes and found decreased sampling of repeated scenes, independent of explicit awareness of these scenes. The current paper differs from this work in several respects. The emphasis here is on visual saliency, and hence we are able to explore the consequences which the saliency distribution might have on scanpaths at repeated viewings. The scanpath measures taken concentrate on similarity between viewings rather than a general assessment of constraint or systematicity. As memory is not the primary concern, memory performance and the time-course of recognition and any implicit effects on eye movements will not be discussed in detail. On the other hand, if prior exposure does produce a “reprocessing effect” (Althoff and Cohen, 1999), then the sampling of visually salient regions should change. Althoff and Cohen (1999) suggest that the shift in processing found when viewing a face which is familiar is designed to be optimally efficient at extracting information from a novel environment. As far as a saliency map represents where such information lies, then saliency should be more correlated with fixation in novel pictures than in previously seen pictures.

In this paper, we record the eye movements of participants viewing photographs of natural scenes in preparation for a memory test. Although we have previously used this task merely to encourage scene exploration, here we also investigate viewing of both previously seen and novel stimuli during a recognition test. Primarily, this study tests the hypothesis that the Itti and Koch model can predict fixation locations and sequences of these fixations, over and above any systematic biases. A number of incidental questions also emerge. Does this relationship vary with the demands of the memory task (due to different top-down factors at encoding and recognition) or with novel and repeated stimuli at test (due to a repetition effect)? Alternatively, do the movements made at first viewing resemble eye movements on subsequent occasions and could a bottom-up model explain this?

**Method**

**Participants**

Twenty-one student volunteers with normal vision took part for payment. Inclusion in the study was contingent on reliable eye tracking calibration.

**Stimuli**

A set of 90, high-resolution color photographs of natural scenes was prepared as stimuli and sourced from both a commercially available collection, and photos were taken using a 5-MP digital camera. Figure 1 shows some examples. They were presented on a color computer monitor at a resolution of 1024 × 768 pixels. A fixed viewing distance of 60 cm gave an image that subtended 31° × 25° of visual angle. All the stimuli showed exterior and interior scenes featuring houses, landscapes, furniture, and other natural objects and were chosen to be similar to those within the same category. Of these pictures, half were designated “old” and shown in both encoding and test phases, whereas the other half will be referred to as “new” and were shown only at test. Each subject saw the same set of old and new pictures, in a random order.

Saliency maps were generated, using Itti and Koch’s (2000) model with standard parameters. These maps were produced for the first five simulated fixations and thus indicate the first five most salient regions for each picture (see Figure 2, top). The only further criterion for stimuli was that all five salient regions were non-contiguous; those pictures where the same or overlapping regions were re-selected within the first five fixations were replaced. In addition, a raw saliency map was computed for each picture to allow more comprehensive analyses of the saliency distribution present in the picture (see Figure 2, bottom, and results below for more details). These maps represent the combined conspicuity of the three feature channels, scaled to a fixed range of 0–255 and before control processes from the model (which favor a restricted number of saliency peaks and so promote a single “winner”) are implemented.

**Apparatus**

Eye position was recorded using an SMI Eyelink system, which consists of a head-mounted camera that samples pupil location at 250 Hz. A 9-point calibration
and validation procedure was repeated several times to ensure that all recordings had a mean spatial error of less than 0.5°. The system parses samples into fixations and saccades using criteria of displacement (saccades >0.15°), velocity (>30°/s), and acceleration (>4000 °/s²) between samples. Head movement was restricted using a chin rest.

Procedure

Following calibration, participants were shown written instructions telling them to “inspect the following pictures in preparation for a memory test.” In a practice phase designed to familiarize participants with the equipment, the displays, and the task, they were shown a set of six photographs with similar characteristics as the test set. They then viewed the same set, mixed with six novel photographs, in a random order. In each case, they made a keyboard response to indicate “old” pictures they had seen before or “new,” unseen pictures.

Following the practice phase, the experiment proper began. In the “encoding” phase, all 45 encoding stimuli were presented in a randomized order. Each picture was preceded by a central drift-correct marker and a fixation cross, which ensured that fixation at picture onset was in the centre of the screen. Each picture was presented for 3 seconds, and participants were free to scan the picture, after which the picture was offset and the next trial began.

After all 45 encoding stimuli had been presented, the “test” phase began. During this phase, all 90 stimuli (45 encoding stimuli which were now “old” and 45 not previously seen) were presented in a random order in exactly the same way as in the encoding phase and participants responded old or new using the keyboard. In order to facilitate an ideal comparison between encoding and test phases, each picture was again shown for 3 seconds, and the response was recorded only if made during this period. This meant that although participants were encouraged to respond quickly, they
were able to inspect each picture for up to 3 seconds before responding, and the picture was not offset by their response.

**Analysis and results**

How well did the saliency map model predict fixations and scanpaths in the different task phases? In order to answer this question, we perform several different analyses. After describing the experimental data, we look at the proportion of all fixations that are targeted at any of the first five “salient regions.” Then, in order to look at model-predicted saliency values in more detail across the whole image and over multiple fixations, we look at the saliency value at fixation.

In both cases, the results need to be compared against chance. One way in which to do this is to compare experimental data against a random model. For example, if more human fixations than randomly generated fixations lie in salient regions, then this would suggest the visual system is selecting based on saliency. However, a uniformly distributed random model might lead to a difference purely due to the systematic bias in eye movements toward the centre. For this reason, a “biased random” model is also used where the random sampling of fixation locations is weighted by the spatial distribution found in the experimental data set. What other biases might affect the results? One way to look at the sequential aspects of the fixations made is to look at the first-order transitions between scene regions (see Descriptives section, and Figure 4). We therefore compute a third model, a “random transition” model. This model samples randomly based on the probability of moving to any one region from the current location. If participants tend to move in a clockwise direction, for example, this bias will be replicated in this model.

It is also desirable to look at full scanpaths and their sequence, rather than just individual fixations. Transition analysis is unsuited to this as the matrices involved explode exponentially when looking at sequences greater than two or three fixations. Instead we compare scanpaths using two different methods. A distance-based algorithm developed by Mannan, Ruddock, and Wooding (1995) for use with eye movements is used to quantify the spatial distance between two scanpaths, whereas a “string-editing” distance method from sequence analysis looks at temporal sequence. Two main comparisons are made. First, are scanpaths at encoding similar to those made when looking at the same pictures at test? Second, are fixation sequences produced by the saliency model similar to experimental scanpaths? As with the other analyses, a number of control comparisons can be computed, which give an indication how much of any correlations are due to systematic biases.

**Response data**

In the test phase, participants responded to indicate whether the current stimulus was one they had seen previously. This task was performed well by all participants (mean percentage of correct trials, 77%, $SD = 12.6$). Those trials that led to errors are excluded from all further analyses. Although it would be interesting to look at incorrect trials in further research (in the hope of relating fixation data to response accuracy), this was not appropriate here for two reasons. First, this comprised less than a quarter of the data; and given the complexities of the analysis, we were concerned about the statistical power available. Second, participants’ behavior in these cases is likely to be heterogeneous and could be interpreted as a true failure of recognition or as a result of confounding activity, boredom, fidgeting, or any number of (potentially idiosyncratic) reasons.

The mean correct reaction time was 1459 ms ($SD = 202$) and 1470 ms ($SD = 212$) for old and new pictures.
respectively. This indicated that although the picture remained on the screen for 3000 ms, an average of approximately five or six fixations were made before responding. As there may have been a change in scanning goals after responding, the eye movement measures looked principally at the first few fixations (except where mentioned otherwise).

**Fixation data**

The raw eye movement data showed fixation locations and durations for each participant on each picture. In all cases, trials were excluded where the fixation at picture onset was not within 1° of the central region, which was ensured by presentation of the drift correct marker. The mean proportion of non-centered trials removed for this reason was 7.8% (SD = 9.1). The first fixation was imposed by the experiment and hence was excluded from all further analysis. Unless otherwise stated, all statistical tests were repeated measures ANOVAs performed on the participant means, with post hoc t-tests (and Bonferroni correction) performed if necessary.

**Descriptives**

The locations of all fixations (excluding the first, which was necessarily in the centre) are plotted in Figure 3 and show a clear tendency for people to fixate in the centre of the display. This is not surprising, especially given that viewing started here, but reinforces the need to take into account this bias when evaluating model predictions. The first five most salient regions, across all stimuli, are also shown, and these are more distributed. In addition, to give an idea of any simple sequential patterns made by subjects, we calculated the first-order transition matrix for all fixations in the experiment. To do this, the display was divided into a grid of 5 × 5 regions (this number achieved a balance between spatial resolution and ease of computation). The matrix shows the probability of moving from each of these 25 regions to any other and is depicted by a contour plot in Figure 4. The high probability along the leading diagonal indicates an increased likelihood to make small movements within the same region. People were also more likely to move from all regions into the center-most region, which is in agreement with the overall tendency to fixate there. Figures 3 and 4 were computed from large numbers of fixations (N = 27,252) collapsed across all stimuli, participants, and task phases. These descriptives make it clear that, across tasks and images, the visual scene is not sampled uniformly. In the next subsection, three random models are explained which take account of this sampling.

Table 1 summarizes measures indicative of the general viewing behavior at encoding and recognition. A repeated measures ANOVA with three levels (encoding, old test items, and new test items) showed no effect of task phase on the number of fixations, $F(2,40) = 2.32$, $MSE = 0.36$, $p = .14$, although the effect on mean fixation duration approached significance, $F(2,40) = 3.65$, $MSE = 632$, $p = .057$. New items at test elicited fixations that were around 15 ms shorter on average than those on old items at encoding or at test.

**Random models**

In order to assess the predictions of the saliency model, three random data sets were generated containing the same number of fixations as the experimental data. The first set, “random,” was produced by a uniformly distributed model which gave a random $x,y$ coordinate for each fixation location. The second random model, “biased random,” weights the fixation location so that some locations are more likely than others. Specifically, the display was split into a grid of 25 sections, and the probability of any one section being selected by the biased random model was equivalent to the proportion of experimental fixations in this region. Placement within a grid section was fully random. This model reproduces the central bias seen in Figure 3, and each modelled fixation is independent of the previous fixation. A biased random model was also used by Henderson, Brockmole, Castelhano, and Mack (2007). A third model, “random transitions,” duplicated the between-fixation regularities depicted in Figure 4. In this model, the probability of a region being selected depended on the previously selected location (for the very first fixation, this was the image.
centre). In each case, the full range of possible regions was sampled based on the experimental probabilities of moving to each region from the previous one.

The use of these random data sets allows a more detailed test of the saliency map model. If human fixations are more likely to be targeted at salient regions than random fixations, then this might be due either to general eye movement biases or to image specific bottom-up guidance. However, if this effect still remains when compared with the biased random and random transition models, it cannot be attributed to general top-down biases, allowing a more confident endorsement of the saliency model. As mentioned previously, turning a correlational observation that fixations and saliency coincide into a causal statement that saliency causes fixation is problematic. Figure 3 shows that, across all images, fixations are not clearly tied to saliency but the remaining analyses will evaluate this more closely. The use of these random models to control for any across-image biases is a conservative effort to test for genuine saliency effects. However, being a correlational study the flip side of this is that if these biases are actually caused by regularities in the distribution of saliency (because photographers place objects in the centre, for example) and are not just incidental, these analyses may underestimate the role of saliency.

### Proportion of fixations on salient regions

As a first step to evaluating the saliency model, fixations were labeled as those on salient regions and those on non-salient regions. There were five non-contiguous salient regions on each picture, corresponding to the first five selected points in the Itti and Koch (2000) model simulation (see Figure 2, top). These regions were defined as all points within 2° of the salient midpoint indicated by the model. The size of this region was chosen to reflect estimates of the size of the fovea and also the “focus of attention” parameter used in Itti and Koch, which determines the region that receives IOR. A smaller region might give fewer salient fixations, but any noise arising from this decision will be replicated in the random

<table>
<thead>
<tr>
<th>Number of fixations</th>
<th>Average fixation duration (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Encoding</td>
<td>10.59</td>
</tr>
<tr>
<td>Old items at test</td>
<td>10.47</td>
</tr>
<tr>
<td>New items at test</td>
<td>10.78</td>
</tr>
</tbody>
</table>

Table 1. Means and standard deviations for general scanning behavior in the different task phases.
models. For each trial, the number of salient fixations was compared to the total number of fixations made. The resulting proportion of salient fixations is shown in Figure 5 for experimental data as a function of task phase and for the random data sets. Across both encoding and recognition, an average of 20% of all fixations landed on salient regions. The standard five salient regions comprised 10.4% of the area of each picture, and the random model produces salient fixations close to this rate.

There was a reliable effect of task phase, $F(2,40) = 75.6$, $MSE = 0.00015$, $p < .001$. A higher proportion of fixations were on salient regions in old trials than in either encoding ($t(20) = 8.71$) or new trials ($t(20) = 13.82$; both $p < .001$). Independent $t$-tests then compared each task phase with each of the random models. In all cases, the comparisons were highly reliable: for random, biased random, and random transitions in order, at encoding ($t(40) = 10.8$, 7.1, and 8.1); for new pictures at test ($t(40) = 14.6$, 9.5, and 10.8), and for old pictures at test ($t(40) = 16.8$, 13.2, and 14.2; all $p s < .001$). Participants made more fixations in salient regions than any of the random or biased models. This is clear evidence that the saliency model is better than chance at identifying regions which will be fixated.

**Saliency at fixation**

The previous analysis looked only at the first five salient regions, and this involved a relatively small subset of the fixations made. An alternative is to look at the saliency map of the whole image that specifies a value for each part of the picture as opposed to just an ordinal rank. This allows the data from many more fixations to be analyzed and so should lead to a fuller description of the relation between fixation and saliency. This analysis is also less dependent on the control processes of the model that simulate a focus of attention and IOR. Perhaps the underlying saliency map rather than dynamic predictions of gaze shifts is more useful in predicting eye movements.

To begin with, values were extracted from a saliency map output from the model (see Figure 2, bottom). This map shows an intermediate stage in the saliency algorithm that highlights the combined conspicuity of all the features but is produced before any inhibitory processes. A map from this stage was chosen so as to give a larger variance of saliency values across the image (the evolving saliency map at a later stage would be likely to reduce most regions to zero and enhance only a few peaks). The model scales all the values to a fixed (and arbitrary) range between 0 and 255 and represents the map at 1/16 the size of the original image. It should be noted that there are arguments as to exactly how the saliency map should be normalized (see for example Parkhurst et al., 2002). Here, these values were considered to be representative of the general saliency-based model being tested.

Firstly, saliency values were extracted from fixation locations, scaled to 1/16, on the map for the corresponding stimulus. These were computed for the first five fixations following the first saccade after picture onset (note that this does not include the very first fixation, which was in the centre) and averaged across trials. Five fixations were a reasonable number to include as at least this many would have occurred on all trials, and the reaction time data indicate that on average this many fixations were made before responding in the test phases.

As previously, the same analysis was carried out using the randomly generated data sets. If the visual system is indeed selecting locations based on saliency, then the saliency values at fixated locations should be higher than
at randomly generated locations. Alternatively, if people fixate regardless of saliency then the saliency values should not differ between experimental and simulated data. Figure 6 shows the mean saliency value at fixated locations. This value did not differ between encoding, new and old trials, $F(2,40) = 2.70, \text{MSE} = 13.7, p = .08$. Accordingly, the mean saliency value at the first five fixation locations was collapsed across task phases, and the resulting grand mean was compared to the three random models. The participant grand mean ($112.5, SD = 2.93$) was reliably different from all three random models (independent $t$-tests; $t(40) = 38.8, 24.1, \text{and } 10.3$ for random, biased random, and random transitions respectively; all $p s < .001$). The results from this analysis therefore are entirely in agreement with those from the proportion of salient fixations; locations with high saliency received preferential fixation and biased random models did not fully capture this.

A supplementary question is whether this advantage changes over the course of scene viewing. Parkhurst et al. (2002) reported that the link between saliency and fixation decreased over the first few fixations. This was based on calculating a “chance-adjusted” saliency value, which took into account the (unbiased) mean saliency value. However, if one assumes that fixations are biased toward the centre, high chance-adjusted saliency will occur without any meaningful relationship. Furthermore, if the central bias of fixations decreases over time, as would seem logical if participants gradually widen their viewing, a drop in chance-adjusted saliency will occur artefactually (Tatler et al., 2005). Figure 7 shows the mean saliency value as a function of ordinal fixation number. A two-way, $3 \times 5$ repeated measures ANOVA was computed for these data and as before found no reliable effect of the task phase, $F(2,40) = 2.70, \text{MSE} = 95.5, p = .08$; saliency at fixation did not differ between inspection in the three task phases. There was a significant main effect of fixation number, $F(4,80) = 10.26, \text{MSE} = 54.9, p < .001$, which paired $t$-tests revealed to be due to saliency at the second fixation that was significantly higher than the third ($t(20) = 4.44$), fourth ($t(20) = 5.13$), and fifth ($t(20) = 4.95$; all $p s < .005$, see Figure 7). There was a marginally significant difference between the first and the fifth fixations ($t(20) = 3.17, p = .043$), but no other comparisons were different and there was no interaction between task phase and fixation number, $F(8,160) < 1, \text{MSE} = 48.5, p = .84$. 

Figure 6. The mean saliency value at the first five fixation locations for experimental (left) and randomly generated (right) data. Error bars indicate one standard error of the mean.

Figure 7. The mean saliency value in the three experimental phases, as a function of ordinal fixation number following the first saccade. Note that saliency remains high over several fixations, and that the task phases show a similar pattern. Error bars show one standard error of the mean.
Thus, although there is a slight suggestion of a downward trend, this is almost entirely due to a high mean saliency value at the second fixation location.

Scanpath data

The results so far have investigated the ability of the saliency map model to predict fixation locations, regardless of the order in which those locations are selected. An alternative way of looking at these locations is to consider the sequential scanpath which viewers make, and this has been of interest to several authors in visual perception (Noton & Stark, 1971; Yarbus, 1967). A particular point of interest has been whether individuals produce repetitive scanpaths on multiple viewings of the same image (Choi, Mosley, & Stark, 1995; Noton & Stark, 1971). If systematic repetitions in eye movements are indeed a feature of natural scene viewing then characterizing them could improve models, such as the saliency map model, which try to account for these movements.

In the next section, we first compare the sequence of fixations made by participants at encoding with the scanpath made when viewing the same stimulus at test. We then compare these scanpaths with predictions from the saliency model. The experimental data consisted of a variable-length sequence of fixations for each subject viewing each stimulus, giving a total of $135 \times 21 = 2835$ scanpaths (for some examples, see Figure 8 and Movie 1). For the saliency map model, the first five predicted locations gave a “saliency scanpath” for each image. This is the scanpath that would be expected if, as the model suggests, the scene is scanned in order of declining saliency value.

There is a certain amount of difficulty in quantifying the comparison between two scanpaths. This difficulty lies in condensing the spatial information of multiple fixations without losing the sequence information inherent in a scanpath. The most popular technique for quantifying the

![Figure 8](http://jov.arvojournals.org/) Example scanpaths from a single subject when viewing a stimulus during encoding (a) and at test (c). A novel stimulus from the same category (a street scene) is also shown (b). In each case, fixations are shown with circles (with diameter proportional to duration) and saccades with arrows. Scanning always started in the centre. (d) Saliency output for the same stimulus, indicating a predicted scanpath.
similarity of such sequences, and the first which is used here, is the Levenshtein or “string-edit” distance.

This technique is described in detail elsewhere (Brandt & Stark, 1997; Choi et al., 1995; Hacisalihzade, Allen, & Stark, 1992) and involves turning a sequence of fixations into a string of characters by segregating the stimulus into labeled regions. The similarity between two strings is then computed by calculating the minimum number of editing steps required to turn one into the other. Three types of operations are permitted: insertions, deletions, and substitutions. Similarity is given by one minus the number of edits required, standardized over the length of the string. The resulting similarity index gives a value between 0 and 1, where 1 indicates identical strings. An algorithm for calculating the minimum editing cost is given in Brandt and Stark (1997), and this was implemented in the present study using a program written in Java. The general scheme is illustrated in Figure 9a. A 5 × 5 grid was overlaid onto the stimuli. Choi et al. (1995) argue that this analysis is robust to changes in the number of regions used. The grid used here was decided on after pilot analyses showed that it segregated the saliency scanpaths efficiently. The resulting 25 regions (rectangles of approximately 6.4° × 4.8°) were labeled with the characters A to Y from left to right. Fixations were then labeled according to their spatial coordinates, resulting in a character string representing all the fixations made in this trial. The first fixation, which was always in the centre or region “M,” was removed. Adjacent fixations on the same regions were condensed into one. This was done as it is the global movements that are of interest here rather than the small re-adjustments that combine to give one gaze on a region.

How much similarity should be expected by chance? The probability that any two fixations will be on the same region is 1/25, although given the constraint that no two consecutive fixations, will be on the same region the actual chance similarity will be slightly higher. A simulation run using a modified version of the Java program compared one thousand randomly generated pairs of 5 letter strings and gave an average similarity of 0.0417. Of course, the spatial biases previously discussed will also increase the similarity of scanpaths. For this reason, a number of control comparisons are also computed which give a more useful baseline against which to measure similarity scores.

Although the string-editing procedure quantifies sequence similarity, it does so at the expense of spatial resolution because the display must be arbitrarily divided into regions. To examine the overall spatial similarity of different scanpaths, we use a method developed by Mannan et al. (1995) for use with eye movements (see Figure 9b). This method computes the mean linear distance between fixations in one scanpath, and their nearest neighbor in the other set. Henderson et al. (2007) refined this method slightly by adding the constraint that each fixation in a scanpath is assigned to only one other in the comparison scanpath.

This “unique-assignment” (UA) version ensures that the scanpath comparison is not disproportionately affected by differences in the overall distribution of fixations. This version therefore also requires that scanpaths have an equal number of fixations. Both the Mannan similarity metric and the UA version are standardized over the mean similarity of a randomly generated set of scanpaths with the same number of fixations as the tested set. This gives an index between 0 and 100, where 100 equals identical scanpaths and 0 equals scanpaths, which are no more similar than chance. A negative index indicates scanpaths that systematically differ.

All three comparison metrics (string-edit, Mannan, and UA) were used to analyze pilot data. The Mannan and UA metrics produced very similar patterns, and as a result, only the former is included in the remaining results.
Comparing scanpaths at encoding and test

Movie 1 shows example scanpaths of a participant looking at the same picture once at encoding and then again when recognizing it at test. The similarity here is striking, and the comparison metrics discussed above aim to quantify this across all subjects and trials. To do this, the scanpath for each subject viewing each image at encoding was compared to that from the same subject viewing the same (old) image at test. Figure 10 shows the mean similarity for both metrics.

There are several factors that might make two scanpaths from the same subject and stimulus on different occasions more similar than chance. If scanpaths are idiosyncratic, similarity might be due to the sequence being generated by the same person with a general scanning strategy rather than due to a sequence for that particular stimulus. In order to investigate this, each old test stimulus was arbitrarily paired with a new picture, and the scanpaths on each were compared. This comparison involves the same person performing the same task of recognition but with a different stimulus. Any similarity here therefore estimates the effect of a personal strategy and cannot be caused by stimulus memory or constant bottom-up factors (as the stimulus in this case is different). In addition, pictures shown at encoding were arbitrarily paired with new pictures at test. This case involves the same person but viewing different stimuli under different task conditions (trying to encode the details of a picture and trying to recognize a previously seen picture). Similar scanpaths here would indicate a participant’s similar scanning behavior independent of stimuli and specific task instructions. Hence, these control comparisons allow us to better interpret the similarity indices found.

There was a significant effect of comparison source on string-edit similarity (Figure 10a), $F(2,40) = 148$, $MSE = 0.00097$, $p < .001$. The similarity between encoding and old scanpaths was significantly higher than the similarity between old and new pictures ($t = 13.14$) and encoding and new pictures ($t = 15.27$, both $p < .001$). The latter two comparisons did not differ, although all comparisons were greater than the randomly generated estimate of 0.0417 (one-sampled $t$-test, all $t_s > 10$, all $p_s < .001$).

An almost identical picture emerges with the Mannan metric (Figure 10b). The three comparisons were reliably

---

**Comparing scanpaths at encoding and test**

Movie 1 shows example scanpaths of a participant looking at the same picture once at encoding and then again when recognizing it at test. The similarity here is striking, and the comparison metrics discussed above aim to quantify this across all subjects and trials. To do this, the scanpath for each subject viewing each image at encoding was compared to that from the same subject viewing the same (old) image at test. Figure 10 shows the mean similarity for both metrics.

There are several factors that might make two scanpaths from the same subject and stimulus on different occasions more similar than chance. If scanpaths are idiosyncratic, similarity might be due to the sequence being generated by the same person with a general scanning strategy rather than due to a sequence for that particular stimulus. In order to investigate this, each old test stimulus was arbitrarily paired with a new picture, and the scanpaths on each were compared. This comparison involves the same person performing the same task of recognition but with a different stimulus. Any similarity here therefore estimates the effect of a personal strategy and cannot be caused by stimulus memory or constant bottom-up factors (as the stimulus in this case is different). In addition, pictures shown at encoding were arbitrarily paired with new pictures at test. This case involves the same person but viewing different stimuli under different task conditions (trying to encode the details of a picture and trying to recognize a previously seen picture). Similar scanpaths here would indicate a participant’s similar scanning behavior independent of stimuli and specific task instructions. Hence, these control comparisons allow us to better interpret the similarity indices found.

There was a significant effect of comparison source on string-edit similarity (Figure 10a), $F(2,40) = 148$, $MSE = 0.00097$, $p < .001$. The similarity between encoding and old scanpaths was significantly higher than the similarity between old and new pictures ($t = 13.14$) and encoding and new pictures ($t = 15.27$, both $p < .001$). The latter two comparisons did not differ, although all comparisons were greater than the randomly generated estimate of 0.0417 (one-sampled $t$-test, all $t_s > 10$, all $p_s < .001$).

An almost identical picture emerges with the Mannan metric (Figure 10b). The three comparisons were reliably
different, $F(2,40) = 185, MSE = 18.6, p < .001$). Encoding and old scanpaths were more similar than the other two comparisons (old vs. new, $t = 11.65$; encoding vs. new, $t = 23.06$; both $ps < .001$). In addition, old and new scanpaths at test were more similar than encoding and new scanpaths ($t = 4.49, p = .001$).

**Comparing experimental and saliency scanpaths**

The scanpath comparison metrics provide a different way of evaluating the saliency map model. How well does this model account for the scanpaths generated by participants? The first five predicted fixations were taken as the “salient scanpath,” and all experimental scanpaths were trimmed to the same length. The results are shown in Figure 11. It should be noted that, unlike the saliency-based analyses already performed, this method explicitly takes into account the sequence information provided by the model rather than just the locations of independent fixations.

For the string-edit metric, there was no significant effect of task phase on similarity with the salient scanpath, $F < 1, p = .59$, indicating that any resemblance with the saliency model was constant at encoding and recognition. The similarity scores are noticeably lower than those in Figure 10 and are much closer to the chance estimate of 0.0417. The Mannan similarity scores are also lower for the saliency comparisons, although they are above zero.

---

Figure 10. Mean similarity scores (with standard error bars) for the three comparisons and for each metric. The scanpaths in each comparison are from the encoding (“enc”) phase or from “old” or “new” items in the test phase. The randomly estimated string-edit similarity is shown (dashed line in panel a); the equivalent value in panel b) is zero.

Figure 11. Mean similarity scores (with standard error bars) for the three task phases compared with the saliency scanpath and for each metric.
that indicates chance in this metric. This is not surprising as the analysis of fixations has previously shown that at least some of the fixated locations lie in salient regions and so should be spatially similar to the salient scanpaths. There was also an effect of task phase on similarity with salient scanpaths. Scanpaths while viewing new items at test were more similar to saliency model scanpaths than those in the other task phases (encoding, \( t(20) = 4.61; \) old \( t(20) = 4.49, \) both \( p < .001 \)).

**Discussion**

This experiment compared model-generated and experimental eye movements, both in terms of the spatial location of individual fixations and their sequential order. It aimed to clarify how well the saliency map model can predict fixation locations and scanpaths. In addition, it asked whether people repeat scanpaths on multiple viewings of the same stimulus. There were a number of interesting findings:

1. There was a tendency for fixations to target the most salient regions, as selected by the model. This could not be explained by biased models that took into account a central fixation distribution.
2. Across multiple fixations the saliency at fixated points was higher than predicted by chance and biased models. Saliency decreased slightly over multiple fixations, but this was mostly due to elevated values on the second fixation.
3. This link between saliency and fixation did not vary significantly with the demands of the task (encoding and recognizing).
4. Scanpaths were most similar when compared between two viewings of the same picture by the same person (encoding vs. old). This was reliable in terms of sequence, as evaluated by the string-edit distance, and linear distance. The similarity was significantly greater than control comparisons (encoding vs. new and old vs. new).
5. The saliency model-predicted scanpaths were not highly similar to human scanpaths.

The results reported here are largely in agreement with those elsewhere that show that high saliency is predictive of fixation (Foulsham & Underwood, 2007; Itti, 2006; Parkhurst et al., 2002; Underwood, Foulsham, van Loon, Humphreys, & Bloyce, 2006). This relationship was evident in multiple findings in the present study: Salient regions were fixated more often than chance, and saliency values at fixation locations were consistently higher than average. The saliency map model was a reliable way of identifying areas likely to be fixated, and it was much better than assuming a uniform distribution of attention. Importantly, it was also better than biased models that incorporated simple spatial and transitional biases. These biases explained more of the variance in where people look than purely random allocation, suggesting that these patterns need to be considered when looking at the relationship between saliency and fixation. Here, saliency had an effect over and above the central distribution of saliency and fixations. Other relatively simple biases, such as a predominance of horizontal saccades, might do better at explaining fixation locations, and this is impetus for further research.

Although this supports the general model, caution should be sounded regarding the correlational nature of these results. On their own, they cannot confirm that high saliency causes fixation or that the visual system is performing a saliency map-based computation. Participants may have fixated regions based on other, top-down or bottom-up, factors that happen to coincide with saliency. When using natural images, where total control over all visual and task factors is impractical, both picture manipulations (Foulsham & Underwood, 2007) and correlations are useful in testing a saliency map model. It should also be stressed that other, potentially simpler, bottom-up models that produce similar predictions to the saliency map model might also outperform chance. When added to the biased random model simple information on the location of edges, for example, might perform as well as the saliency map model and render some of its other aspects unnecessary. Although it is beyond the scope of this report to compute such possibilities, the relatively small difference between saliency model performance and chance is encouraging for those trying to explain this variance in other ways. For example, Tatler et al. (2005) find that edges and contrast are more related to fixations than luminance, whereas Jost, Ouerhani, von Wartburg, Murr, and Hugli (2005) quantify the contribution of color. Such research points toward variations in the feature channels which go into Itti and Koch’s (2000) model. Aside from saliency maps, Raj, Geisler, Frazor, and Bovik (2005) devised a model where fixation selection strives to minimize entropy or uncertainty in terms of local contrast. In theory, this would be an efficient way of gaining spatial information in a memory task such as that presented here.

There are two further questions regarding the role of saliency in this experiment. Firstly, did the effect of saliency decline over multiple fixations? Although this is a popular framework for bringing together bottom-up and top-down factors in scene viewing, such that bottom-up saliency effects are gradually overridden by task knowledge, the findings here were not conclusive. The general shape of the saliency function over the first five fixations suggests a decrease, but even on the fifth fixation, saliency was still higher than expected by chance. Although a significant change occurred, this was wholly as a result of one high value on the second fixation. It is possible that
the saliency map takes time to accumulate and therefore only becomes important at the second fixation, but analysis over longer durations would be needed to fully explore this effect. Tatler et al. (2005) have demonstrated that decreases in saliency over time can be an artefact of central biases in saliency and fixation. Of course if salient regions are inhibited, then shifts to less salient regions will become more frequent after the first few fixations. This study provides no clear evidence that there is any change in the bottom-up allocation of attention over scene viewing.

A second and related issue is whether saliency was predictive of fixation in all phases of the task. Henderson et al. (2007) argue that saliency cannot account for eye movements in a real-world search task involving counting the people in a scene, and Foulsham and Underwood (2007) and Underwood et al. (2006) found no experimental effect of saliency in search. In the current experiment, salient regions were fixated more often than chance, and saliency values at fixation were higher than chance expectancy in all task phases. Interestingly, there was evidence that this was even more significant when old pictures were seen again during recognition.

There was a small effect of repetition on the mean fixation duration at test, such that people made longer fixations when viewing pictures they had seen before (with a fixed viewing time, this implies fewer fixations as the two measures are inversely proportional, although this was not reliable). These observations replicate findings by Ryan et al. (2000) and so support the idea of memory-dependent changes in eye movements. Previous research has identified a reprocessing effect, which leads to optimized scanning of novel items in comparison to previously seen items (Althoff & Cohen, 1999; Ryan et al., 2000). This might suggest that salient regions would be more potent at attracting attention in new pictures, as saliency might highlight more informative regions, which would be more optimal to fixate. However, it was not the case that people looked at visually salient regions more often in new pictures than in old pictures at test.

The fact that saliency was predictive while preparing for a memory test is concordant with other results (Foulsham & Underwood, 2007) and is reasonable for a task in which there were no pressing task demands as to what to look at. The demands when recognizing pictures are quite different. It might be that participants are guided top-down to search out remembered features which could be defined, for example, by location within the picture. A raw saliency map would not trigger these top-down shifts. However, there was no evidence that the recognition task was less driven by saliency. This might suggest that top-down shifts at recognition were not common, or that when they did occur they were made to remembered regions based on saliency. The relationship with saliency when viewing pictures for the second time suggests that memory independent of saliency was not a significant factor in eye guidance. Of course, if the first viewing is largely driven by saliency, then the best remembered features will often be the most salient ones, making the two factors difficult to unravel in this experiment. It has also been suggested that scanning and the frequency of fixations in salient regions could be explained by the semantic meaning of these regions. The interaction of saliency and scene semantics has been explored in some recent research (Torralba, Oliva, Castelhano, & Henderson, 2006) and is an interesting avenue for further work.

It is worth noting that the saliency effects, although reliable, are not large. On average, only one of the first five fixations landed in a salient region, suggesting that there is plenty of variance in natural scene viewing unaccounted for. One aspect of oculomotor control that might explain some of this variance is temporal sequence information, and the present study explored this. Is the sequence of fixations made when inspecting a picture for the second time highly similar to that made on the first occasion? The present results suggest that scanpaths are indeed much more similar than would be expected from randomly generated sequences, and so support findings from simpler stimuli (Brandt & Stark, 1997; Noton & Stark, 1971).

Why is this the case? Scanpath theory (Noton & Stark, 1971) suggests that visual patterns are represented in memory as a network of features and the attention shifts between them. This network is then replayed and compared to the external stimulus when recognizing the image later. By this account, the scanpaths at recognition were similar to those at test because they were stored and recalled top-down. The similarity seen here, though reliable, is much less than previously reported (Brandt & Stark, 1997, report mean string-edit similarity as high as 0.75 in one subject), raising the question as to why there is still so much variance unaccounted for. Previous demonstrations of scanpath similarity have largely used simple patterns or line drawings, with fewer and larger regions of interest. It is likely that the much more complex photographs used here led to less scanpath repetition, possibly due to a greater influence of top-down scene knowledge. A closer analysis of the similarity seen here shows that a proportion of the resemblance can be explained by the consistent strategies and idiosyncrasy of individual subjects (as estimated by comparison with encoding and new picture similarity). These strategies need to be quantified further.

There remains significant similarity between first and second viewings, over and above this, that is to be explained. There are two different possible explanations. Similarity might be explained by a version of scanpath theory, or by some other top-down strategy. This strategy would need to be tied to the stimulus; otherwise, it would also lead to increased similarity between encoding and new trials. It is possible that fixation locations were ordered in terms of decreasing semantic relevance, perhaps in combination with scene gist. The debate as to whether semantic violations attract attention might also be
relevant to this possibility (Loftus & Mackworth, 1978; Underwood & Foulsham, 2006; Underwood et al., 2007).

Alternatively, scanpath similarity could be explained by bottom-up allocation. By this account, scanpaths are similar because in both encoding and recognition phases, fixation locations are at least partly determined by saliency, which remains constant over viewings. In principle, this seems likely as the analysis of fixation locations shows that participants were equally likely to target salient regions, which would remain the same, in encoding and old recognition items.

Top-down and bottom-up explanations produce the same prediction, partially similar scanpaths, and so are hard to distinguish from this alone. However, by using the saliency model and sequence analysis, we can test explicitly whether scanpaths resemble the saliency model’s predicted sequence at encoding and test. If this were the case, then there would be no need to posit top-down scanpath generation. However, model-generated scanpaths were not highly similar to those shown by participants. It appears that the string-editing method is particularly sensitive to sequence and that the model does not replicate the order in which salient regions are selected. This may explain the failure here to find a convincing change in saliency over multiple fixations. The similarity between scanpaths made at encoding and test was much greater than the comparison of either to the saliency model. Different models of bottom-up allocation might still explain repetitive scanpaths, but at the very least, this finding suggests that such sequences are important for modeling eye movements.

Conclusions

This study offers partial support for a saliency map model. The model was much better than uniform or biased random models in marking out fixation locations. However, the model was poor at predicting the order in which these locations were selected. Interestingly, scanpaths were partially repeated on multiple exposures to the same stimulus, and our analyses suggest this is not due to saliency. Models may be missing out on sequential aspects of oculomotor control that could potentially predict fixation much better than saliency alone.

Acknowledgments

We are grateful to the EPSRC for project award EP/E006329/1 to GU. We also thank Laurent Itti and colleagues for making the saliency map model available and Ben Tatler and two anonymous reviewers for insightful comments on a previous version of this paper.

Commercial relationships: none.
Corresponding author: Tom Foulsham.
Email: lpxtf@psychology.nottingham.ac.uk.
Address: School of Psychology, University of Nottingham, Nottingham NG7 2RD, UK.

References


