Everyone knows what is interesting: Salient locations which should be fixated

Christopher Michael Masciocchi Department of Psychology, Iowa State University, Ames, IA, USA

Stefan Mihalas Zanvyl Krieger Mind/Brain Institute and Department of Neuroscience, Johns Hopkins University, Baltimore, MD, USA

Derrick Parkhurst Thirty Sixth Span Internet Technologies, Oklahoma City, Oklahoma, USA

Ernst Niebur Zanvyl Krieger Mind/Brain Institute and Department of Neuroscience, Johns Hopkins University, Baltimore, MD, USA

Most natural scenes are too complex to be perceived instantaneously in their entirety. Observers therefore have to select parts of them and process these parts sequentially. We study how this selection and prioritization process is performed by humans at two different levels. One is the overt attention mechanism of saccadic eye movements in a free-viewing paradigm. The second is a conscious decision process in which we asked observers which points in a scene they considered the most interesting. We find in a very large participant population (more than one thousand) that observers largely agree on which points they consider interesting. Their selections are also correlated with the eye movement pattern of different subjects. Both are correlated with predictions of a purely bottom–up saliency map model. Thus, bottom–up saliency influences cognitive processes as far removed from the sensory periphery as in the conscious choice of what an observer considers interesting.

Keywords: attention, saliency, eye movements, fixations, interest points, interesting locations, model

Citation: Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision, 9*(11):25, 1–22, http://journalofvision.org/9/11/25/, doi:10.1167/9.11.25.

Introduction

Due to limitations of the human visual system, only a small amount of visual information can be fully processed at any given time. Light reflected from that area in a scene that is fixated will fall on the fovea, and thus be processed in the highest spatial detail. Visual attention can also be used to select objects or locations for preferential processing. While the exact function of attention differs according to various models, attention can generally be thought of as facilitating processing of certain areas or objects along with the inhibition of unattended regions.

An important question, then, is how attention is guided in artificial and natural scenes. This topic has been investigated using both response time and eye movement measurements. For instance, in visual search tasks participants are instructed to manually respond to a target presented among a number of distractors. Early studies demonstrated that response times are faster if the subjects are first cued to the location where the target subsequently appeared (Posner, 1980), and that they depend little on the number of distractors if the target is defined by a unique feature (Treisman & Gelade, 1980). It is generally agreed that items can be selected in two separate ways. The first is bottom–up attentional selection, which is a fast, automatic, stimulus-driven mechanism that operates based on the uniqueness or salience of an item's features. Irrelevant distractor items with a unique color (Theeuwes, 1992) or those that appear abruptly (Yantis & Jonides, 1984) have been shown to delay response time to the target, suggesting that those items capture attention automatically. However, response times to targets are sometimes not affected by the appearance of unique distractors, particularly when the target and distractors are defined by different properties (e.g., Folk, Remington, & Johnston, 1992), or when the features of the target and distractors are known ahead of time (Bacon & Egeth, 1994). This implies that items can also be selected via top–down attention, which is a slower, goal-oriented method of selection that operates based on the observer's intentions or expectations.

The use of eye movements to study attention is based on the assumption that there is a close link between where

doi: 10.1167/9.11.25 Received April 27, 2009; published October 27, 2009

individuals fixate and where they attend. The premotor theory of attention, for instance, states that visual attention becomes directed to a spatial location before eye fixations are generated to that area (Rizzolatti, Riggio, Dascola, & Umiltá, 1987; Rizzolatti, Riggio, & Shelgia, 1994). Support for this theory comes from findings that show a close link between attention and fixations. For instance, Hoffman and Subramaniam (1995) showed that detection of peripheral targets is enhanced when a saccade is set to be programmed to the target's location, even when participants are informed ahead of time that the target has a greater probability of occurring at a different location. This suggests that attention precedes the deployment of eye movements leading to faster responses to targets appearing at to-be fixated regions. Thus, although covert attention and eye movements can be voluntarily dissociated (as shown in the classical Posner, 1980, task), in many cases fixation locations are a good indicator of where people are attending, and consequently which objects or locations are receiving the most detailed processing.

Saliency

Several computational models of attention have been proposed which make predictions about where individuals attend when viewing complex scenes. The main premise of these models is that the entirety of the visual scene initially receives coarse, preattentive processing (Treisman & Gelade, 1980). This global information is insufficient for object recognition, however, and attention must be directed serially to regions in the scene to bind feature information from different dimensions (e.g., color, orientation) into a usable representation, e.g., for visual search (Wolfe, 1994). The question of how attention is deployed is, in general, a very complex one since it involves the internal state of the observer, including short-term and long-term goals, memory contents, expectations, etc. Substantial progress was made when Koch and Ullman (1985) removed these difficulties by focusing on the bottom–up (data-driven) part of attentional control, i.e., that part which is determined by the sensory input alone. Notably, they introduced the concept of a saliency map, defined as a topographically organized feature map that represents the instantaneous saliency of the different parts of a visual scene. Assuming that the most salient stimuli deserve the most attention, the bottom–up part of attentional control is thus reduced to the computation of the saliency map and then finding its maxima. In other words, the initial mechanism for selecting regions of the scene for higher-level processing is based on the conspicuity of each location in the scene (Koch & Ullman, 1985). This predicts that participants should attend to the most salient regions of a scene first, then to lower local maxima in the saliency map in the approximate order of their prominence.

The concept of the saliency map led to a computational implementation (Itti, Koch, & Niebur, 1998; Niebur & Koch, 1996) which allowed its predictions to be tested quantitatively. This implementation identifies the most salient, or locally distinctive, areas in an image across three dimensions (and many spatial scales): color, orientation, and intensity. The values of each dimension are initially stored separately in three distinct feature maps, which are later combined to form a master saliency map. Attention is proposed to be directed serially in a winner-take-all fashion to the location in the image corresponding to the highest saliency value in the master map, and then in decreasing order to the location with the next highest saliency value. An inhibition-of-return mechanism (Posner, Rafal, Choate, & Vaughan, 1985) discourages attention from immediately returning to previously attended areas. Since the observer's goals or expectations are not taken into account, and the computation of saliency is based solely on the visual properties of the image, this model's predictions rely purely on bottom–up information.

Eye movements

Several studies have explored the predictive value of the saliency map model. The most direct test of the model would compare covert attentional selection choices with model predictions. It is technically easier to instead use eye fixations as the dependent measure, i.e., to test where humans attend against the model's predictions and, of course, how humans control their eye movements is a question of great interest by itself. In the first published study applying the saliency map model towards understanding human eye movements, Parkhurst, Law, and Niebur (2002) recorded participants' eye movements as they free-viewed a series of complex images (natural scenes and fractals, with image statistics comparable to those of natural scenes). They then compared the fixation locations with the saliency model's predictions and found that the model predicted participants' fixations significantly better than chance. As expected, the prediction was better for the first fixation than for later fixations since top–down influences are likely weaker for the first fixation when less is known about the image contents. Areas of high texture contrast, another example of bottom–up influences, were also shown to attract fixations (Parkhurst & Niebur, 2004). Because the saliency model's predictions are based solely on the bottom–up features of the images, Parkhurst et al. concluded that eye movements, and hence attention, are drawn to salient locations in scenes. In more recent work, Foulsham and Underwood (2008) asked observers to first view natural scenes with the goal of memorizing them and then, on a second presentation, decide whether a given scene had been viewed previously or not. They confirmed that saliency is a significantly better predictor of fixation locations than random models, both during the memorization and the recognition phase of the experiment. This does not seem to be the case when participants are performing a visual

search task for a known target: in this case, they are able to cognitively override low-level features such that their eye movements are not preferentially directed to salient distractors (Underwood & Foulsham, 2006) although the presence of salient distractors increases reaction times (Foulsham & Underwood, 2009). This latter effect is presumably due to covert attention being directed to salient items without resulting in eye movements. Thus, saliency directly influences eye movements during tasks involving a "scanning" of the scene (as in memory encoding and retrieval tasks), and low-level saliency influences covert though not necessarily overt attention even during top–down dominated visual search tasks.

Interest

A related body of research has explored whether interesting locations in scenes draw attention and eye movements. Using a change detection paradigm, Rensink, O'Regan, and Clark (1997) found that participants were faster to detect changing items which were rated by a separate group of participants as being of central interest to that image. Rensink et al. suggested that interesting locations or objects attract peoples' attention, resulting in improved change detection in those areas of the scene.

In an early study, Mackworth and Morandi (1967) had one group of participants subjectively rate the most informative regions in a set of images, while a separate group viewed the same images as their eye movements were monitored. The main finding was that participants in the eye monitoring group fixated longer on the areas of the images that participants in the rating group independently rated as more informative. This method of asking participants to identify the most informative regions of images likely biases them to select regions high in top–down information, for instance areas or objects that are semantically important for identifying the scene. This suggests, then, that areas high in top–down or semantic information also attract attention and fixations. More recent studies have confirmed this link. For instance, Henderson, Weeks, and Hollingworth (1999) found that participants spend more time fixating semantically inconsistent objects in scenes, but that initial fixations were unaffected by the presence or absence of scene-inconsistent items. These results are consistent with the conclusions of Parkhurst et al. (2002) that early fixations are predominately influenced by image saliency, while later fixations are more influenced by top–down factors.

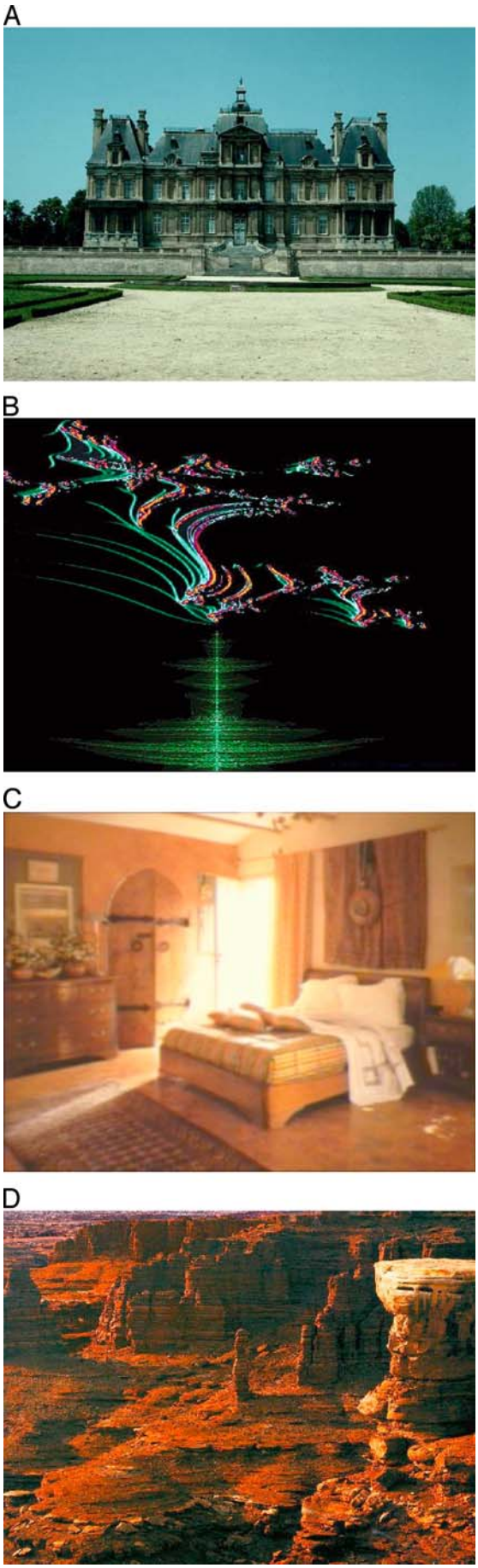
Correlating attention, eye movements, and interest

We use an approach that is inspired by that employed by Mackworth and Morandi (1967). We instructed participants to select, order and mark (by mouse clicks) the five most "interesting locations" in a series of images, without time pressure. This method of defining interesting regions has several advantages. It is a rather natural behavior and we consider it likely that it simultaneously reveals bottom–up and top–down influences on attention, although we have no control about their relative contributions. Second, the selected interesting locations are inherently sorted by importance. This is an improvement over methods that do not provide a relative ranking, and matches well with eye movement studies that use fixation number as a ranking system.

The purpose of the present study was to examine subjectively determined interest points. The first question asked is the degree to which participants agreed upon which points they perceived as interesting. While the outcome might have been different, we show that this agreement is surprisingly strong. This, then, allows us to ask our second main question, how well the interesting locations are correlated with other measures of attention. In the computer vision literature the term "interest points" typically refers to those areas of an image which are important for object or scene identification and are defined computationally rather than subjectively. These regions are generally highly salient (Wolf & Deng, 2005) or invariant across different viewpoints or lighting conditions (Schmid, Mohr, & Bauckhage, 2000). While these interest points appear to attract attention and fixations (Privitera & Stark, 2000), they are often determined using the bottom–up features of images and thus may not take top–down factors into account.

In Experiment 1 participants in an online study clicked on what they deemed to be the five most interesting locations in each of a series of images. Due to our ability to collect data from a large number of participants (more than one thousand), we first examined whether interest points are determined idiosyncratically, or if large numbers of participants tended to agree upon which areas of a scene are the most interesting. High consistency would be expected if interesting locations are primarily based on properties of the image itself, such as bottom–up saliency, while low consistency would be expected if interest points are determined based on individually different preferences (or if substantial variability is added by external noise processes). To determine the correlation between saliency and interesting locations, we also created a saliency map for each image using the saliency model of Itti et al. (1998) and compared the model's predictions to participants' interest point selections.

In Experiment 2, a separate set of subjects participated in an eye tracking experiment and free-viewed the same set of images presented in Experiment 1. The purpose of this experiment was to determine whether interesting locations attract attention by comparing participants' fixations with the regions of the scene that participants in Experiment 1 rated as interesting.



Experiment 1

The first purpose of Experiment 1 was to establish whether interesting locations are determined idiosyncratically, that is whether different people find different locations interesting, or whether selections are consistent across participants. If selections are consistent, then one would expect the majority of interest points to be grouped, or clustered, around a small number of locations per image. Such a finding would imply that some feature inherent in the image is accounting for the clustering and would be inconsistent with the claim than interest points are determined based on individual, varying preferences. Secondly, to examine whether bottom–up factors could account for interest point selections, we created a saliency map for each image using the implementation of Itti et al. (1998), and we compared its predictions to participants' interest point selections. If interest points are based primarily on bottom–up features, we expect a high correlation between the saliency maps and interest points.

Method

Participants

A total of 1395 entries were collected from a Web site affiliated with The Johns Hopkins University. The experimental methods were approved by the Johns Hopkins Institutional Review Board. These entries were initially filtered based on the responses to a demographic questionnaire at the beginning of the experiment. Only data from individuals who reported normal or corrected to normal vision, normal color vision and being 18 to 99 years old were included. Next, to minimize the number of repeat participants, we removed those subjects who reported having participated in the experiment before. Also excluded were participants with the same screen names (see below) or IP addresses as those of previous participants. This resulted in a total of 802 (425 males, 377 females) unique participants with a mean age of 27.4 years (*SD* = 9.8 years). All demographic information was self-reported.

Stimuli and apparatus

A total of 100 images were used, with 25 images selected from each of four categories: buildings, fractals, home interiors, and landscapes (see Figure 1 for an example of an image from each category). These stimuli are a subset of the images used in Parkhurst et al. (2002), and spatial frequency analyses of the images are provided in Parkhurst and Niebur (2003). On average, each image was viewed by 183 participants.

Figure 1. One example of each image from the four categories: A) buildings, B) fractals, C) home interiors, and D) landscapes.

A custom designed Java application was used to conduct the online interest point experiment. Participants viewed 15 images selected randomly from the whole image set and presented sequentially at a resolution of 640 × 480 pixels at the top-center of their browser window. Due to the online nature of the study, there was no control of the image display (e.g., properties of the monitor, background illumination, etc.) or the visual extent of the images as viewed by the participants, which varied as a function of screen size, resolution and viewing distance. Note that this considerable variability in image presentation conditions will decrease whatever clustering effect we observe. Thus, our measures for clustering as well as for the correlation between the bottom–up model and the responses of the participants must be considered conservative estimates.

Procedure

Participants first read an online consent document and gave their consent to participate by providing a screen name. Next, they answered a series of demographic questions, including whether they could see their computer screen without any difficulties, whether they were colorblind, their age, and whether they had participated in the experiment before. Participants were then instructed that they would view 15 sequentially presented images, and for each image they should, "Click the 5 points that are the most interesting to you." Each time a participant clicked on a location, a red circle appeared at that location and remained visible until all five locations for that image were selected. The next image then appeared automatically. The experiment was self-paced and took approximately 5 minutes. At the end of the experiment participants could elect to repeat the procedure and view another 15 randomly selected images as many times as they chose, which many elected to do. Only responses from the first 45 trials, in each of which the participant saw a unique image, were included in the analyses. Thus, no participant ever saw the same image more than once. We used only the first 45 trials (3 complete sets) because we wanted to obtain a representative sample of normal observers. We did not use all the data from the (few: 46 total) participants who did more than 3 complete sets of images because we were concerned that those observers might have atypical traits (e.g., obsessive-compulsive).

An alpha level of .05 was used in all experiments, and the Greenhouse–Geisser adjustment was used when applicable. We note that the *p*-values we found in the randomization tests were usually well below this level.

Results and discussion

Interest point selections

Reaction times

Figure 2 shows the mean reaction times for the five interest point selections. An analysis of variance (ANOVA) revealed a main effect of selection number, $F(4, 3204) = 134.13$, $MSE = 8.39$, $p < .001$. Paired-samples *t*-tests showed that first selections were the slowest, second selections were the fastest, and subsequent selections showed a steady increase in reaction time, all $p < .01$. One explanation for this pattern is that participants initially viewed the entire scene for an extended time to determine the most interesting location, as well as to locate the next few interesting regions before making their first selection. Consequently, the choice to make the subsequent selections was facilitated and the time shortened.

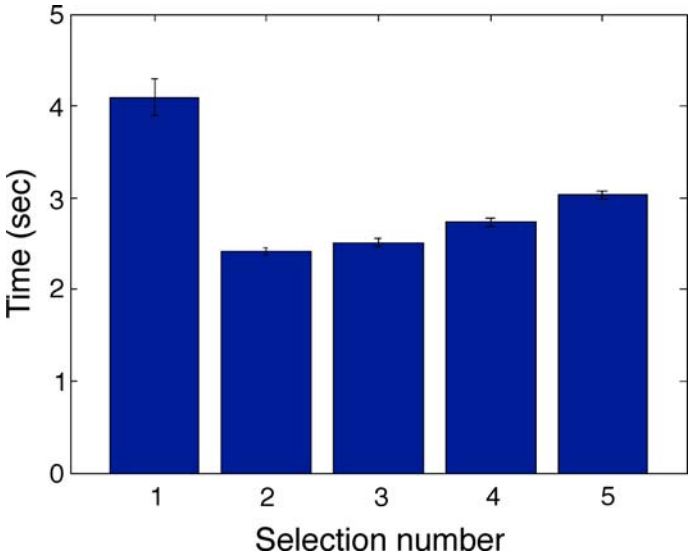


Figure 2. Mean reaction time to make interest point selections. Error bars represented plus and minus one standard error.

Clustering of interest points

We now address one of the central questions of this study, whether different participants select the same regions in the scenes as being most interesting, or if interesting locations are largely determined idiosyncratically. Because of its central importance (any conclusions comparing interest points with fixations and bottom–up saliency that are made in the second part of this report would be meaningless if interest points were idiosyncratic, i.e., would not show inter-individual commonalities), we measure clustering using four different methods. The first is a qualitative illustration, showing selection patterns for representative images from all image types. We then apply three different quantitative measures of clustering. The first of these compares the mean distances between interest points which, as is shown, is significantly lower than would be expected in the absence of clustering. The next method measures the *number* of interest points close to others; we show that these numbers are significantly higher than would be expected in the absence of clustering. These assessments of clustering were chosen since they are similar to those used in previous work in

this field. The fourth, and final, method we used is a standard k-means algorithm.

To qualitatively present the clustering, Figure 3 shows two images from each of the four image categories, with interest point selections plotted as colored dots (red for first and blue for following selections) superposed on the image. Results shown are representative for the whole data set: participants tended to select similar regions as interesting, resulting in around six to nine clusters of interest points per image with a moderate number of interest points falling outside the clusters. Thus, different participants select the same regions in the scene as being interesting, which would not be the case if interest point selections were determined idiosyncratically.

A more quantitative technique of determining clustering is to calculate whether interest points are closer together than would be expected by chance. If they are, this would suggest that different participants select similar regions as interesting. To determine whether there was more clustering of interest points than would be expected by chance, we first found the (x, y) coordinate of each interest point location, and we determined its distance from every other interest point in the same image. We then calculated the fraction of interest points that were separated by a given number of pixels. We used bins of 10 pixels, i.e., we determined the fraction of interest points that were between 1 and 10 pixels from another interest point, then 11 to 20 pixels away, and so on up to 50 pixels (we chose this maximal distance since beyond it the clustering effects begin to be counteracted by the compensation that has to occur at sufficiently large distances).

Next, we quantified the amount of clustering expected by chance. When computing the chance values, we have to be careful to avoid effects of systematic correlations that may be present in all images. For instance, participants may weigh the center of the image higher, either because this is the default position at the center of each trial or because they expect that important items have been placed by the image creator in the center (photographer's bias) or for other reasons (top–down and bottom–up contributions that generate a center bias were recently analyzed by Tseng, Carmi, Cameron, Munoz, & Itti, 2009). Indeed, we found that both interest-point selections and eye fixations (determined in Experiment 2, see below) were heavily biased centrally, as shown in Figure 4. Following Parkhurst and Niebur (2003), we therefore created a shuffled data set by comparing each interest point in one image to all interest points from a randomly selected image. For instance, the interest points for Image 1 were compared to those of Image 35, etc. This method ensured that any clustering we did observe was not due to inherent and systematic subject biases, for instance to select regions in the center of the image as interesting.

Figure 5 shows the mean percentage of interest points as a function of distance from each interest point selection. To determine whether interest points were closer together thank what would be expected by chance, an analysis of variance (ANOVA) with distance (groups of 10 pixels) and clustering data set (actual, shuffled) as within-subjects variables and image type (buildings, fractals, interiors, landscapes) as a between-subjects variable examined the fraction of interest points which were within the first 5 bins (of 10 pixels width) from each interest point. A main effect of data set was found, $F(1, 96) = 710.04$, $MSE = 0.297$, $p < .001$, which was due to more clustering in the actual than shuffled data set, as well as a main effect of distance, $F(4, 384) = 65.70$, $MSE = 0.123$, $p < .001$. Furthermore, there was a main effect of data type, $F(3, 96) = 4.41$, $MSE = 0.324$, $p < .01$. There was also an interaction between distance and data set, $F(4, 384) = 177.75$, $MSE = 0.116$, $p < .001$, as well as an interaction between distance and image type, $F(12, 384) = 11.69$, $MSE = 0.123$, $p < .001$. All of these effects were qualified by a Distance × Data Set × Image Type interaction, $F(12, 384) = 14.44$, $MSE = 0.116$, $p < .001$. Overall, the main finding is that for all image types interest point selections were closer together than what was predicted by the shuffled data set.

Our third method of examining interest point clustering was to find the fraction of interest points that fall within a cluster. An interest point was defined to be part of a cluster if 35 or more interest points were within 50 pixels of the location of that selection. The exact choice of these parameters is not critical, while these are the values chosen in the final analysis, similar patterns of results were found when different criteria were used for the required number of interest points to form a cluster, as well as the maximum distance of a cluster. The actual and shuffled data sets were created in similar ways as the previous analysis. In the actual data set, interest points from the same image were examined. In the shuffled data set, each interest point in one image was compared to the interest points from another image.

Figure 6 shows the fraction of interest point selections that fell within a cluster for each image type. An ANOVA with data set (actual, shuffled) as a within-subjects factor and image type as a between-subjects factor revealed a main effect of data type, $F(1, 96) = 1044.88$, $MSE = 101.60$, $p < .001$, showing more clustering in the actual versus shuffled data set, a main effect of image type, $F(3, 96) = 7.98$, $MSE = 114.97$, $p < .001$, as well as a Data Type × Image Type interaction, $F(3, 96) = 8.65$, $MSE = 101.60$, $p < .001$. Independent samples t-tests showed that the interaction was based on there being more clustering in the building $t(48) = 2.35$, $SE = 4.07$, $p < .05$, $t(48) = 4.40$, $SE = 3.38$, $p < .001$ and home interior image $t(48) = 3.29$, $SE = 3.92$, $p < .05$, $t(48) = 5.69$, $SE = 3.19$, $p < .001$ types, compared to the fractal and landscape image types, respectively. Buildings and home interiors did not differ from each other, $p > .10$ nor did fractals and landscapes, $p > .25$. However, the main finding is that for all image types a larger fraction of interest points were part of a cluster than would be predicted by the shuffled data set.

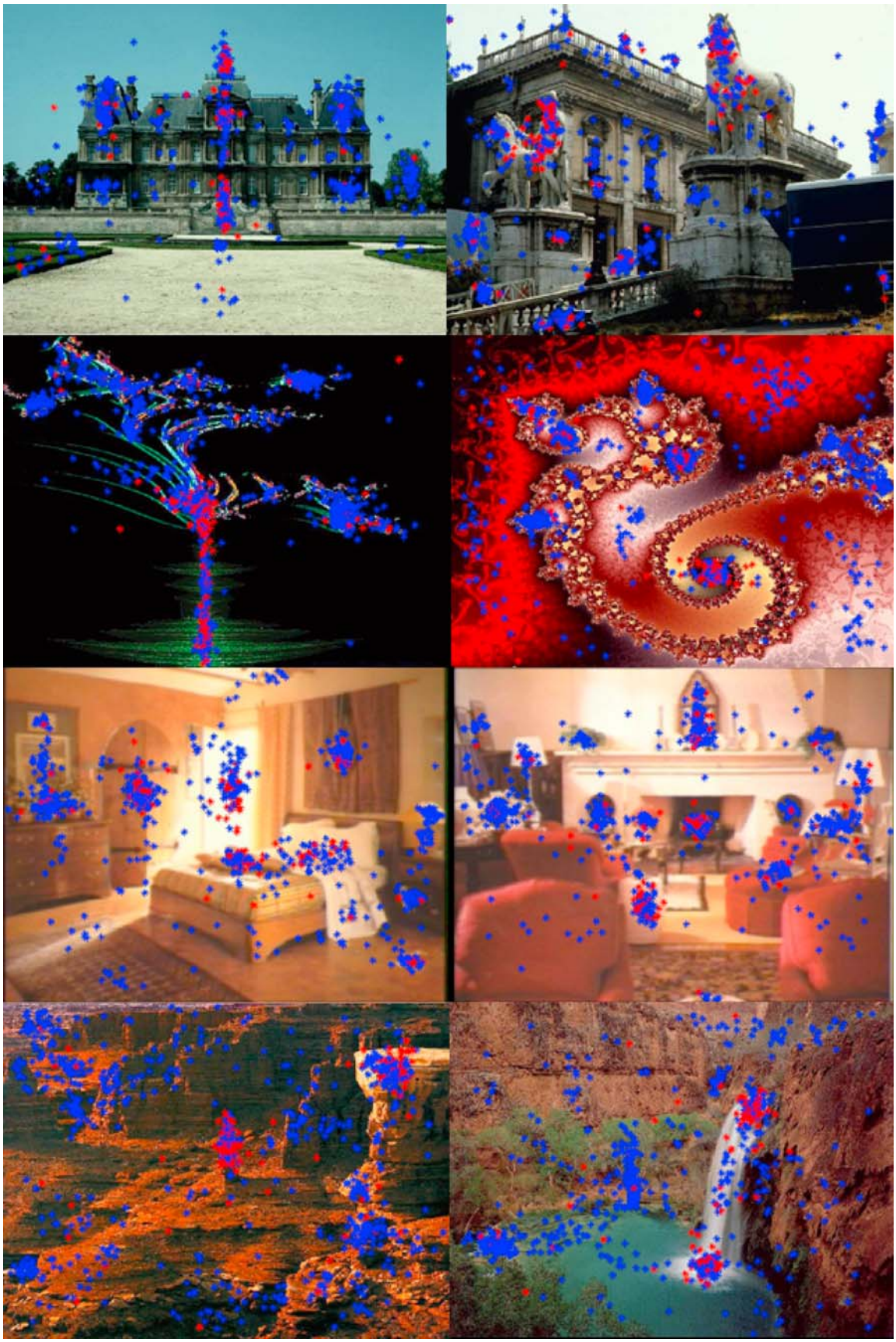


Figure 3. Two example images from each image category demonstrating the clustering of interest points. First selections are shown as red dots, selections two to five as blue dots.

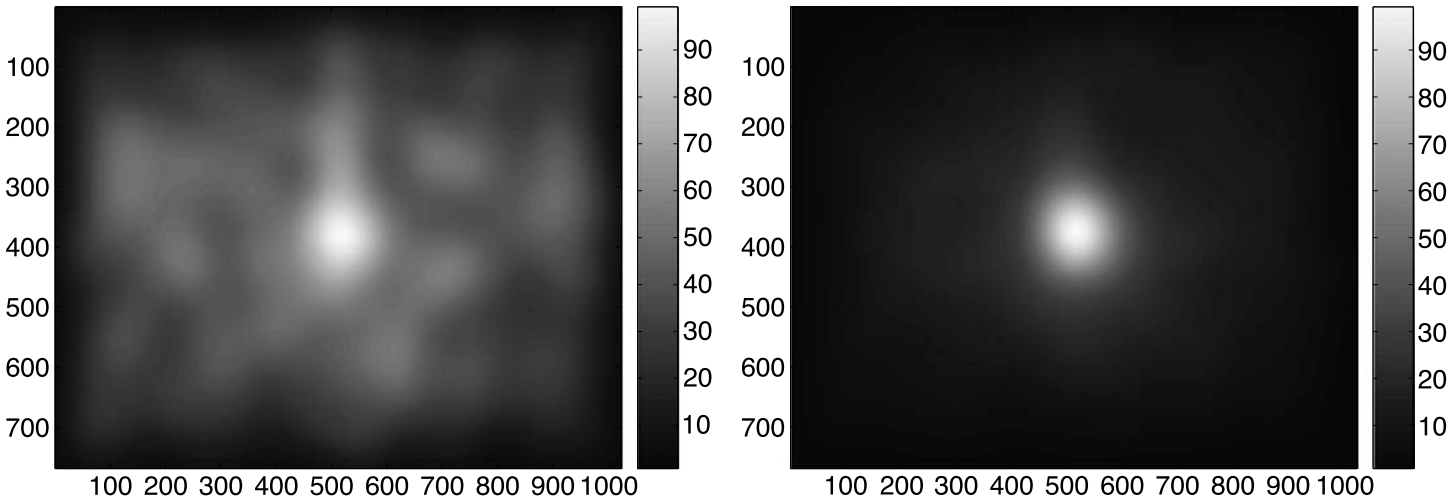


Figure 4. Grand average (over all images) of interest point selections (left) and fixations (right). The strong central bias is discussed in the Summary of results of Experiment 2.

The fourth method we used to test for clustering was a standard k-means cluster analysis, which is designed to partition the data into a given number of clusters. To determine the most appropriate number of clusters, we maximized the mean value of the silhouette (Rousseeuw, 1987). A range of 2 to 15 clusters was examined per image, and an average of 5 repetitions per cluster value was used to smooth the data. The number of clusters for which the mean silhouette value was highest was noted, and its corresponding mean silhouette value was used as a goodness-of-fit measure. Then, to determine the level of clustering expected by chance, 1000 matrices were created with 915 randomly defined (x, y) locations, equal to the average number of interest points per actual image. The mean maximum silhouette value for participants' interest point selections for the 100 images was 0.587 (SD = 0.061). The mean maximum silhouette value for the random distribution was 0.415 (SD = 0.005). For each image, the best mean silhouette value for participants' data was above the 95th percentile of the random data, suggesting that for every image a greater degree of clustering was observed than expected by chance. Furthermore, the number of clusters corresponding to the maximum silhouette value was also different. For the selections made by the participants, the average number of clusters which best fit the data was 10.92 (SD = 3.10), while for the random distribution it was 2.17 (SD = 0.55). Once again, these results strongly suggest that the interest point selections were best described as clustering around several independent locations. Next, we examined whether clustering differed as a function of image type. A one-way ANOVA on the silhouette values was significant, $F(3, 96) = 7.91$, $MSE = 0.003$, $p < .001$. Pairwise comparisons revealed that the difference was caused by smaller values for the landscapes compared to buildings $t(48) = 2.18$, $SE = 0.84$, $p < .05$ and home interiors $t(48) = 1.70$, $SE = 0.91$, $p < .1$, consistent with the previous clustering results.

In summary, our results provide strong and converging evidence that interest point selections cluster heavily around a few areas in each image. Overall, this strongly suggests that the selection of interest points is not idiosyncratic, and that participants agree on which locations in scenes are interesting.

Comparing interest points and image saliency

Having determined that interest points do cluster together, and the majority of interest points are part of clusters in every image type, we examined whether the location of interest point selections are correlated with image saliency, as determined by the saliency map model (Itti et al., 1998). Two methods were used to quantify the relationship between saliency and interest points. The first, the value comparison method, involved creating a saliency map and then identifying the values of that map at the locations of interest points. The second method, the cross-correlation comparison method, involved calculating the cross-correlation between the saliency maps and the interest maps. The techniques for creating the two types of maps, and the results of the two comparisons, are discussed below.

Value comparison

The first method of comparing saliency to interest points involved determining whether regions of the scene that participants deemed to be interesting had higher saliency values than would be expected by chance. To test this, we created a saliency map (SM) for each image in the database, using the algorithm in Itti et al. (1998). The model was run for each image individually, and the maps

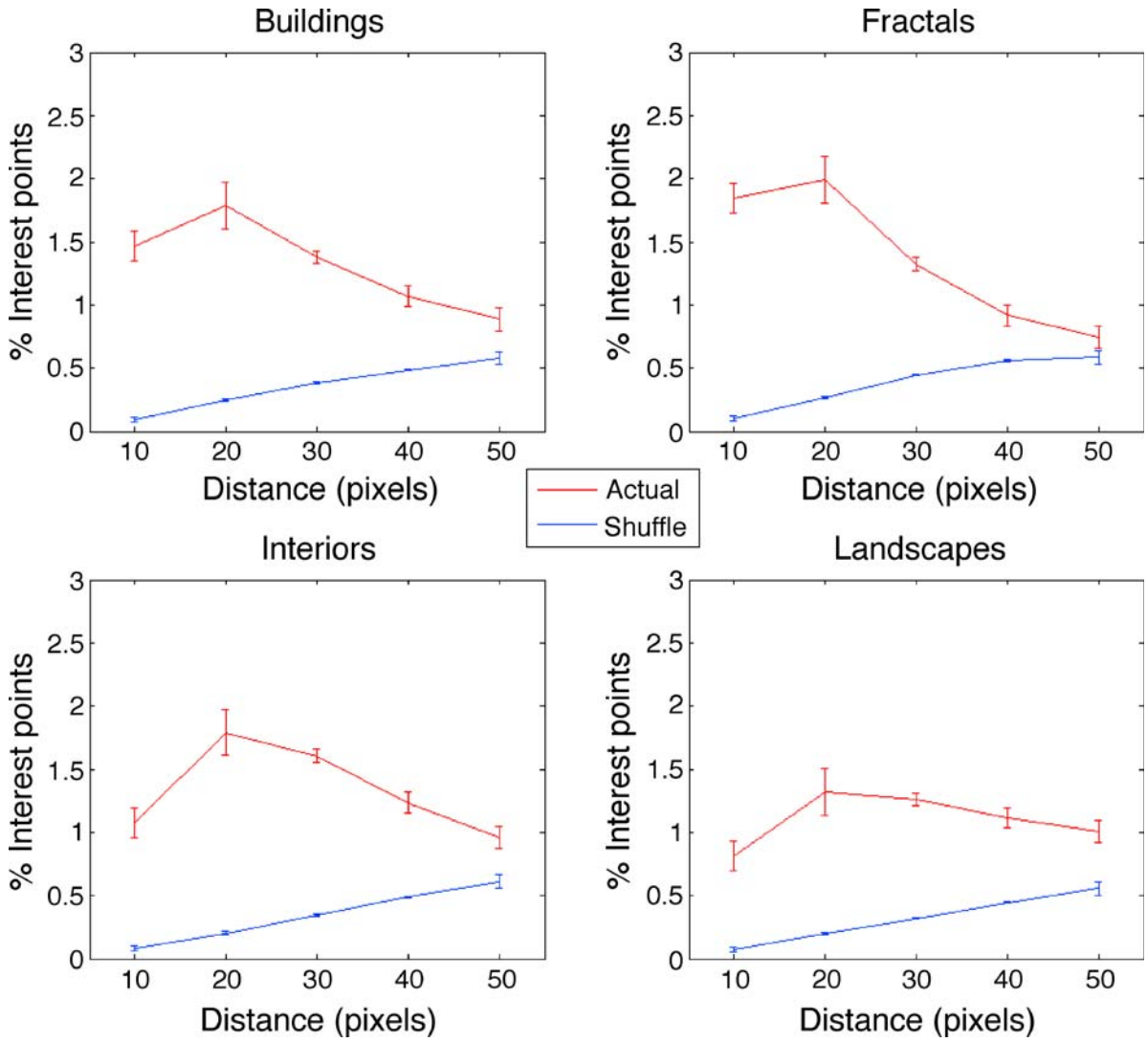


Figure 5. Clustering of interest points determined by the percent of interest point selections a given distance (pixels) away from each other interest point for the four image categories. Error bars represent plus and minus one standard error.

were normalized by dividing all values by the maximum value of that map, and multiplying by 100, thus ensuring that the values for each saliency map ranged between 0 and 100.

To determine if regions of high salience were selected as interesting, we extracted the value of the SM at the location of participants' interest point selections. Specifically, the (x, y) coordinates of all interest point selections for each image were determined, and the saliency values from that image's SM were extracted separately for the five selections. These values formed the actual distribution.

We used a similar technique to that of Parkhurst et al. (2002) to determine whether the values in the actual distribution were higher than would be expected by chance. Specifically, we compared the mean values for each selection number in the actual distribution to a chance sampling distribution. As in the creation of the shuffled data for the determination of interest point clustering (Interest point selections section), the chance sampling distribution was created by using the interest point selections from all other images to extract the values from each SM (Parkhurst & Niebur, 2003). For instance, the values of the random sampling distribution for Image 1 were obtained by calculating the mean saliency values from Image 1's SM using participants' interest point selections from Images 2 to 100.

Finally, to determine whether the selected regions were higher in saliency than would be expected by chance, we calculated the difference between the mean of the actual distribution and the mean of the chance sampling distribution for each image, which, following Parkhurst et al. (2002), we will refer to as the chance-adjusted saliency for interest

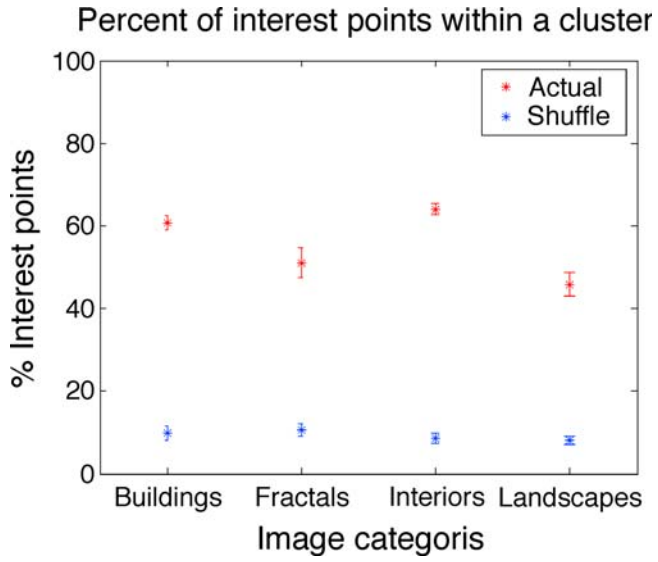


Figure 6. Clustering of interest points determined by the percent of interest points that fell within an interest cluster for the four image categories. Error bars represent plus and minus one standard error.

selections. If this value is positive, then participants selected areas of the image that were higher in saliency than expected by chance; if this value is negative, then participants selected areas that were lower in saliency than expected by chance.

The means of the actual and chance sampling distributions are plotted in Figure 8A. The chance-adjusted saliency for interest selections is the difference between the distributions for each selection number. To determine if participants selected regions of high saliency as interesting above chance, five one-sample *t*-tests tested whether the chance adjusted-saliency was greater than zero for the first five selections. All differences were significant, all $p < .001$, showing that for all selections participants selected areas higher in saliency than would be expected by chance. Next, a repeated measures ANOVA was conducted with selection number (1, 2, 3, 4, 5) as a within-subjects variable and image type as a between-subjects variable. The only reliable effect was a main effect of selection number, $F(4, 384) = 12.83$, $MSE = 70.403$, $p < .001$, as earlier selections showed a higher chance-adjusted saliency value than later selections. Overall, these results show that participants do select salient locations as being interesting more so than would be expected by chance, and that earlier selections are more influenced by image saliency than later selections.

Cross-correlation of interest and saliency maps

The second method used to compare interest point selections and image saliency involved calculating the cross-correlation value between the saliency map (SM) and the interest map (IM) for each image. The latter was computed, for each image, by convolving each interest point location with a Gaussian with a standard deviation of 27 pixels, truncated for efficiency at 3 standard deviations (81 pixels). As explained below (Experiment 2), this value represents the estimated spatial precision of the eye tracker. It is also the value used for creation of the fixation maps, below. Next, the total value of all points in the map were normalized so that the sum of all values for a given map equaled unity, and the mean of all values subtracted (see Figure 7). We also expanded the size of the saliency map to that of the interest maps to simplify the computation of the correlations.

In order to obtain a simple expression for the correlation between the maps, the SM was computed as discussed above but normalized and expanded, as for the IM.

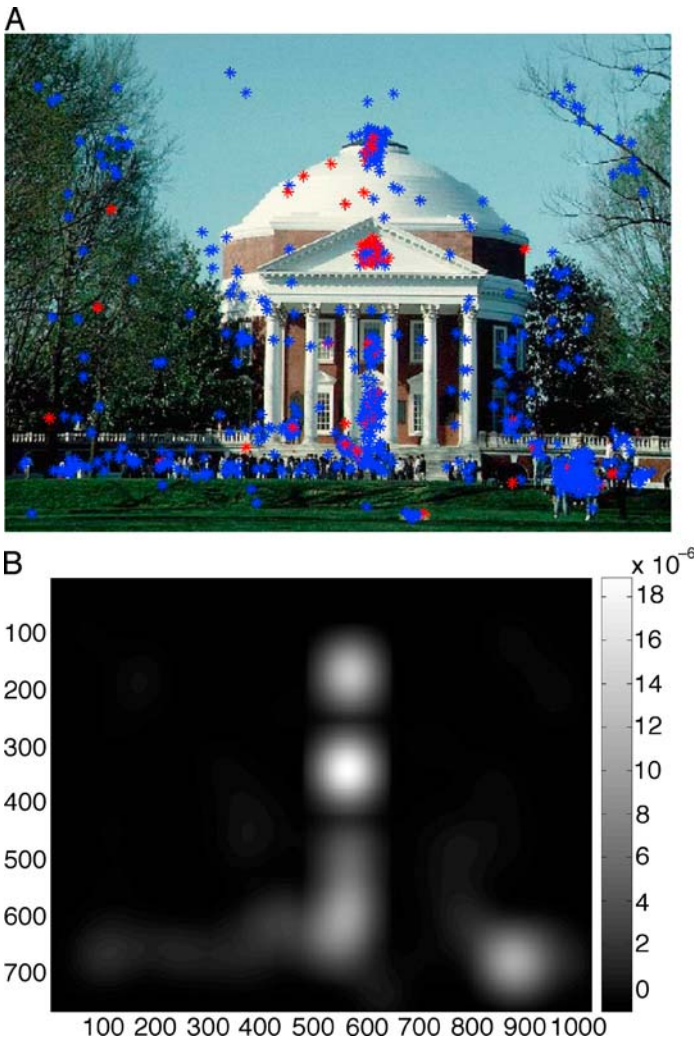


Figure 7. Creation of interest maps. (A) Original image with interest points plotted on top. Color coding as in Figure 3. (B) Interest selections with Gaussian intensity "blobs" centered on each interest point and superposed. Values have been normalized such that the sum of all values equals unity, and the overall mean has been subtracted.

Let us define the un-normalized cross-correlation between two maps P and Q, both of dimension M, N, as

$$C_{P,Q} = \sum_{x'=1,M} \sum_{y'=1,N} p(x', y')Q(x', y'). \quad (1)$$

The normalized cross-correlation between the interest map IM and the saliency map SM is then computed as the un-normalized correlation (setting $P = IM$ and $Q = SM$ in Equation 1) divided by the square root of the product of the autocorrelations,

$$C_{IS} = \frac{C_{IM,SM}}{\sqrt{C_{IM,IM}}\sqrt{C_{SM,SM}}}. \quad (2)$$

To determine the cross-correlation expected by chance, every SM was cross-correlated with all other IMs and the actual cross-correlation was compared to the resulting distribution of random values. Specifically, the 100 cross-correlation values between an IM and its corresponding SM were compared to the 95th percentile value in the random cross-correlation distribution, corresponding to the value expected by a p-value of .05. The mean value of the actual distribution was 0.368 (SD = 0.141). Overall, 61 out of the 100 comparisons were above the random sampling distribution curve at the 95th percentile, suggesting that over half of the cross-correlation values between the IMs and SMs are greater than what would be expected by chance (see Figure 8B). If we assume sampling from a Bernoulli process, with 61 out of 100 instances that had a probability of 0.05 (corresponding to 61 out of 100 images being above the 95th percentile), we would obtain this result with a probability of 5×10^{-53}.

Summary of results of Experiment 1

There are two important findings in Experiment 1. The first is that there is a high degree of consistency between participants' interest point selections. Specifically, we observed substantial clustering for interest point selections, signifying that many participants found the same few areas to be the most interesting regions in that image. This suggests that the selection of interest points is not idiosyncratic and argues against the possibility that each participant adopted a unique strategy for selecting interest points. Instead, participants base their selections on a particular set of features of the image that are available to all of them. The high consistency of our results also demonstrates the absence of extraneous variability in stimulus presentation and the behavioral responses, in spite of the low degree of environmental control that our online paradigm allows for.

For the complex natural scenes employed in the experiment, we hypothesized that participants select interest points based on a combination of bottom–up and top–down criteria. We demonstrate that bottom–up saliency does, indeed, play a significant role in their selections by showing that predictions of a purely bottom–up model, the saliency map, correlate significantly with the selections. To the extent that the saliency model is a predictor of bottom–up attention, interest point selections are biased by bottom–up factors. Since bottom–up saliency, as computed by the saliency map model, is common to all participants, it contributes to the high degree of consistency between participants. It should be noted, however, that while participants do select regions of high saliency as being interesting, the correlation is far from perfect. This is to be expected since the participants' task was to select "interesting" regions which likely includes taking top–down factors into account, such as regions of the scene, or individual objects in the case of home interiors or buildings, that are semantically important. Overall, bottom–up factors alone can only account for part of the strong consistency in the selection of interest points, suggesting that features of the image other than saliency are contributing to the participants' interest point selections.

In Experiment 2, we investigated the relationship between the interest point data from Experiment 1 and eye movement data recorded from a different sample of participants, as well as their relationship with bottom–up saliency measures.

Experiment 2

To further investigate the relationship between interesting regions and attention, in Experiment 2 we monitored participants' eye movements as they free-viewed the same set of images used in Experiment 1. We had two main goals.

First, we wanted to confirm that participants would indeed preferentially fixate salient regions of images used in our database. Previous studies have established a significant though modest correlation between eye fixations and saliency (e.g., Parkhurst et al., 2002), which we wanted to replicate with our images.

Second, we examined the correlation between fixations and interesting regions, specifically whether interesting regions serve as effective predictors of participants' eye movements. Experiment 1 revealed a moderate correlation between saliency and interest point locations. However, as discussed previously, the saliency model's predictions are based solely on bottom–up features, and it is likely that the selection of interest points involves some top–down influences. Fixations during scene viewing tend to be idiosyncratic and influenced by both bottom–up and top–down factors (see Henderson & Hollingworth, 1999, for a review). Thus, we predicted that interesting regions might

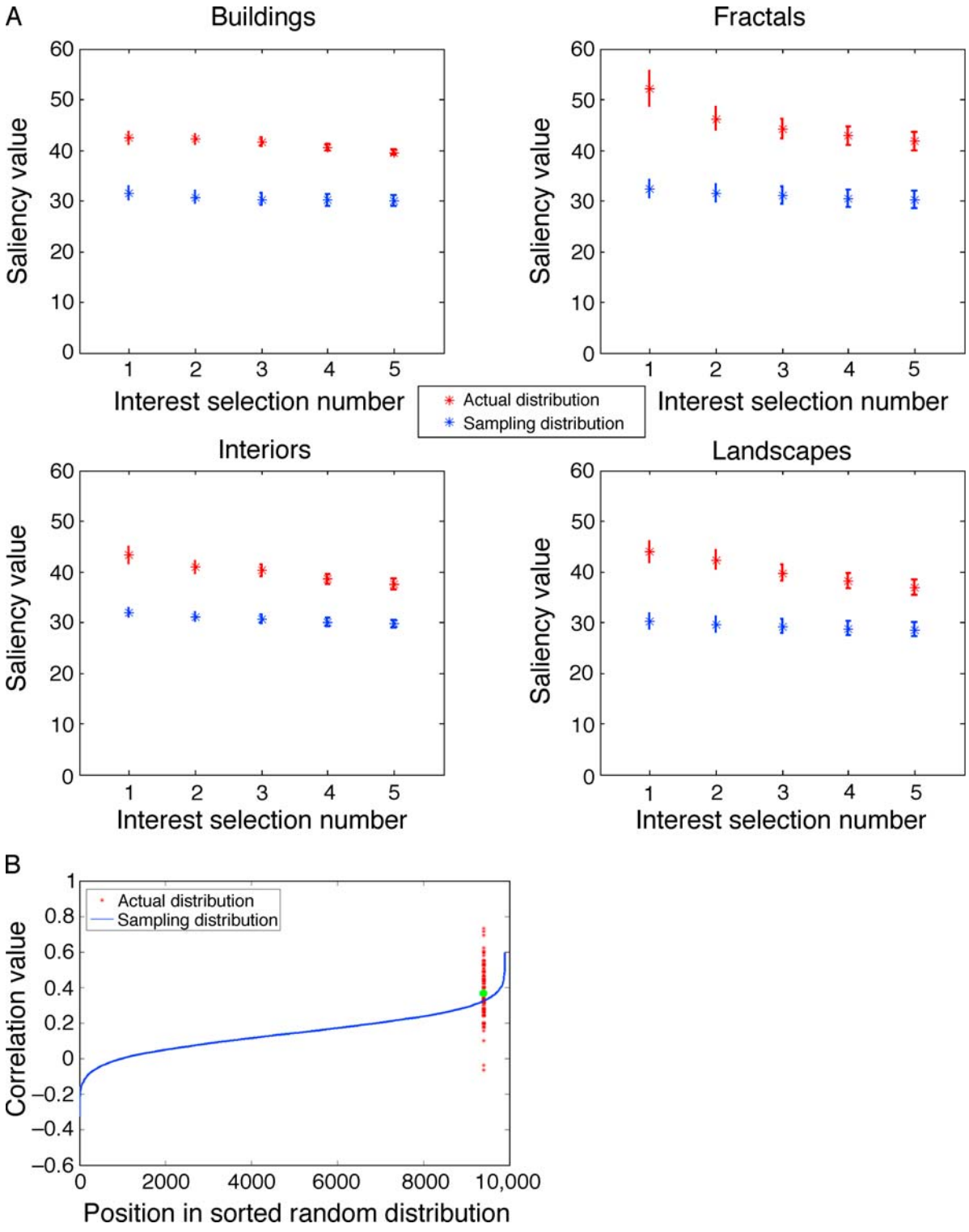


Figure 8. (A) Mean saliency values at different interest point locations for the four image types and the actual and chance sampling distributions. Error bars represent plus and minus one standard error. Note that the distance between the actual and chance sampling distributions represents the chance-adjusted saliency value. (B) Cross-correlation between interest maps and saliency maps. Values from the random cross-correlation distribution are sorted from weakest to strongest and from negative to positive correlations. The mean value of the actual distribution is plotted in green.

serve as an equal or even more robust predictor of fixations.

Method

Participants

All experimental methods were approved by the Institutional Review Board of Iowa State University. Twenty-one Iowa State University undergraduates (12 male) participated in the eye tracking task for course credit. All reported normal vision.

Apparatus

Eye movements were recorded by an ASL eye tracker (Model R-HS-P/T6 Remove High Speed Optics), with a sampling rate of 120 Hz. Images were displayed in full color on a Samsung SyncMaster 910T LCD monitor, with a viewing area of 38 × 30 cm. A chin rest was used to maintain a viewing distance of approximately 70 cm. A separate monitor, which was only viewable by the experimenter, indicated the participants' approximate fixation position.

Stimuli

Stimuli were identical to those used in Experiment 1, and each participant viewed all 100 images in random order. The images were expanded to encompass the entire screen (1024 × 768 pixels) and subtended approximately 30.4° × 24.2° of visual angle.

Procedure

Participants began by signing an informed consent document and were instructed that their task would be to, "Look around the images as you naturally would." They were then seated with their chin in a chinrest. The experiment was divided into two blocks of 50 images. Each block began and ended with a 9-point calibration sequence to calibrate the eye tracker, where a verbal command was given for participants to fixate on each number sequentially. Eye tracker error was determined by taking the mean distance between the calibration points and participants' fixation locations during the pre-experiment and post-experiment calibrations for both blocks. The mean difference was approximately 27 pixels (approximately 0.8°).

A trial began with the presentation of the fixation cross at the center of the screen, which participants were instructed to fixate until the image appeared. The experimenter initiated each trial when the view of the secondary monitor indicated that the participant was indeed fixating the cross. After a delay of approximately 1.5 seconds, a randomly selected picture from any of the four categories was presented for 5 seconds. The central fixation cross then reappeared to signal the beginning of the next trial. The experiment lasted approximately 35 minutes.

Eye fixations were defined by a combination of distance and time. Specifically, eye movements that traveled less than 1° in 25 ms, and were longer than 100 ms, were counted as a single fixation. To improve the statistical estimates of the eye tracking data, we identified trials where a failure in eye tracking occurred by summing the total time of all fixations for each trial and removing those trials that had a total fixation time less than three standard deviations below the mean trial time (4.23 seconds), which resulted in the rejection of 8.4% of trials. Overall, the mean fixation duration was 290 ms (SD = 21), and there were a mean of 12.89 (SD = 3.11) fixations per trial.

Results and discussion

Comparing image saliency and fixations

The same methods for comparing the relationship between saliency and interest point selections employed in Experiment 1, value comparison and cross-correlation, were used to compare the relationship between saliency and fixations.

Value comparison

To determine whether participants fixated salient regions above chance, we calculated the value of the SMs at the (x, y) coordinates of the participants' first 10 fixation locations. These values formed the actual distribution. The SMs were identical to the ones used in Experiment 1. The chance sampling distribution was also created in the same way as Experiment 1. That is, the values of the SM for each image were extracted from the fixation locations from all other images. Finally, to determine whether participants fixated more salient regions than would be expected by chance, we calculated the difference between the mean of the actual distribution and the mean of the chance sampling distribution for each image, the chance-adjusted saliency for fixation locations.

The means of the actual and chance sampling distributions are plotted in Figure 9A. To determine if participants fixated regions of high saliency above chance, 10 one-sample t-tests tested whether the chance adjusted saliency was greater than zero for the first 10 fixations. All differences were significant, all $p < .001$, showing that for the first 10 fixations participants fixate areas higher in saliency than would be expect by chance. Next, a repeated measures ANOVA was conducted with fixation number (1–10) as a within-subjects variable and image type as a between-subjects variable. The only reliable effect was a main effect of fixation number, $F(4, 384) = 3.50$, $MSE = 25.783$, $p < .001$, as the first fixation had a lower chance-adjusted salience value than all other fixations. The main

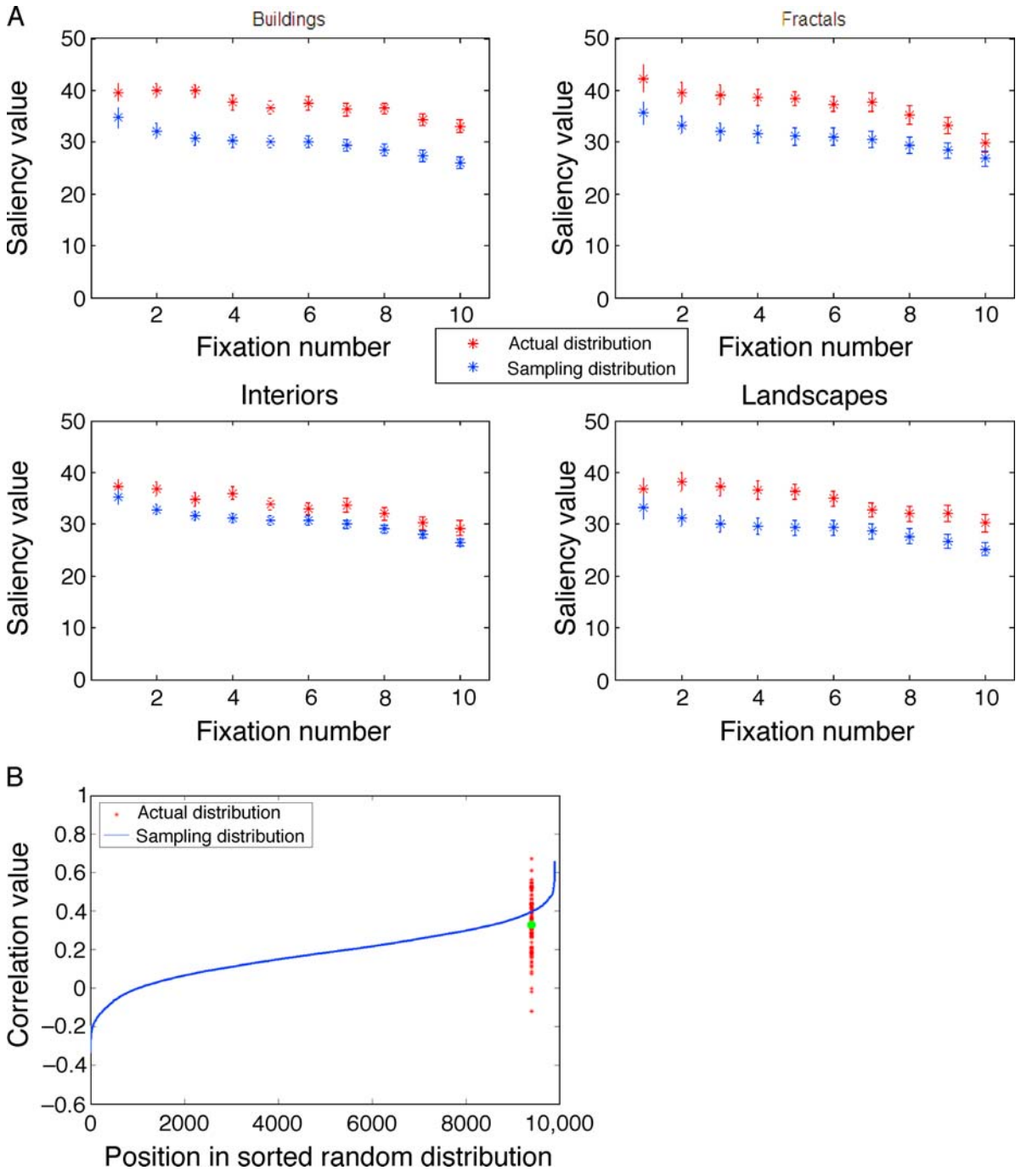


Figure 9. (A) Mean saliency values at different fixation locations for the four image types and the actual and chance sampling distributions. Error bars represent plus and minus one standard error. Note that the distance between the actual and chance sampling distributions represents the chance-adjusted saliency value. (B) Cross-correlation between fixation maps and saliency maps. Values from the random cross-correlation distribution are sorted from weakest to strongest and from negative to positive correlations. The mean value of the actual distribution is plotted in green.

effect of image type did approach significance, $F(3, 96) = 2.62$, $MSE = 293.246$, $p = .055$, and pairwise comparisons showed that the chance adjusted saliency value was smaller for home interiors compared to buildings and fractals. Overall, our results show that for our image set participants do fixate regions of high saliency above chance.

Cross-correlation of saliency and fixation maps

Next, we computed the cross-correlation between the SMs, which were identical to those used in Experiment 1, and fixation maps (FMs). FMs for each image were created in a similar way to the IMs in the Comparing interest points and image saliency section. Specifically, a Gaussian distribution with a standard deviation of

27 pixels (estimated eye tracker precision), truncated at 81 pixels (3 *SD*), was placed around each fixation location. Instead of the fixed weight used for interest maps, for the computation of fixation maps the Gaussian was weighted proportionally to the length of that fixation. Thus, longer fixations received a higher total value in the FM than shorter fixations. Then, the total values of each map were normalized so that the sum of all values for a given map equaled unity, and the overall mean subtracted.

The actual and random cross-correlation distributions were created in the same way as in Experiment 1, and the computations were identical to those performed there. The mean value of the actual distribution was 0.327 (*SD* = 0.145). Overall, 33 out of the 100 actual comparisons were above the random cross-correlation distribution curve at the 95th percentile, suggesting that about one third of the cross-correlation values between fixation and saliency maps were higher than expected by chance (see Figure 9B). The probability of obtaining this result by chance, again computed from the binomial distribution as in the Comparing interest points and image saliency section, is 1×10^{-18}.

Comparing fixations and interest points

The same two techniques as before were used to compare the relationship between fixations and interest point selections.

Value comparison

The interest maps were constructed as in the Comparing interest points and image saliency section and the statistical tests constructed as in the Comparing image saliency and fixations section, changing "saliency" to "interest" and "saliency map" to "interest map."

The means of the actual and chance sampling distributions are plotted in Figure 10A. To determine if participants fixated regions of high interest above chance, 10 one-sample *t*-tests tested whether the chance adjusted-interest value was greater than zero for the first 10 fixations. All differences were significant, all $p < .001$, showing that for the first 10 fixations participants fixate areas higher in interest value than would be expected by chance. Next, a repeated measures ANOVA was conducted with fixation number (1–10) as a within-subjects variable and image type as a between-subjects variable. A main effect of fixation number was found, $F(4, 864) = 11.22$, $MSE = 48.295$, $p < .001$, as interest values were higher for fixations after the first. The main effect of image type was marginally significant, $F(3, 96) = 2.18$, $MSE = 296.770$, $p = .095$. The Fixation Number × Image Type interaction was also significant, $F(27, 864) = 2.00$, $MSE = 48.295$, $p < .01$. This interaction could be understood by comparing the interest values for the first and second fixations for fractals to all other image types. For fractals, the chance adjusted interest value for the first two fixations was not significantly different (paired samples *t*-test, $t(24) = 0.294$, $SE = 2.25$, $p = .771$). For the other three image types, the chance adjusted interest value was significantly higher for the second compared to the first fixation, all $p < .01$. Overall, these results show that participants in Experiment 2 were likely to fixate image regions that participants in Experiment 1 found interesting.

Cross-correlation of interest and fixation maps

The cross-correlation between the IMs and FMs was computed, in the same way as described previously. As was done for the SM in the Comparing interest points and image saliency section, the size of the IM was expanded to be the same as that of the FM to simplify the computation of the correlations. The mean value of the actual distribution was 0.656 (*SD* = 0.113). Overall, 98 out of the 100 actual cross-correlations were larger than the value at the 95th percentile of the random sampling distribution (see Figure 10B), suggesting that practically all of the images had a higher cross-correlation between their corresponding interest and fixation maps than would be predicted by chance. The probability of obtaining this result by chance as computed from the binomial distribution, as in the Comparing interest points and image saliency section, is 1×10^{-124}.

Summary of results of Experiment 2

The first set of results in Experiment 2 largely replicated previous findings (Parkhurst et al., 2002), which revealed a moderate correlation between fixations and image saliency. Parkhurst et al. (2002) found that for the average over all image categories (results for individual categories were not provided), it was the first fixation that carried the highest saliency value (see their Figure 5). In Experiment 2, we likewise find that the chance-adjusted saliency (the difference between the red and blue symbols in Figure 9A) when averaged over all four image categories is higher at the first fixation than at the second ($t(99) = 3.07$, $SE = 0.63$, $p < .01$). A closer look at each image category separately reveals that this holds also individually for the three natural image types (landscapes, interiors, buildings; *t*-test, all $p < .05$) though not for fractals. Note that this is not the case for the raw (non-adjusted) saliency values which were found to be not significantly different between first and second fixations, a result that holds for each of the image categories separately as well as when they are collapsed (*t*-tests, all $p > .15$). Furthermore, in both studies the chance-adjusted saliency for fractals is found to be greater for the first fixation than for natural scenes (collapsed over the three classes of natural scenes, $t(98) = 2.14$, $SE = 1.51$, $p < .05$). This is consistent with an interpretation in which the top–down effects gain in importance as additional information from previous fixations is processed, and also with the intuitively

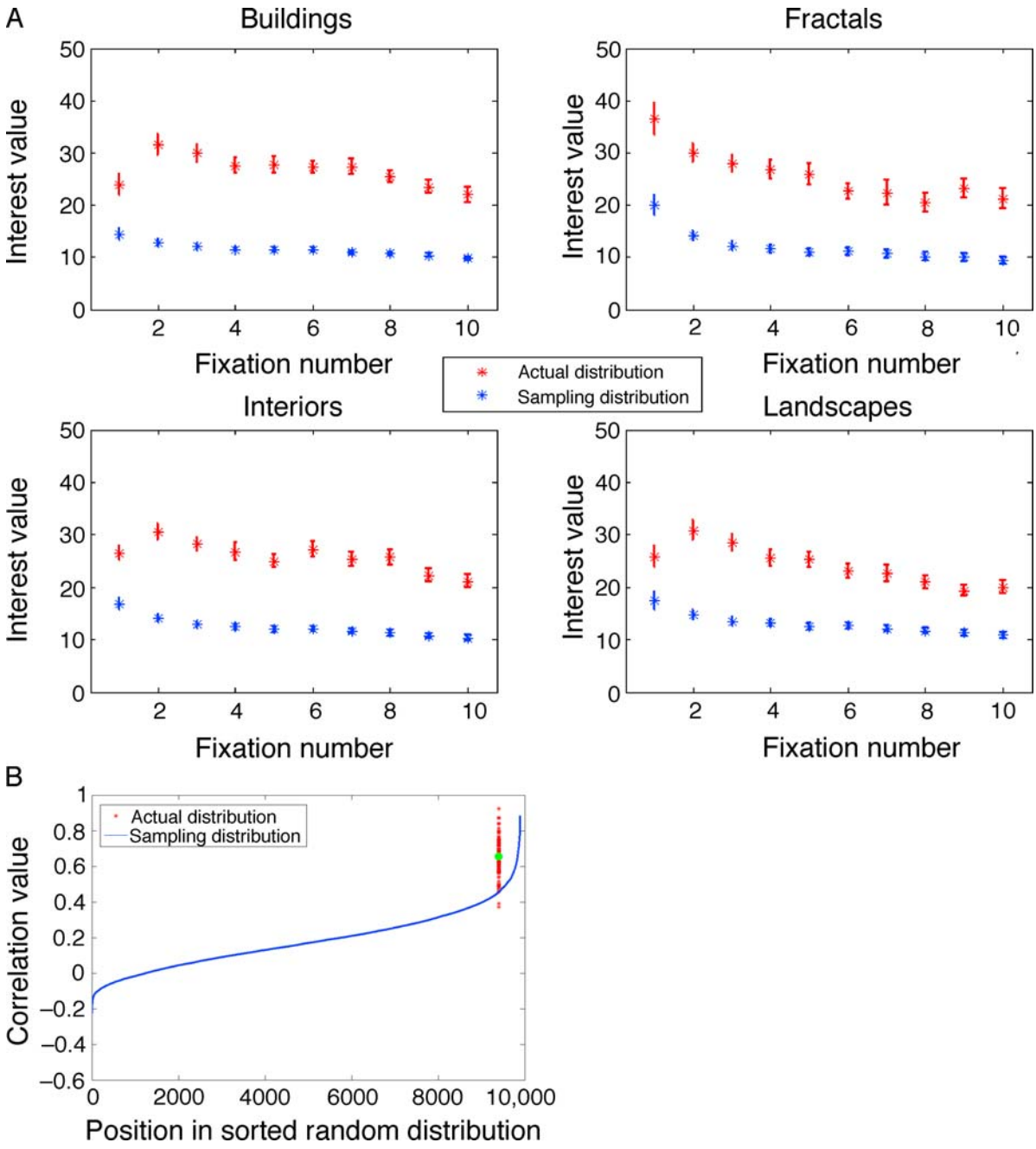


Figure 10. (A) Mean interest values at different fixation locations for the four image types and the actual and chance sampling distributions. Error bars represent plus and minus one standard error. Note that the distance between the actual and chance sampling distributions represents the chance-adjusted interest value. (B) Cross-correlation between fixation maps and interest maps. Values from the random cross-correlation distribution are sorted from weakest to strongest and from negative to positive correlations. The mean value of the actual distribution is plotted in green.

plausible assumption that top–down effects play a greater role in the semantically "meaningful" natural scenes (landscapes, buildings, interiors) than in fractals. Therefore, the purely bottom–up saliency map model performs better on earlier than later fixations, and better on fractals than on natural scenes.

The situation is more complex for the relationship between conscious selection of interesting points and eye movements. For fractals, the chance adjusted interest value decreases monotonically with fixation number (Figure 10), as was the case for salience. For natural scenes (Buildings, Interiors and Landscapes) that presumably have more "semantic" contents than fractals, we find that the chance adjusted interest value is lower at the first fixation than at the second (see the Comparing fixations and interest points section), and it is monotonically decreasing

for the following fixations, as expected. One possible explanation is that semantically controlled top–down influences require more information about image contents than is the case for fractals and more than can be acquired without at least one fixation. Therefore, some of this information is acquired during the first fixation and it is only at the second fixation that the center of gaze is on the most "interesting" location. Although the time scales of eye movements and interest point selections differ by an order of magnitude, this interpretation is at least consistent with the observation that the selection of the first interest point takes significantly longer than that of the second (Figure 2), implying that some amount of time is needed to analyze the image for interesting locations. Furthermore, the short time for the first fixation of fractals is reflected in the fact that the first interest point selection is selected fastest for this image type (3.75 vs. 3.97 seconds for landscapes, 4.33 seconds for interiors, and 4.37 seconds for buildings). We note that a similar pattern, with the highest saliency encountered at the second fixation, was found by Foulsham and Underwood (2008) while participants scanned natural scenes of a similar nature as our buildings, interiors, and landscapes (they did not use fractals) with the goal of memorizing the images. We also acknowledge, however, that the true first fixation a participant makes after the image appears can be difficult to identify. In most experimental paradigms, including ours and that of Foulsham and Underwood, the trial starts by the participant fixating a marker in the center of the screen. The first fixation away from this default position is frequently quite small (possibly because of the center bias) and whether a particular eye movement is classified as a fixation depends then somewhat sensitively on the parameters of the fixation–detection algorithm. If the algorithm misclassifies a fraction of the first fixations (e.g., by including a small drift of the eye among the fixations), these will on average fall on a part of the image that has low saliency, and these misclassified fixations will then lower the mean saliency of the whole set of first fixations. Of course, such a misclassification will also contribute to the central bias and this may contribute to the central bias of the fixation map (Figure 4, right), which is much stronger than that of the interest map (Figure 4, left).

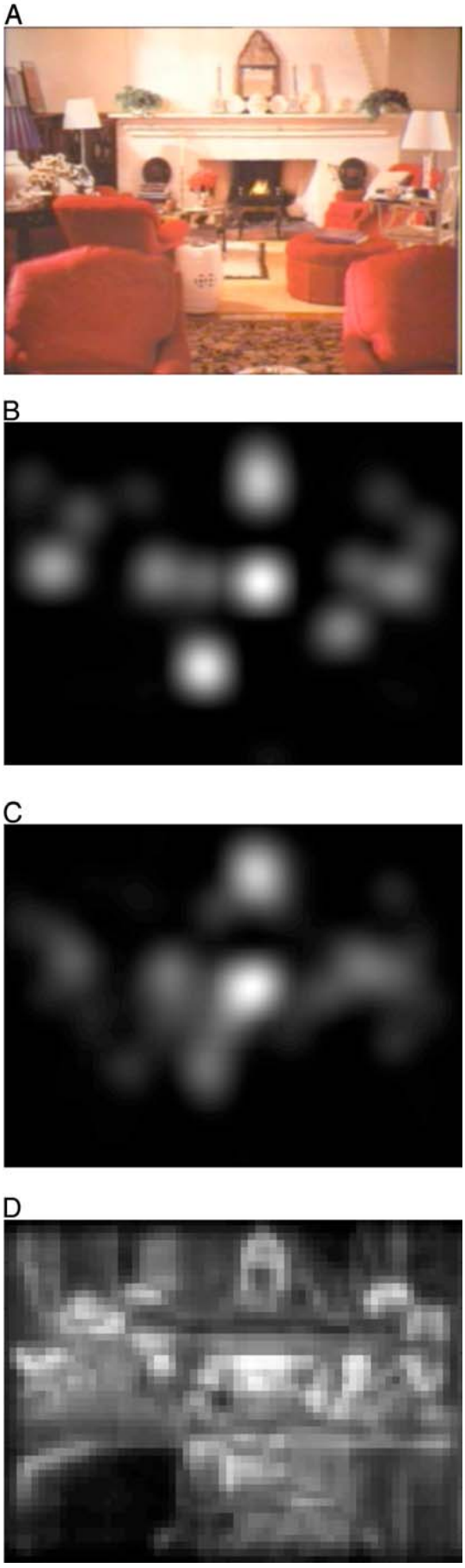
But independent of these variations, one result of crucial importance remains that even after many fixations and for all scene types considered, bottom–up saliency influences eye movements: In both studies, saliency values are significantly above chance levels not only for the first but for all fixations for the duration of the experiment (5 seconds).

A novel finding is that there is a strong correlation between subjectively defined interest point locations and fixations. Specifically, participants in Experiment 2 tended to fixate the same regions of the scene that participants in Experiment 1 found interesting. To the extent that eye movements are influenced by bottom–up and top–down attentional factors, it seems reasonable to suggest that interest points are thus influenced by both types of factors as well.

General discussion

Among internal (mental) states, selective attention has been recognized as being particularly amenable for objective study nearly a century and a half ago (von Helmholtz, 1867). Psychophysical experimentation has led to a thorough quantitative characterization of attentional selection and a number of psychological models. More recently, models of the neural basis of selective attention were developed. Of particular importance is the saliency map concept introduced by Koch and Ullman (1985). Rather than attacking head-on the extremely complex general problem of selective attention, which is so central to perception and cognition that solving it in its entirely would probably involve understanding most of an individual's mental life, these authors decided to focus on the "low-hanging fruit"-type question of bottom–up attention. Not only does this have the practical advantage of being amenable to experimental study much more directly than top–down attention (since it can, by definition, be manipulated by instantaneous sensory input), it also avoids the complexities of influences by personal history (long-term memory) and of other mental states (e.g., goals) which are in general difficult to measure and to quantify. Furthermore, the concept of a saliency map can be implemented in algorithmic form (Itti et al., 1998; Niebur & Koch, 1996), which makes it immediately amenable to quantitative hypothesis testing (and has the additional benefit of making it useful for technical and practical applications).

Previous experimental tests of predictions of the saliency map model of *covert* attention have used *overt* attention, i.e., eye movements (e.g., Parkhurst et al., 2002), demonstrating that bottom–up scene-contents influences attentional selection. In the present work, we go one step further away from the sensory periphery and investigate whether the purely bottom–up predictions of the saliency map model have predictive value even for consciously made decisions about what constitutes an "interesting" area of an image. Given the high-level of abstraction of this concept (whose precise definition is deliberately left to the individual human observers), it was expected that inter-individual differences will be substantial. We therefore developed a novel, Internet-based approach that allowed us to use a very large population of observers (Experiment 1). Together with results from a more traditional study of eye-movements (Experiment 2) on the same image set, we can constructs maps for bottom–up saliency (SM), fixations (FM), and conscious selection of interest (IM). For the sake of illustration, we show one example image together with all these three



maps derived from this image (Figure 11). Our major results, to be discussed below, are the relationships between these three maps, and the quantitative determination of the inter-individual differences between points selected as interesting.

Limited inter-individual differences in selection of interest points

In our first experiment, we addressed the question of which points in complex images (natural scenes and fractals) human observers consider "interesting," using their own interpretation of this term. Although observers have been questioned about features or objects they find of interest (e.g., Rensink et al., 1997, also see the discussion in the Interest section and Correlating attention, eye movements, and interest), we believe that, surprisingly, the very simple task of prioritizing which points in a complex visual scene observers consider the most interesting using their own criteria has never been addressed in a systematic and quantitative way. Before performing the experiment, one might have expected that different observers would make vastly different choices in what they consider personally interesting. One might have argued that different human observers are sufficiently different from each other and that, therefore, there should be little agreement between their conscious determination of what is "interesting." Furthermore, any source of random variations, both within observers and within the experimental setup which by its nature is much less closely controlled than traditional psychophysical experimental paradigms will add to the variability between the recorded choices the observers make.

Our results show that the selection of interest points was very consistent across participants. Figure 6 shows the remarkable result that for each of our four image classes, more than half of all selected locations clustered in areas that comprise only a very small fraction of the image. We consider this a truly surprising result. The complexity of human behavior, in combination with an experimental setting with a large number of uncontrolled variables, could have resulted in a very "noisy" distribution of interest points. Instead, we found a highly structured pattern of selections that tended to cluster around a few regions in each image. We can safely conclude that selection of interest points is *not* idiosyncratic but reflects something inherently significant about certain areas of scenes. This is even more remarkable considering the differences in viewing angle, monitor quality, and other uncontrolled factors that could have lead to vastly different responses between participants who responded on their personal computers via the Internet. These factors are

Figure 11. Example image (A) with associated interest map (B), fixation map (C), and saliency map (D).

nearly certain to have added variability and "noise" to the data so that the true degree of agreement between observers is likely even higher than what we measured. Significant consistency between observers was also found in a recent study (Cerf, Cleary, Peters, Einhäuser, & Koch, 2007) where observers rated the saliency of a whole image relative to other images of the same category (rather than parts of a image relative to other parts of the same image, as in the present study). Significant inter-observer consistency was also observed when the same image set was tested again a year later when participants reported not having explicit memory of the specific images they saw. These results support our hypothesis that it is objective criteria, rather than idiosyncratic decisions, that determine what people find subjectively interesting.

Relation between bottom–up saliency, eye movements and subjective interest

The high consistency we found between choices of interesting points is not only an important finding by itself but it also allows us to constructs meaningful maps of interest attributed to different areas of an image. The question then arises naturally how the subjective interest thus determined is correlated with eye movements on the one hand, and with bottom–up saliency on the other hand. Between the three maps (IM, SM, and FM), there are thus three comparisons to be made. We found statistically highly significant correlations (all $p < .001$) for all these three relations (Figures 8, 9, and 10).

Thus, not only are subjective interest and eye movements correlated with each other (even in different subjects), but both are furthermore correlated with salient regions as determined by the saliency map model (Itti et al., 1998). This suggests that the selection of interest points is biased by bottom–up factors, and that interest points can serve as a useful indicator of bottom–up attention. We note in passing that these results cannot be explained by a correlation between some generic correlations between the members of our image set. Such correlations do indeed exist, as is shown by the fact that the mean correlation between shuffled selections is positive (the averages of the blue curves in Figures 8B, 9B, and 10B are all positive), but the correlations are significantly stronger than can be explained by this effect (red dots).

As expected, the correlation was not as strong as one would expect if interest points were determined solely by saliency. While it is certainly possible that other models of bottom–up attention may show stronger correlations with interesting locations, we believe it is more likely that other factors besides saliency were involved in the selection of interesting regions. It is likely that top–down information plays a role in the selection of interesting locations. This would explain why the saliency model, which is based on solely bottom–up information, is only moderately correlated with interesting regions. This is supported by our result showing that interest point selections from Experiment 1 were a very robust predictor of participants' eye fixations in Experiment 2. In other words, participants fixate interesting locations (even though our criterion of "interesting" is that they were found interesting by *other* individuals).

Previous research suggests that fixations are influenced by top–down and bottom–up factors (Henderson & Hollingworth, 1999), which could explain why interest points tend to correlate better with fixations than saliency. Fixations are also correlated (though imperfectly) with detected changes in change-detection paradigms (O'Regan, Deubel, Clark, & Rensink, 2000), and these changes are on their turn correlated with subjectively perceived salience (Wright, 2005). Perceived salience is most likely a combination of bottom–up and top–down contributions (Wright, 2005). Thus, interest points (and fixation points) seem to reflect the combination of bottom–up and top–down attention.

The question of the relative contributions of bottom–up and top–down influences was also addressed in two independent recent studies (Foulsham, Barton, Kingstone, Dewhurst, & Underwood, 2009; Mannan, Kennard, & Husain, 2009) that recorded eye movements of agnosia patients while they either free-viewed natural scenes (Mannan et al., 2009) or performed a search task in them (Foulsham et al., 2009). Visual agnosia patients have severe problems recognizing objects and understanding global scene properties and it is believed that they are impaired in applying top–down guidance when confronted with a visual scene because they are unable to link visual input with top–down knowledge. As a result, it was expected that their eye movements should conform more closely with the predictions of the saliency map model since top–down influences found in healthy observers, which necessarily degrade the model's performance, would be absent. Indeed, this was found in both studies. Notably, while normal observers performing the search task in the Foulsham et al. (2009) study were able to override low-level saliency effects (as discussed previously), this was not the case for the general agnosia patient studied who was apparently unable to use top–down guidance and relied on strategies dictated by bottom–up information. Fixation patterns in the agnosia patient were therefore significantly better predicted by the saliency map model (Itti et al., 1998) than for normal observers. A notable result of the Mannan et al. (2009) study is that the first few eye movements of their agnosia patients were indistinguishable from those made by the healthy observers and well-predicted by the saliency map model. It was only later that the fixation patterns of normal observers deviated from those of the agnosia patients (and the saliency map model), presumably because of top–down influences in the control group that were not available to the patients.

Related to top–down influences is the possibility that participants are selecting objects, rather than locations, as interesting. In a study that is complementary to ours, Elazary and Itti (2008) did not give observers an explicit instruction to search for "interesting" locations (our approach) but instead assumed that observers who were asked to label "objects" in a natural scene would predominantly do so for objects that they found "interesting." Using a large existing database of labeled images (available at http://labelme.csail.mit.edu), they find that low-level saliency, as computed from the Itti et al. (1998) model, is a highly significant predictor of which objects humans chose to label. The saliency map model finds a labeled object 76% of the time within the first three predicted locations. An anecdotal review of our results, with interest points plotted on top of each image, confirms that selections tend to cluster on and around objects, or complex structures in the case of fractals and natural scenes (e.g., Figure 3).

As has been theorized (e.g., Rensink, 2000a, 2000b; Treisman & Gelade, 1980), the formation of a stable object representation requires attention. However, information that is important for object processing, such as figure–ground segmentation (Driver & Baylis, 1996) or invariant features (Lowe, 2004), is not accounted for by the saliency model. This information is not top–down, *per se,* in the sense that the information is part of the image and does not reflect an idiosyncratic preference or task goal. For instance, the features of a car in an image would be the same whether or not the person is searching for a car, but these features are part of the image itself and could be pre-selected to aid in the detection of that object if that was the observer's goal. Thus, attention in natural scenes, as measured by interest point selections, may be guided by an intermediate stage between bottom–up and top–down information, which is important for the formation of stable object representations. Furthermore, recent physiological results show that figure–ground segregation does not depend on selective attention (Qiu, Sugihara, & von der Heydt, 2007). A theoretical model for a neural substrate of this and other mechanisms of intermediate vision has been suggested recently in a computational model (Craft, Schütze, Niebur, & von der Heydt, 2007).

Conclusions

The selection of interest points appears to be quite consistent across subjects, highly correlated with fixations, and influenced by both bottom–up and top–down attentional factors. More work is required to elucidate the relationship between interesting locations and semantically important objects in scenes. Interest points could also serve as a useful step for computationally modeling top–down influences on attention.

Acknowledgments

This work was supported by NIH-NEI 5R01EY016281-02. Please address correspondence to EN.

Commercial relationships: none.
Corresponding author: Ernst Niebur.
Email: niebur@jhu.edu.
Address: Johns Hopkins University, Mind/Brain Institute, 3400 N. Charles St., Baltimore, MD 21218, USA.

References

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics, 55,* 485–496. [PubMed]

Cerf, M., Cleary, D. R., Peters, R. J., Einhäuser, W., & Koch, C. (2007). Observers are consistent when rating image conspicuity. *Vision Research, 47,* 3052–3060. [PubMed]

Craft, E., Schütze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure–ground organization. *Journal of Neurophysiology, 97,* 4310–4326. [PubMed] [Article]

Driver, J., & Baylis, G. (1996). Edge-assignment and figure–ground segmentation in short-term visual matching. *Cognitive Psychology, 31,* 248–306. [PubMed]

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3):3, 1–15, http://journalofvision.org/8/3/3/, doi:10.1167/8.3.3. [PubMed] [Article]

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control setting. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 1030–1044.

Foulsham, T., Barton, J. S., Kingstone, A., Dewhurst, A., & Underwood, G. (2009). Fixation and saliency during search of natural scenes: The case of visual agnosia. *Neuropsychologia, 47,* 1994–2003. [PubMed]

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*(2):6, 1–17, http://journalofvision.org/8/2/6/, doi:10.1167/8.2.6. [PubMed] [Article]

Foulsham, T., & Underwood, G. (2009). Does conspicuity enhance distraction? Saliency and eye landing position when searching for objects. *Quarterly Journal of Experimental Psychology, 62,* 1088–1098. [PubMed]

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243–271. [PubMed]

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228.

Hoffman, J., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57,* 787–795. [PubMed]

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based fast visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1254–1259.

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

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60,* 91–110.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2,* 547–552.

Mannan, S. K., Kennard, C., & Husain, M. (2009). The role of visual salience in directing eye movements in visual object agnosia. *Current Biology, 19,* R247–R248. [PubMed]

Niebur, E., & Koch, C. (1996). Control of selective visual attention: Modeling the "where" pathway. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (vol. 8, pp. 802–808). Cambridge, MA: MIT Press.

O'Regan, J. K., Deubel, H., Clark, J. J., & Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition, 7,* 191–211.

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

Parkhurst, D., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16,* 125–154. [PubMed]

Parkhurst, D., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19,* 783–789. [PubMed]

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3–25.

Posner, M. I., Rafal, R. D., Choate, L., & Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology, 2,* 211–228.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 970–982.

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure–ground mechanisms provide structure for selective attention. *Nature Neuroscience, 10,* 1492–1499. [PubMed] [Article]

Rensink, R. A. (2000a). Seeing, sensing, and scrutinizing. *Vision Research, 40,* 1469–1487. [PubMed]

Rensink, R. A. (2000b). The dynamic representation of scenes. *Visual Cognition, 7,* 17–42.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368–373.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25,* 31–40.

Rizzolatti, G., Riggio, L., & Shelgia, B. M. (1994). Space and selective attention. In C. Umiltá & M. Moscovitch (Eds.), *Attention and performance* (vol. XV, pp. 231–265). Cambridge, MA: MIT Press.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53–65.

Schmid, C., Mohr, R., & Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision, 37,* 151–172.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51,* 599–606. [PubMed]

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

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7):4, 1–16, http://journalofvision.org/9/7/4/, doi:10.1167/9.7.4. [PubMed] [Article]

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59,* 1931–1949. [PubMed]

von Helmholtz, H. (1867). *Handbuch der physiologischen optik*. Leipzig: Voss.

Wolf, H., & Deng, D. (2005). How interesting is this? Finding interest hotspots and ranking images using an MPEG-7 visual attention model. In *17th Annual Colloquium of the Spatial Information Research Centre* (pp. 67–76). Dunedin, New Zealand.

Wolfe, J. M. (1994). Guided Search 2.0—A revised model of visual search. *Psychonomic Bulletin & Review, 1,* 202–238.

Wright, M. J. (2005). Saliency predicts change detection in pictures of natural scenes. *Spatial Vision, 18,* 413–430.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 601–621.