

**Visual Saliency Does Not Account for Eye Movements during
Visual Search in Real-World Scenes**

John M. Henderson^{1,3}, James R. Brockmole^{1,3},
Monica S. Castelhana^{1,3}, Michael Mack^{2,3}

¹Department of Psychology, ²Department of Computer Science,
³Cognitive Science Program

Chapter to appear in:

Roger van Gompel, Martin Fischer, Wayne Murray, and Robin Hill (Eds). *Eye Movement Research: Insights into Mind and Brain*. Elsevier.

Revision: December 19, 2005

Corresponding Author:

John M. Henderson
Department of Psychology
285B Psychology Building
Michigan State University
East Lansing, MI 48824-1116
USA

john@eyelab.msu.edu
Voice: +1-517-432-3367
Fax: +1-517-353-3745

Abstract

We tested the hypothesis that fixation locations during scene viewing are primarily determined by visual salience. Eye movements were collected from participants who viewed photographs of real-world scenes during an active search task. Visual salience as determined by a popular computational model did not predict region-to-region saccades or saccade sequences any better than did a random model. Consistent with other reports in the literature, intensity, contrast, and edge density differed at fixated scene regions compared to regions that were not fixated, but these fixated regions also differ in rated semantic informativeness. Therefore, any observed correlations between fixation locations and image statistics cannot be unambiguously attributed to these image statistics. We conclude that visual saliency does not account for eye movements during active search. The existing evidence is consistent with the hypothesis that cognitive factors play the dominant role in active gaze control.

During real-world scene perception, we move our eyes about three times each second via very rapid eye movements (*saccades*) to reorient the high-resolving power of the fovea. Pattern information is acquired only during periods of relative gaze stability (*fixations*) due to a combination of central suppression and visual masking (Matin, 1974; Thiele, Henning, Buishik, & Hoffman, 2002; Volkman, 1986). *Gaze control* is the process of directing the eyes through a scene in real time in the service of ongoing perceptual, cognitive, and behavioral activity (Henderson & Hollingworth, 1998, 1999; Henderson, 2003).

There are at least three reasons that the study of gaze control is important in real-world scene perception (Henderson, 2003; Henderson & Ferreira, 2004a). First, human vision is active in the sense that fixation is directed toward task-relevant information as it is needed for ongoing visual and cognitive computations. Although this point seems obvious to eye movement researchers, it is often overlooked in the visual perception and visual cognition literatures. For example, much of the research on real-world scene perception has used tachistoscopic display methods in which eye movements are not possible (though see Underwood, this section; Gareze & Findlay, this section). While understanding what is initially apprehended from a scene is an important theoretical topic, it is not the whole story; vision naturally unfolds over time and multiple fixations. Any complete theory of visual cognition, therefore, requires understanding how ongoing visual and cognitive processes control the direction of the eyes in real time, and how vision and cognition are affected by where the eyes are pointed at any given moment in time.

Second, eye movements provide a window into the operation of selective attention. Indeed, although internal (covert) attention and overt eye movements can be dissociated (Posner & Cohen, 1984), the strong natural relationship between covert and overt attention has recently led some investigators to suggest that studying covert visual attention independently of overt attention is misguided (Findlay, 2004; Findlay & Gilchrist, 2003). For example, as Findlay and Gilchrist (2003) have noted, much of the research in the visual search literature

has proceeded as though viewers steadfastly maintain fixation during search, allocating attention only via an internal mechanism. However, visual search is virtually always accompanied by saccadic eye movements (e.g., see the chapters by Hooge, Vlaskamp, & Over, this section; Shen & Reingold, this section). In fact, studies of visual search that employ eye tracking often result in different conclusions than do studies that assume the eye remain still. As a case in point, eye movement records reveal a much richer role for memory in the selection of information for viewing (e.g. McCarley, Wang, Kramer, Irwin, & Peterson, 2003; Peterson, Kramer, Wang, Irwin, & McCarley, 2001) than research that uses more traditional measures such as reaction time (e.g. Horowitz & Wolfe, 1998). To obtain a complete understanding of the role of memory and attention in visual cognition, it is necessary to understand eye movements.

Third, because gaze is typically directed at the current focus of analysis (see Irwin, 2004, for some caveats), eye movements provide an unobtrusive, sensitive, real-time behavioral index of ongoing visual and cognitive processing. This fact has led to enormous insights into perceptual and linguistic processing in reading (Liversedge & Findlay, 2000; Rayner, 1998; Sereno & Rayner, 2003), but eye movements are only now becoming a similarly important tool in the study of visual cognition generally and scene perception in particular.

Fixation Placement during Scene Viewing

A fundamental goal in the study of gaze control during scene viewing is to understand the factors that determine where fixation will be placed. Two general hypotheses have been advanced to explain fixation locations in scenes. According to what we will call the *visual saliency hypothesis*, fixations sites are selected based on image properties generated in a bottom-up manner from the current scene. On this hypothesis, gaze control is to a large degree a reaction to the visual properties of the stimulus confronting the viewer. In contrast,

according to what we will call the *cognitive control hypothesis*, fixation sites are selected based on the needs of the cognitive system in relation to the current task. On this hypothesis, eye movements are primarily controlled by task goals interacting with a semantic interpretation of the scene and memory for similar viewing episodes (Hayhoe & Ballard, 2005; Henderson & Ferreira, 2004a). On the cognitive control hypothesis, the visual stimulus is of course still relevant: The eyes are typically directed to objects and features rather than to uniform scene areas (Henderson & Hollingworth, 1999); however, the relevance of a particular object or feature in the stimulus is determined by cognitive information-gathering needs rather than inherent visual salience.

The visual saliency hypothesis has generated a good deal of interest over the past several years, and in many ways has become the dominant view in the computational vision literature. This hypothesis has received primary support from two lines of investigation. First, computational models have been developed that use known properties of the visual system to generate a *saliency map* or landscape of visual salience across an image (Itti & Koch, 2000, 2001; Koch & Ullman, 1985). In these models, the visual properties present in an image give rise to a 2D map that explicitly marks regions that are different from their surround on image dimensions such as color, intensity, contrast, and edge orientation (Itti & Koch, 2000; Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002; Torralba, 2003), contour junctions, termination of edges, stereo disparity, and shading (Koch & Ullman, 1985), and dynamic factors such as motion (Koch & Ullman, 1985; Rosenholtz, 1999). The maps are generated for each image dimension over multiple spatial scales and are then combined to create a single saliency map. Regions that are uniform along some image dimension are considered uninformative, whereas those that differ from neighboring regions across spatial scales are taken to be potentially informative and worthy of fixation. The saliency map approach serves an important heuristic function in the study of gaze control because it provides an explicit model that generates precise quantitative predictions about fixation locations and their

sequences, and these predictions have been found to correlate with observed human fixations under some conditions (e.g., Parkhurst et al., 2002).

Second, using a *scene statistics* approach, local scene patches surrounding fixation points have been analyzed to determine whether fixated regions differ in some image properties from regions that are not fixated. For example, high spatial frequency content and edge density have been found to be somewhat greater at fixated than non-fixated locations (Mannan, Ruddock, & Wooding, 1996, 1997b). Furthermore, local contrast (the standard deviation of intensity in a patch) is higher and two-point intensity correlation (intensity of the fixated point and nearby points) is lower for fixated scene patches than control patches (Krieger, Rentschler, Hauske, Schill, & Zetsche, 2000; Parkhurst & Neibur, 2003; Reinagel & Zador, 1999).

Modulating the evidence supporting the visual saliency hypothesis, recent evidence suggests that fixation sites are tied less strongly to saliency when meaningful scenes are viewed during active viewing tasks (Land & Hayhoe, 2001; Turano, Gerguschat, & Baker 2003). According to one hypothesis, the modulation of visual salience by knowledge-driven control may increase over time within a scene-viewing episode as more knowledge is acquired about the identities and meanings of previously fixated objects and their relationships to each other and to the scene (Henderson, Weeks, & Hollingworth, 1999). However, even the very first saccade in a scene can often take the eyes in the likely direction of a search target, whether or not the target is present, presumably because the global scene gist and spatial layout acquired from the first fixation provide important information about where a particular object is likely to be found (Antes, 1974; Brockmole & Henderson, under review; Castelhamo & Henderson, 2003; Henderson et al., 1999; Mackworth & Morandi, 1967).

Henderson and Ferreira (2004a) sorted the knowledge available to the human gaze control system into several general categories. Information about a specific scene can be learned over the short term from the current perceptual encounter (*short-term episodic scene*

knowledge) and over the longer term across multiple encounters (*long-term episodic scene knowledge*). Short-term knowledge underlies a viewer's tendency to refixate areas of the current scene that are semantically interesting or informative (Buswell, 1935; Loftus & Mackworth, 1978; Henderson et al., 1999; Yarbush, 1967), enables the prioritization of newly appearing or disappearing objects from a scene (Brockmole & Henderson, 2005, in press), and ensures that objects are fixated when needed during motor interaction with the environment (Land & Hayhoe, 2001). Long-term episodic knowledge involves information about a particular scene acquired and retained over time. Recent evidence suggests that good memory for the visual detail of fixated regions of a viewed scene is preserved over relatively long periods of time (Castelhano & Henderson, 2005; Henderson & Hollingworth, 2003; Hollingworth & Henderson, 2002). The contextual cueing phenomenon shows that perceptual learning of complex visual images can take place relatively rapidly over multiple encounters (Chun & Jiang, 1998), and this effect has been shown to influence eye movements (Peterson & Kramer, 2001). We have recently found that this same type of learning can take place even more rapidly for real-world scenes (Brockmole & Henderson, 2006). Furthermore, we have shown that these learned representations can facilitate eye movements during search in real-world scenes (Brockmole & Henderson, submitted). Another interesting example of the influence of episodic scene knowledge on gaze control is the finding that viewers will often fixate an empty scene region when that region previously contained a task-relevant object (Altmann, 2004; Richardson & Spivey, 2000).

A second source of information that can guide gaze is *scene schema knowledge*, generic semantic and spatial knowledge about a particular category of scene (Biederman, Mezzanotte, & Rabinowitz, 1982; Friedman, 1979; Mandler & Johnson, 1977). Schema knowledge includes information about the objects likely to be found in a specific type of scene (e.g., bedrooms contain beds), and spatial regularities associated with a scene category (e.g., pillows are typically found on beds), as well as generic world knowledge about scenes (e.g.,

beds do not float in the air). Scene identity can be apprehended and a scene schema retrieved very rapidly (Potter, 1976; Schyns & Oliva, 1994), and schema knowledge can then be used to limit initial fixations to scene areas likely to contain an object relevant to the current task (Henderson et al., 1999).

A third source of information important in gaze control is *task related knowledge* (Buswell, 1935; Yarbus, 1967). Task related knowledge can involve a general *gaze control policy* or strategy relevant to a given task, such as periodically fixating the reflection in the rear-view mirror while driving, and moment-to-moment control decisions based on ongoing perceptual and cognitive needs. Gaze control differs during complex and well-learned activities such as reading (Rayner, 1998), tea and sandwich making (Land & Hayhoe, 2001), and driving (Land & Lee, 1997). The distribution of fixations over a given scene changes depending on whether a viewer is searching for an object or trying to memorize that scene (Henderson et al., 1999). Gaze is also strongly influenced by moment-to-moment cognitive processes related to spoken language comprehension and production (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see Henderson & Ferreira, 2004b).

Present Study

As reviewed above, there is abundant evidence that fixation placement during scene viewing is strongly affected by cognitive factors. Most proponents of the visual saliency hypothesis acknowledge that cognitive factors play a role in gaze control, but they tend to focus on the adequacy of a saliency-based approach to account for much of the data on fixation placement (e.g., Parkhurst et al., 2002).

In the present study, we investigated further the degree to which fixation location is related to image properties during scene viewing. First, we collected eye movement data from participants who viewed full-color photographs of real-world outdoor scenes while engaged in

a visual search task in which they counted the number of people who appeared in each scene. We then analyzed the fixation data in three ways to investigate the adequacy of the visual saliency hypothesis. First, we compared the fixation data against the predictions generated from an established visual saliency model. Second, we conducted an image statistics analysis to determine whether image properties differ at fixated and non-fixated locations. Third, we tested whether any observed correlations between fixation locations and image statistics might be due to the meaning of fixated locations. Our conclusion is that the evidence supporting the visual saliency hypothesis is weak, and that the existing evidence is consistent with the hypothesis that cognitive factors play the dominant role in gaze control.

Method

The eye movements of eight Michigan State University undergraduates were monitored as they viewed 36 full-color photographs of real-world outdoor scenes displayed on a computer monitor. The photographs were shown at a resolution of 800 by 600 pixels and subtended 16 deg horizontally x 12 deg vertically at a viewing distance of 113 cm. Eye position was sampled at a rate of 1000 Hz from a Fourward Technologies Generation 5.5 Dual Purkinje Image Eyetracker, and raw eyetracking data were parsed into fixations and saccades using velocity and distance criteria (Henderson et al., 1999). The subject's head was held steady with an anchored bite-bar made of dental impression compound. Prior to the first trial, subjects completed a procedure to calibrate the output of the eyetracker against spatial position on the display screen. This procedure was repeated regularly throughout the experiment. Observers were instructed to count the number of people in each photograph. Each participant saw all 36 scenes in a different random order. Each photograph contained between 0 and 6 people and was presented until the participant responded or for 10 sec maximum. Across all search photographs, accuracy on the counting task was 82%, with greater accuracy for scenes with fewer targets present. Accuracy was below 100% because

some targets were well hidden and difficult to find in the scenes.

Analysis 1: Comparing Saliency Model Predictions to Human Fixations

A benchmark for the visual saliency hypothesis is the saliency map model of Itti & Koch (2000, 2001). This model produces explicit predictions about where viewers should fixate in complex images. The Itti and Koch model has been shown to predict human fixations reasonably well under some conditions (e.g., Parkhurst et al., 2002), though Turano et al. (2003) demonstrated that the correlations between the model and human fixations were eliminated when the viewing task was active. However, one could argue that this latter result was a consequence of the dynamic interaction between a moving viewer and the real world, a situation for which the model was not specifically developed. In Analysis 1, we examined the degree to which the Itti and Koch saliency map model is able to predict fixation locations in static scenes (the situation for which it was developed) during an active visual search task.

Do Human Fixations Fall on the Most Visually Salient Regions? In a first analysis we compared the number of saccadic entries into, and the number of discrete fixations in the scene regions that the saliency map model specified as most salient. For this analysis, the Itti and Koch bottom-up saliency model posted on the website [<http://ilab.usc.edu/toolkit/downloads.shtml>] May 16, 2005, was used to determine the bottom-up salient regions in each of our test scenes. The model “viewed” the scenes for 10 seconds each (the same amount of time given to the participants) with a foveal radius of 1 degree. While viewing each scene, the model generated a cumulative saliency map showing the scene regions that it found most salient over the 10 seconds of viewing. Any region in these cumulative saliency maps that held a value greater than zero was defined as a salient region in the eye movement analysis. To better understand the relationship between the regions the Itti and Koch model found salient and the regions participants fixated while viewing the scenes, two measures of the participants' eye movements were examined, *Salient*

Region Entries and Salient Region Fixations. Salient Region Entries was defined as the proportion of all participant-generated saccades that started outside of a given salient region and landed in that region. This measure captures the degree to which the eyes tended to saccade to salient regions. Salient Region Fixations was defined as the proportion of all participant-generated fixations that fell in a given salient region. This measure reflects all fixations in a salient region regardless of whether the fixation was due to a saccade from beyond that region or within that region. As control contrasts, random fixations of equal number to the participants' were generated by two random models. The first model (*pure random*) simply sampled from all possible fixation locations. To control for the participants' bias to fixate in the lower central regions of the scenes (see Figure 2), a second control contrast (*biased random*) used randomly generated fixations based on the probability distribution of fixation locations from the participants' eye movement behavior across all the scenes.

Salient Region Entries. The first analysis was carried out by taking the proportion of saccades that entered a salient region (see Figure 1 for an example). All saccades from the participant trials and an equal number of saccades from the pure- and biased-random models were included in the analysis. A higher proportion of salient region entries means that a greater number of saccades were made into salient regions. If the saliency map model is able to identify regions that capture attention better than chance, the proportion of salient region entries made by participants should be higher than the proportion of salient region entries made by the random models. If the proportion of salient region entries does not differ between participants and the random models, participants are no more likely to saccade to regions identified by the model than they are by chance. As can be seen in Figure 3A, although the saliency map model predicted entries better than did a pure random model, $t(34) = 3.64, p < .001$, it did not predict entries better than a biased-random model that took into account participants' general tendency to fixate the lower and central regions of all scenes, $t(34) < 1$.

Contrary to the visual saliency hypothesis, participants' tendency to move their eyes to specific scene regions was not accounted for by the saliency model.

Salient Region Fixations. The second analysis was carried out by taking the proportion of all fixations that landed in a salient region (see Figure 1 for an example). All fixations from the participant trials and an equal number of fixations from the two random models were included in the analysis. Once again, if participants were to have a higher proportion of salient region fixations than the random model, this would suggest that the model is finding regions that capture attention better than chance. On the other hand, if the proportion of salient region fixations does not differ between participants and the random models, this would suggest that the model is finding regions that are no more likely to be fixated than by chance. Observers fixated salient regions identified by the model more often predicted by the pure random model, $t(34) = 6.81, p < .001$, and the biased-random model, $t(34) = 4.02, p < .001$, indicating that the saliency model predicted the number of fixations in salient regions more accurately than models based on chance.

Summary. The Salient Region Entries analysis demonstrates that viewers were no more likely to saccade to a salient scene region (as identified by the saliency map model) than they were by chance. On the other hand, the Salient Region Fixations analysis shows that viewers fixated salient regions more often than would be expected by chance. Together, these data suggest that although the eyes are not specifically attracted to salient regions, they do tend to stick to them once there. The latter result might be taken as at least partial support for the saliency control hypothesis. However, because this hypothesis is supposed to account for the movement of the eyes through a scene rather than the tendency to dwell in a given region, the support is weak. Furthermore, as detailed below, the latter result is also consistent with the possibility that saliency is correlated with "object-ness", and that viewers tend to gaze at objects.

Do Human Fixations Correspond with Model-Generated Fixation Predictions?

In addition to generating a map of salient scene regions, the saliency model also produces a set of fixations. Therefore, a second way to test the ability of the model to predict human fixations is to compare directly the human- and model-generated fixation locations. We quantified the distance between these fixation locations in two ways, one based on a similarity metric devised by Mannan, Ruddock, & Wooding (1995), and a second that we developed as an extension of this metric.

Mannan, Ruddock, & Wooding (1995) Similarity Metric. The fixation location similarity metric introduced by Mannan et al. (1995) compares the spatial proximity of fixations derived from two unique fixation sets (e.g. model generated and observer generated). The location similarity metric compares the linear distance from one set of fixation locations to the closest fixation in the other set, and vice versa. A high score indicates high similarity. As a control, we also computed the same similarity metric for all pairwise comparisons among participants. If the saliency map model is able to predict the locations of human fixations, its similarity to human observers should be comparable to the similarity of one human viewer to another.

The index of similarity (I_s) introduced by Mannan et al. is based on the squared distances between corresponding fixations in two gaze patterns (D_m and D_{mr}) and is defined in the following manner:

$$I_s = 100 \left[1 - \frac{D_m}{D_{mr}} \right], \quad (1)$$

with

$$D_m^2 = \frac{n_1 \sum_{j=1}^{n_2} d_{2j}^2 + n_2 \sum_{i=1}^{n_1} d_{1i}^2}{2n_1n_2(w^2 + h^2)}, \quad (2)$$

where n_1 and n_2 are the number of fixations in the two gaze patterns, d_{1i} is the distance between the i th fixation in the first gaze pattern and its nearest neighbor fixation in the second

gaze pattern, d_{2j} is the distance between the j th fixation in the second gaze pattern and its nearest neighbor fixation in the first gaze pattern, and w and h are the width and height of the image of the scene. The calculation of D_{mr} is the same as D_m but with randomly generated gaze patterns of the same size being compared. Similar to a correlation, identical gaze patterns produce an I_s score of 100, random gaze patterns produce an I_s score of 0, and systematically different gaze patterns generate a negative score (Mannan et al., 1995). For our analysis, we examined the first seven fixations each participant produced when viewing each scene and compared them against the first seven fixations produced by the saliency model.

Figure 4A shows the mean similarity score I_s for each participant against all other participants (left bar) and all participants against the model (right bar). As can be seen in the figure, the participants' fixations were significantly less similar to those generated by the saliency model than they were to each other, $t(35) = 7.87, p < .001$.

A Unique Assignment Variant of the Mannan et al. (1995) Metric. A potential concern with the Mannan et al. (1995) similarity metric is that it does not take into account the overall spatial variability in the distribution of fixations over an image. For example, if all of the fixations in Set 1 are clustered in one small region of a scene, and there is at least one fixation in that same region in comparison Set 2, all the Set 1 fixations will be compared against that single Set 2 fixation. Another way to compute similarity in the same spirit as the Mannan et al. method that corrects for this issue is to require that each fixation in each set be assigned to a unique fixation in the other set. A metric can then be computed based on the distance of each point in Set 1 to its assigned point in Set 2. Intuitively, this unique-assignment metric better takes into account the overall spatial distributions of fixations. (Unlike the Mannan et al. analysis, this method requires that there be an identical number of fixations in each set.) In our unique-assignment analysis, all possible assignments of each fixation in Set 1 to a unique fixation in Set 2 were examined to find the single assignment that produced the smallest average deviation. This assignment was then used to compute the

similarity metric, which is the squared deviation of each fixation point in Set 1 to its mate in Set 2.

More precisely, unique-assignment distance (W_s) between two gaze patterns (D_m and D_{mr}) was defined as:

$$W_s = 100 \left[1 - \frac{D_w}{D_{wr}} \right], \quad (3)$$

with

$$D_w = \frac{1}{n} \sum_{j=1}^n p_j^2, \quad (4)$$

where n is the number of fixations in the gaze patterns, p_j is the distance between the j th unique pair of one fixation from the first set and one fixation from the second set. The calculation of D_{wr} is the same as D_w except that randomly generated gaze patterns of the same size are compared. Identical gaze patterns produce an W_s score of 100, random gaze patterns produce an W_s score of 0, and systematically different patterns generate negative scores.

Again, as a contrast, we also computed the unique-assignment similarity metric for all participants against all other participants. If the saliency model is able to predict human fixations, its similarity to human observers should be comparable to the similarity for all pairwise comparisons of participants. As above, we restricted the analysis to the first seven fixations each participant produced when viewing each scene and the first seven fixations produced by the saliency model.

Figure 4B shows the mean similarity score W_s for each participant against all other participants (left bar) and for all participants against the model (right bar). As can be seen in the figure, as with the first similarity metric, the fixations generated by the saliency model were significantly less similar to those of the participants than were those of the participants to each other, $t(35) = 5.27, p < .001$.

Are Human Fixation Sequences Predicted by Model Fixation Sequences? Both

the original Mannan et al. (1995) similarity metric and our unique-assignment variant of it ignore information about fixation sequence. In the case of the Mannan et al. (1995) metric, there is no requirement that fixations be assigned in a one-to-one correspondence across sets, and in the unique-assignment variant, the correspondence is based purely on spatial proximity and so does not take into account the temporal order in which the fixations were produced. It could be that the saliency model does a better job of predicting fixation order (scan pattern) than it does the exact locations of fixations. To investigate this possibility, we computed the Levenshtein Distance, a similarity metric specifically designed to capture sequence. The analysis uses a set of basic transformations to determine the minimum number of steps (character insertion, deletion, and substitution) that would be required to transform one character string into another. This general method is used in a variety of situations including protein sequencing in genetics (Levenshtein, 1966; Sankhoff & Kruskal, 1983). To conduct the analysis, we divided each scene into a grid of 48 regions of about 2 deg x 2 deg each. This division allowed some noise in fixation location so that minor deviations from the model would not disadvantage it. Each of the 48 regions was assigned a unique symbol. Each fixation was coded with the symbol assigned to the region in which it fell. We again analyzed the first seven fixations, so each fixation sequence produced a 7-character string. The similarity metric between two strings was the number of steps required to transform one string into another. Identical strings generated a value of 0, and the maximum value was 7. As in the first two analyses, we computed the similarity of each subject's fixation sequence for each scene to the sequence generated for that scene by the model. Again, as a control, we also computed the string metric for all participants against all other participants for each scene. If the saliency model is able to predict human fixations, its similarity to human observers should be comparable to the similarity of the human participants to each other. Figure 4C shows the mean distance score for each participant against all other participants (left bar) and for all participants against the model (right bar). The fixation sequences generated by the saliency

model were significantly less similar to those of the participants than were those of the participants to each other, $t(35) = 10.2, p < .001$.

Saliency Map Model Comparison Summary. In a first set of analyses, we tested the ability of an implemented saliency map model to predict human fixation locations during an active viewing task. Overall, the results suggested that the model does not do a particularly good job. Human fixations did not land in regions of a scene that the model considered to be visually salient, and the similarity of the participants' fixations to each other were much greater than the similarity of the participants' fixations to model-generated fixations. Of course, the ability of a given model to predict human performance is a function both of those aspects of the model that are theory-inspired and other incidental modeling decisions required for the implementation. One could argue that the spirit of the model is correct, but that the implementation is not. Similarly, one might argue that the implementation is correct, but that specific parameter choices are not. However, it is important to remember that this version of the model has been reported to predict human fixation locations reasonably well under other conditions (Parkhurst et al., 2002; Parkhurst & Neibur, 2003, 2004). The model seems to do a particularly good job with meaningless patterns (such as fractals) and in relatively unstructured viewing tasks. In this context, the present results can be taken to suggest that whereas visual salience (as instantiated in the Itti and Koch saliency map model) does a reasonable job of accounting for fixation locations under some circumstances, it does a poor job when the viewing task involves active search and the image is a real-world scene.

Analysis 2: Measuring Local Image Statistics at Fixated Locations

Several studies have demonstrated that the image properties of fixated regions tend to differ in systematic ways from regions that are not fixated (Krieger, et al., 2000; Mannan, et al., 1995, 1996; Mannan, Ruddock, & Wooding, 1997a; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). Specifically, fixated scene regions tend to be lower in intensity but higher in

edge density and local contrast, and are more likely to contain third-order spatial relationships such as T-junctions and curves, than non-fixated regions. These results have been taken to suggest that such regions act as “targets” for fixations. Do these results generalize to an active visual task with real-world scenes?

Scene Statistics Method. In the present study, we measured the local image statistics associated with the fixations generated by our viewers, and compared those values to the values associated with randomly selected scene locations (see Parkhurst et al., 2002). For each scene image, ensembles of *image patches* were created. These patches had a radius of 1 degree of visual angle, approximating the spatial extent of foveal vision. Three different types of ensembles were created. In the *subject ensemble*, patches were defined by the subject-selected fixation positions within each image. That is, the center of each patch was defined by the (x,y) coordinates of each fixation. Thus, the subject ensemble was completely constrained by subject behavior. In the *random ensemble*, patches were centered on randomly selected positions within each scene. Thus, the patches in the random ensemble were completely unconstrained and every point in the image was equally likely to be selected. In the *shuffled ensemble*, patches were derived by “shuffling” subject-selected fixation locations from one image onto a different, randomly selected image. Like the biased random control condition in Analysis 1, this shuffled ensemble was used to account for the participants’ bias to fixate more centrally in an image (see Parkhurst & Niebur, 2003).

For each ensemble, several measures of local image statistics were calculated. Analyses then focused on evaluating the similarity of the image statistics within each type of ensemble. Image statistics within the subject ensemble are characteristic of those image properties that are fixated. Since it is a random sampling, image statistics within the random ensemble are characteristic of the image properties in the scenes overall. Image statistics within the shuffled ensemble are characteristic of the image properties in those scene regions that tend to be fixated across scenes, such as the lower scene center (Figure 2). The degree to

which the scene statistics of the subject ensembles differ from the random and shuffled ensembles indicates the extent to which fixation location is correlated with particular image statistics.

Three common measures of local image statistics were examined: intensity, contrast, and edge density. These image statistics characterize different properties of image luminance. The luminance of each scene was extracted by converting the scene's RGB values to the CIE $L^*a^*b^*$ colorspace (Oliva & Schyns, 2000) which separates the luminance information of an image into a distinct dimension (L^*). The chromatic information in the a^* and b^* dimensions was discarded, and analyses focused on the values in the L^* dimension. Intensity was defined as the average luminance value of the pixels within a patch (see Mannan et al., 1995). Greater intensity is associated with higher luminance, or a higher degree of subjectively perceived brightness. Local contrast was defined as the standard deviation of luminance within a patch (see Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). Local contrast, then, is a measure of how much the luminance values of pixels within a patch vary from each other. More uniform patches have less contrast. Edge density was defined as the proportion of edge pixels within an image patch. Edge pixels were found by filtering the scenes with a Sobel operator that responds to contours in scenes represented by steep gradients in luminance (see Mannan et al., 1995, 1996). Greater edge density is associated with image patches containing a greater number of contours.

Results. Representative patches from the subject, random, and shuffled ensembles are depicted in Figure 5. Quantitative analyses of the local image statistics available in patches from each ensemble are summarized in Figure 6. For all analyses, the local image statistics observed at fixation (the subject ensemble) were tested against the random and shuffled ensembles using one-sample t-tests.

Consistent with the prior findings in the literature cited earlier, patches derived from the subject ensembles were reliably different from those from the shuffled and random

ensembles for all three local image statistics. Intensity within the subject ensemble patches was 6% lower than that within the shuffled patches [$t(284) = -3.88, p < .001$], and 8% lower than that in the random patches [$t(284) = -6.55, p < .001$]. Local contrast within the subject ensemble patches was 14% higher than that within the shuffled patches [$t(284) = 6.72, p < .001$] and 9% higher than that in the random patches [$t(284) = 7.59, p < .001$]. Edge density within the subject ensemble patches was 19% higher than that within the shuffled patches [$t(284) = 8.03, p < .001$], and 29% higher than that in the random patches [$t(284) = 15.5, p < .001$].

Summary. Replicating prior results, observers fixated regions that were lower in intensity and higher in local contrast and edge density than either control regions selected randomly or based on fixations from another image. On the face of it, these data could be taken to suggest that regions marked by differences in local image properties compared to the remainder of the scene act as “targets” for fixation, irrespective of the semantic nature of the information contained in those regions (Parkhurst et al., 2002; Parkhurst & Neibur, 2003). However, because these analyses only establish a correlation between fixation locations and image properties, it is also possible that the relationship is due to other factors. In the following section we explore the hypothesis that region meaning is such a factor. Specifically, we measured the semantic informativeness of the subject, shuffled, and random ensembles to determine whether meaning was also correlated with fixation location.

Analysis 3: Are Fixated Scene Regions More Semantically Informative?

The purpose of this analysis was to determine whether fixated regions that have been shown to differ from non-fixated regions in intensity, contrast, and edge density, also differ in semantic informativeness. To investigate this question, an independent group of observers rated the degree to which patches from each ensemble were semantically informative (Antes, 1974; Mackworth & Morandi, 1967).

One hundred patches were selected from each of the subject, shuffled, and random ensembles generated from the scene statistics analysis reported above. These patches met two constraints. First, patches had to be representative of their ensemble supersets (subject, shuffled, random) in terms of local image statistics (as determined above) so that the reliable statistical differences observed would be preserved. Second, a minimum distance of 2 degrees of visual angle was established between the center points of any two patches originating from the same scene so that selected patches could not spatially overlap. Within these constraints, patches were chosen randomly. Patches from all scenes and subjects were represented in the final subset used in Analysis 3.

Seven Michigan State University undergraduates viewed all 300 selected patches on a computer monitor. Stimuli for presentation were created by placing each patch in the center of a uniform gray background that subtended 16 degrees horizontally and 12 degrees vertically. Each individual patch subtended 2 degrees horizontally and vertically. Presentation order of patches was randomly determined. Using a 7-point Likert-type scale, observers were instructed to rate how well they thought they could determine the overall content of the scene from the small view they were shown.

Results. Mean ratings for each patch type are illustrated in Figure 7. Patches from the subject, shuffled, and random ensembles received mean ratings of 4.65, 4.25, and 4.11, respectively. A one-way repeated-measures ANOVA demonstrated a reliable effect of patch type, $F(2, 12) = 19.4$, $p < .001$, with all pairwise comparisons reliable. Critically, patches from the subject-ensemble were judged to be more informative of scene identity than those from the shuffled and random ensembles. This analysis demonstrates that observers in the eye tracking experiment fixated scene regions that were more likely to provide meaningful information about the scene. These results challenge the hypothesis that local scene statistics and semantic informativeness are independent.

Summary. Image statistics of areas selected for eye fixation within scenes differ in

systematic ways from areas that are not fixated. A possible interpretation of these results is that fixation position can be accounted for by low-level image statistics (Krieger et al., 2000; Mannan et al., 1995, 1996, 1997a; Reinagel and Zador, 1999; Parkhurst & Niebur, 2003). The present results, however, call this interpretation into question. We conclude that examining the relationship between image statistics and fixation location without also measuring the semantic content of fixated regions can provide a partial or even misleading characterization of bottom-up influences on gaze control. Though it is possible that image properties in a scene directly influence gaze control, the results from scene statistics analyses cannot be taken as strong support for it.

General Discussion

Gaze control during scene perception is critical for timely acquisition of task-relevant visual information. In this study, we tested the hypothesis that the selection of locations for fixation during complex scene viewing is primarily driven by visual salience derived from a bottom-up analysis of image properties. To test this visual saliency hypothesis, we collected eye movement data from participants who viewed full-color photographs of real-world scenes during an active visual search task. We then analyzed the eye movement data in a variety of ways to test the visual saliency hypothesis. In an initial set of analyses, we compared the fixation data against the predictions generated from what is arguably the standard among computational visual saliency models. We found that visual salience, as instantiated by the model, did a poor job of predicting either fixation locations or sequences. In a second set of analyses, we examined whether image properties differ at fixated and non-fixated locations. Consistent with other reports in the literature, we found clear differences in intensity, contrast, and edge density at fixated scene regions compared to regions that were not fixated. However, in a third analysis, we showed that fixated regions also differ in rated meaning compared to regions not fixated. Therefore, any observed correlations between fixation locations and image

statistics could be due to the informativeness of fixated locations rather than to differences in the image statistics themselves. Our conclusion is that the evidence supporting the visual saliency hypothesis is weak, and that the existing evidence is consistent with the hypothesis that cognitive factors play the dominant role in gaze control.

Visual Saliency or Cognitive Control?

To what extent is there strong evidence that gaze is primarily controlled by visual salience? As we have shown, the main sources of evidence, correlation of fixation positions with model-determined visual saliency, and differences in scene statistics at fixated and non-fixated locations, are both problematic. In the case of correlations with saliency model output, there is very little evidence that for active viewing tasks in the real world, existing saliency models do a good job of predicting fixation locations (Turano et al, 2003). In the present study, we showed that an existing saliency model also does a poor job of predicting fixation locations during active visual search in static images.

In the case of analyses showing that image statistics differ at fixated and non-fixated locations, our results suggest that previously reported effects may just as well be due to differences in region meaning as to differences in the image statistics themselves. This confound is probably unavoidable: meaningful objects differ from scene background in their image properties. The important conclusion is that showing differences in image properties at fixated and non-fixated regions cannot be used as unambiguous support for the hypothesis that those properties are driving eye movements. It is at least as likely that the meaning of an image region is responsible for the fact that it was fixated.

The few prior attempts to investigate a causal link between image statistics and fixation point selection have produced mixed results. Mannan and colleagues (1995, 1996) demonstrated that fixation locations in normal and low-pass filtered scenes are similar over the first 1.5 s of viewing. Because the objects in many of the low-pass filtered scenes were not

identifiable, these data suggest that human gaze control does not initially select fixation sites based on object identity information. However, Einhauser and König (2003) demonstrated that perceptually detectable modifications to contrast have no effect on fixation point selection, suggesting that contrast does not contribute causally to fixation point selection, though this study has been criticized on methodological grounds (Parkhurst & Niebur, 2004). Nevertheless, Einhauser and König (2003) concluded that top-down, rather than bottom-up, factors determined attentional allocation in natural scenes.

Given that local image statistics associated with semantically informative regions such as objects undoubtedly differ systematically from those of backgrounds, and given our demonstration that such relationships exist for fixated scene regions, the results obtained by investigations linking local image statistics and gaze are entirely consistent with the conclusion that cognitive factors guide eye movements through scenes.

The Special Case of Sudden Onsets

Are there any conditions in which stimulus-based factors have priority over cognitive factors in controlling fixation location during scene viewing? We know of only one definitive case: The top-down direction of the eyes can be disrupted by the abrupt appearance of a new but task-irrelevant object, a phenomenon called oculomotor capture (Irwin, Colcombe, Kramer, & Hahn, 2000; Theeuwes, Kramer, Hahn, & Irwin, 1998). We have recently found that during real-world scene viewing, the transient motion signal that accompanies an abruptly appearing new object attracts attention and gaze quickly and reliably—up to 60% of fixations immediately following the onset are located on the new object (Brockmole & Henderson, 2005, in press). This effect is just as strong when the new object is unexpected as when the observer's goal is to search for and identify new objects, suggesting that the allocation of attention to transient onsets is automatic. Thus, visually salient scene regions marked by low-level transient motion signals introduced by sudden changes to a scene can influence gaze in a

manner divorced from cognitive control.

How should stimulus-based and knowledge-based information be combined?

The fact that gaze control draws on stored knowledge implies that image properties about potential fixation targets must somehow be combined with top-down constraints. How is this accomplished? One approach is to construct the initial stimulus-based saliency map taking relevant knowledge (e.g., visual properties of a search target) into account from the outset (Rao, Zelinsky, Hayhoe, & Ballard, 2002). A second approach is to combine independently computed stimulus-based and knowledge-based saliency maps so that only salient locations within knowledge-relevant regions are considered for fixation. For example, Oliva, Torralba, Castelhana, & Henderson (2003) filtered an image-based saliency map using a separate knowledge-based map of scene regions likely to contain a specific target. A yet more radical suggestion would be to move away from the concept of an image-based saliency map altogether, and to place primary emphasis on knowledge based control. For example, in what we might call a *Full Cognitive Control* model, objects would be selected for fixation primarily based on the types of knowledge discussed in the Introduction, such as episodic and schema-based scene knowledge. The visual image in this type of model would still need to be parsed to provide potential saccade targets, but unlike the assumption of the salience hypothesis, the objects and regions would not be ranked according to their inherent visual saliency, but rather would be ranked based on criteria generated from the cognitive knowledge base. For example, if I am searching for the time, and I know I have a clock on the wall in my office, I would rank non-uniform regions in the known location on the wall highly as a saccade target. In this view, saliency itself would play no direct role in saccade location selection. Image properties would only be directly relevant to the extent that they were able to support the sorts of processes need to segregate potential targets from background. (And of course they would be necessary as input for determining that I'm in my office, how I'm

oriented in my office, where the wall is, and so on.) We call this idea that inherent visual saliency plays no role the *Flat Landscape Assumption* and contrast it with the differentially peaked saliency landscapes assumed by the saliency hypothesis.

Conclusion

What drives eye movements through real-world scenes during active viewing task?

Despite the recent popularity of the visual saliency hypothesis as an explanation of gaze control, the evidence supporting it is relatively weak. Cognitive factors are a critical and likely dominant determinant of fixation locations in the active viewing of scenes.

References

- Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the “blank screen paradigm”. *Cognition*, *93*, B79-B87.
- Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*, 62-70.
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene Perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143-177.
- Brockmole, J. R., & Henderson, J. M. (2005). Prioritization of new objects in real-world scenes: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 857-868.
- Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, *13*, 99-108.
- Brockmole, J. R., & Henderson, J. M. (in press). Object appearance, disappearance, and attention prioritization in real-world scenes. *Psychonomic Bulletin & Review*.
- Brockmole, J. R., & Henderson, J. M. (submitted). Contextual cueing in real-world scenes: Eye movement evidence that gist initially guides attention to targets during search.
- Buswell, G. T. (1935). *How People Look at Pictures*. Chicago: University of Chicago Press.
- Castelhano, M. S., & Henderson, J. M. (2003). *Flashing scenes and moving windows: An effect of initial scene gist on eye movements*. Presented at the Annual Meeting of the Vision Sciences Society, Sarasota, Florida.
- Castelhano, M. S., & Henderson, J. M., (2005). Incidental visual memory for objects in scenes. *Visual Cognition*, *12*, 1017-1040.
- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28-71.

Einhauser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, 17, 1089-1097.

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson, and F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108, 316-355.

Gareze & Findlay (this section)

Hayhoe, M. M., & Ballard, D. H. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9, 188-194.

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

Henderson, J. M., and Ferreira, F. (2004a). Scene perception for psycholinguists. In J. M. Henderson, and F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Henderson, J. M., and Ferreira, F. (2004b) (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing: An overview. In G Underwood (Ed.), *Eye guidance while reading and while watching dynamic scenes*. (pp. 269-293). Oxford: Elsevier.

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

Henderson, J. M., & Hollingworth, A. (2003). Eye movements and visual memory: Detecting changes to saccade targets in scenes. *Perception & Psychophysics*, 65, 58-71.

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

Hooge, Vlaskamp, & Over (this section)

Horowitz, T. S. & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, 394, 575-577.

Irwin, 2004. Fixation Location and Fixation Duration as Indices of Cognitive Processing. In J. M. Henderson, & F. Ferreira (Eds). *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

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

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews: Neuroscience*, 2, 194-203.

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

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, 13, 201-214.

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

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, 369, 742-744.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physice – Doklady*, 10, 707-710.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4, 6-14.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 565-572.

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

Mandler, J. M., & Johnson, N. S. (1977). Some of the thousand words a picture is worth. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 529-540.

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision*, 9, 363-386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10, 165-188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997a). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision*, 11, 157-178.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997b). Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26, 1059-1072.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81, 899-917.

McCarley, J. S., Wang, R. F., Kramer, A. F., Irwin, D. E., & Peterson, M. S. (2003). How much memory does oculomotor search have? *Psychological Science*, 14, 422-426.

Oliva A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176-210.

Oliva, A., & Torralba, A., Castelano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. *IEEE Proceedings of the International Conference on Image Processing*.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 6*, 125-154.

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

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

Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception & Psychophysics, 63*, 1239-1249.

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. *Psychological Science, 12*, 287-292.

Posner, M. I., & Cohen Y. (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds). *Attention and Performance X*. Hillsdale: Erlbaum.

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

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

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39*, 3157-3163.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computer and Neural Systems, 10*, 1-10.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76*, 269-295.

Sankhoff, D. and J. B. Kruskal (Eds.) (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-

Wesley Publishing Company, Inc.

Schyns, P., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*, 195-200.

Sereno, S. C., & Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends in Cognitive Sciences*, *7*, 489-493.

Shen & Reingold (this section)

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632-634.

Thiele, A., Henning, M., Buischik, K., Hoffman, P. (2002). Neural mechanisms of saccadic suppression. *Science*, *295*, 2460-2462.

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

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

Underwood (this section)

Volkman, F. C. (1986). Human visual suppression. *Vision Research*, *26*, 1401-1416.

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

Acknowledgements

This research was supported by the National Science Foundation (BCS-0094433 and IGERT ECS-9874541) and the Army Research Office (W911NF-04-1-0078) awarded to John M. Henderson. Affiliations of co-authors are now as follows: James Brockmole, University of Edinburgh; Monica Castelhana, University of Massachusetts at Amherst; Michael Mack, Vanderbilt University. We thank John Findlay and the volume editors for comments on an earlier draft of this chapter. Please address correspondence to John M. Henderson, 285B Psychology Building, Michigan State University, East Lansing, MI 48824-1116, or to john@eyelab.msu.edu.

Figure Captions

Figure 1. Top left: Original scene. Top middle: Model-determined salient regions in the scene. Top right: Fixation locations from all participants. Bottom: Scene with salient regions and participant fixations overlaid. Red dots show participant fixations within a salient region. Red tails mark saccade paths that originated in a non-salient region. Green dots denote participant fixations outside of the salient regions.

Figure 2. Spatial distribution of all participant-generated fixations across all scenes. The figure shows an overall bias for fixations to be placed along a lower central horizontal band.

Figure 3. Top: Mean proportion (with standard error) of all participant-generated saccades that moved from outside to inside salient regions, compared with those generated by a pure random model and a biased random model. Bottom: Mean proportion (with standard error) of all participant-generated fixations that fell within salient regions, compared with those generated by a pure random model and a biased random model.

Figure 4. Similarity of participant fixation locations to model-generated locations (left bars) and to each other (right bars) for the Mannan et al. Index of Similarity (Top Panel), our Unique-Assignment “Warping” similarity index (Middle Panel), and the Levenshtein Distance metric for sequence similarity (Bottom Panel).

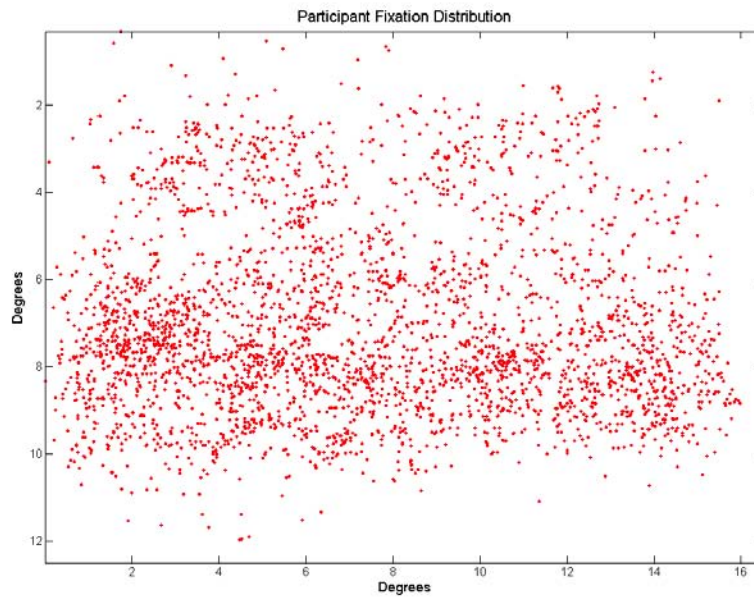
Figure 5. Representative patches from the subject-ensembles (determined by participant fixation locations) and for the random- and shuffled-ensembles used as control conditions. Originals were presented in color.

Figure 6. Mean intensity, contrast and edge density (with standard errors) for the subject-, random-, and shuffled-ensembles.

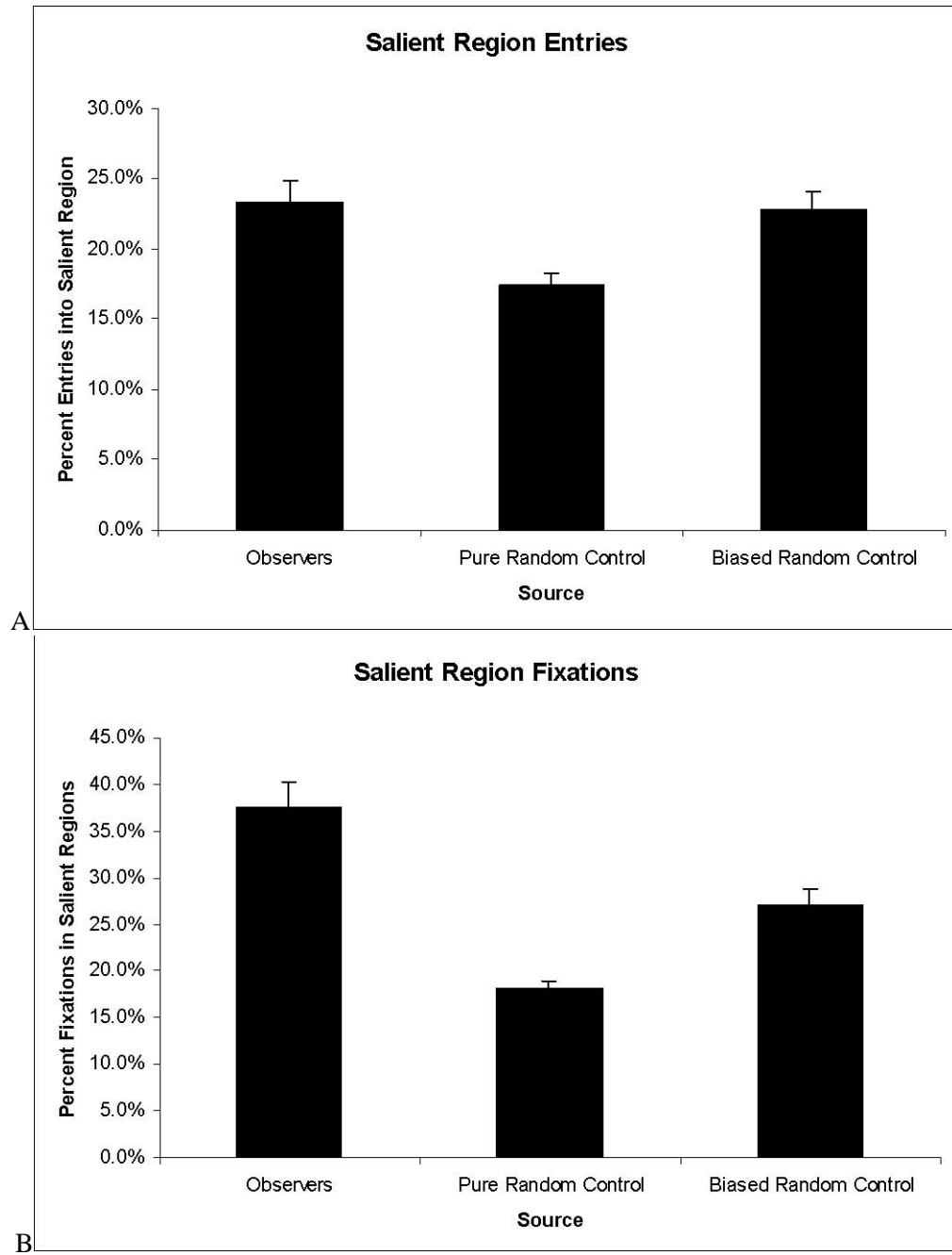
Figure 7. Mean semantic informativeness ratings (with standard errors) for the subject-, random-, and shuffled-ensembles.



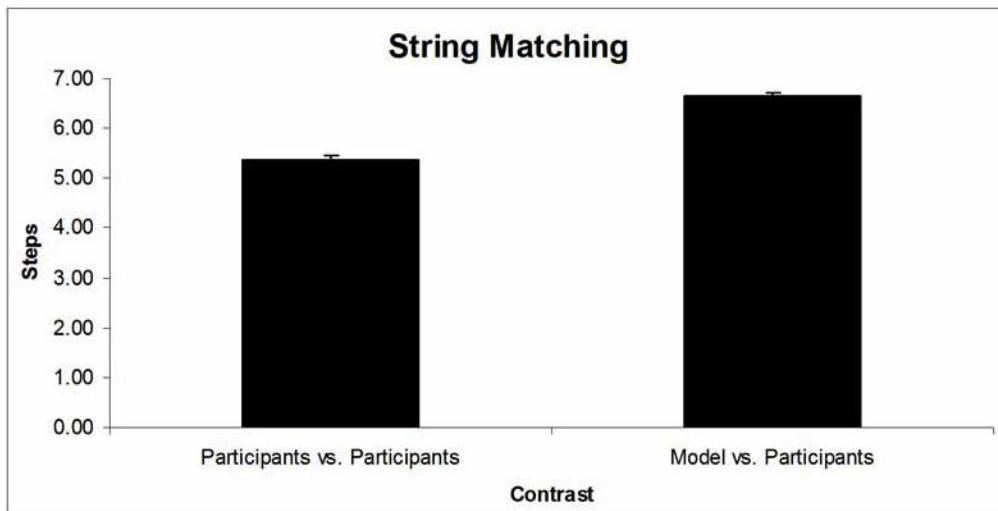
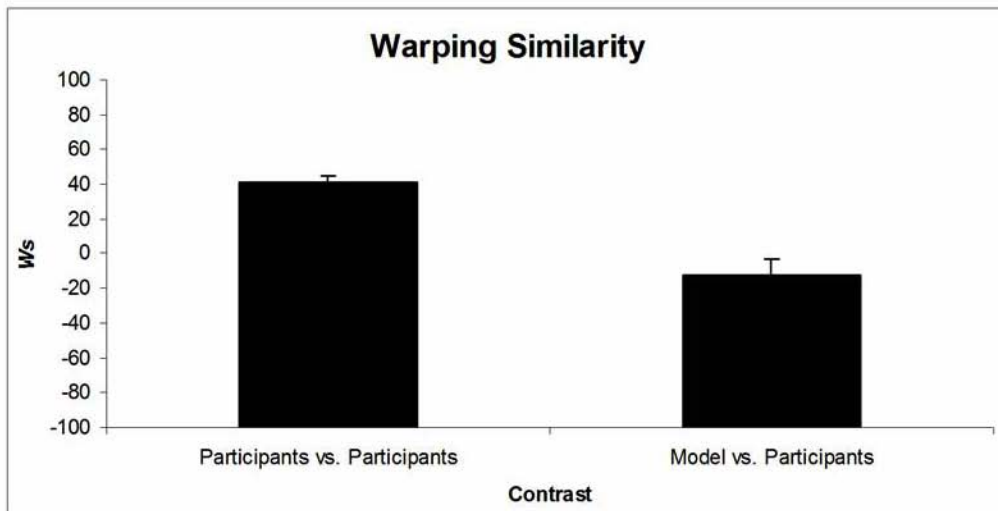
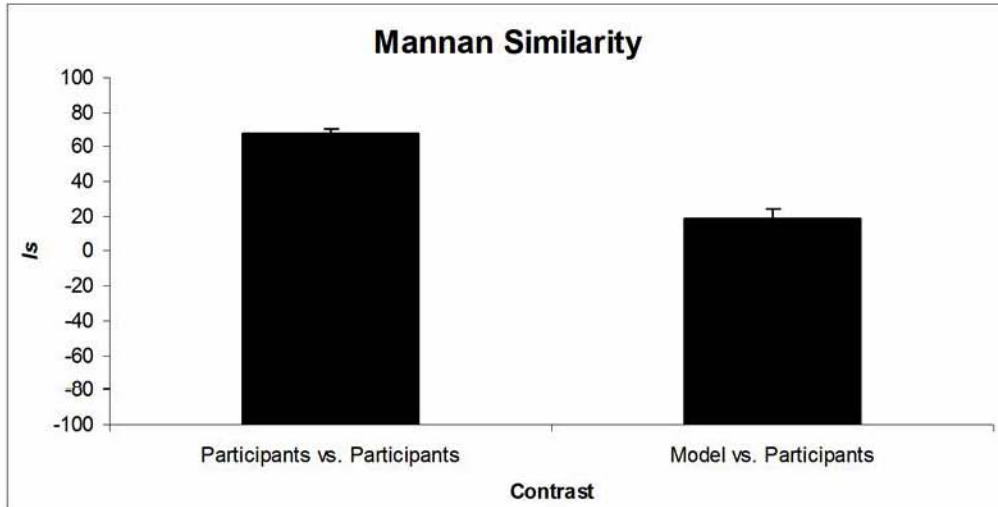
(1)



(2)



(3)



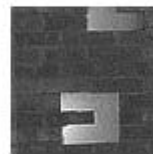
(4)

Patch Exemplars

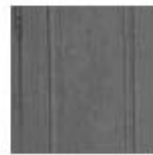
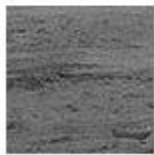
Subject Ensemble



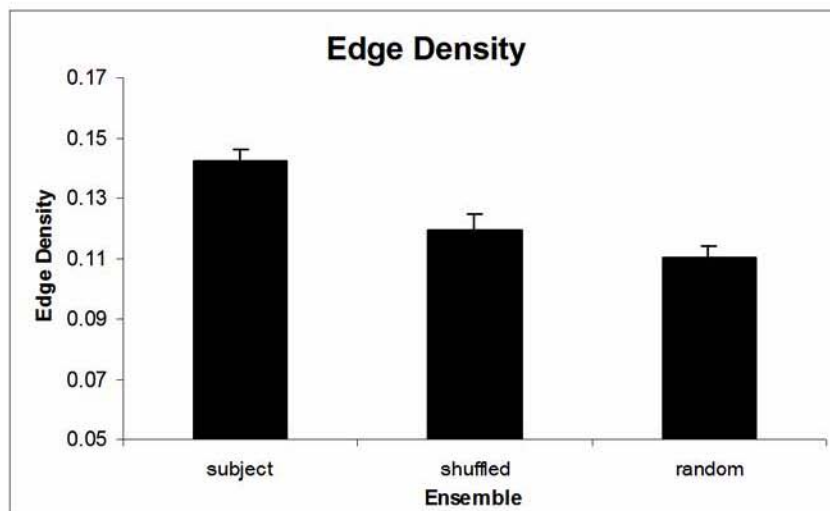
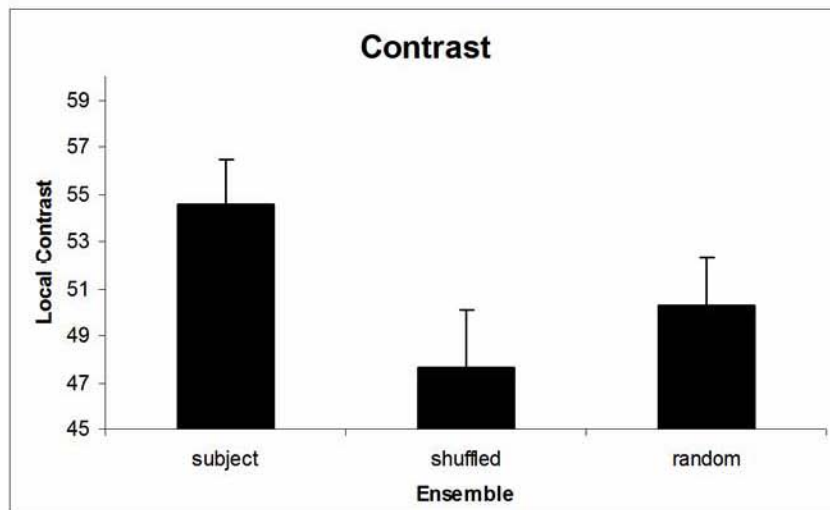
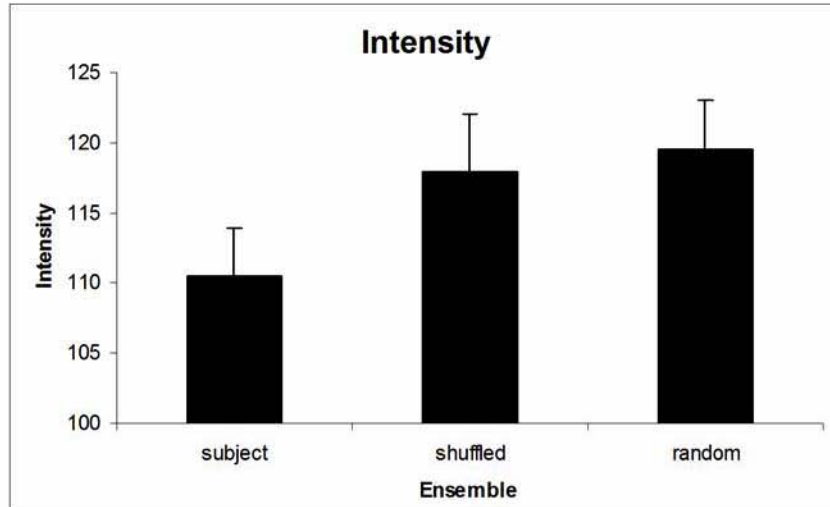
Shuffle Ensemble



Random Ensemble



(5)



(6)

