Interesting objects are visually salient

Lior Elazary

Department of Computer Science,
University of Southern California,
Los Angeles, CA, USA

Laurent Itti

Department of Computer Science,
and Neuroscience Graduate Program,
University of Southern California,
Los Angeles, CA, USA

How do we decide which objects in a visual scene are more interesting? While intuition may point toward high-level object recognition and cognitive processes, here we investigate the contributions of a much simpler process, low-level visual saliency. We used the *LabelMe* database (24,863 photographs with 74,454 manually outlined objects) to evaluate how often interesting objects were among the few most salient locations predicted by a computational model of bottom-up attention. In 43% of all images the model's predicted most salient location falls within a labeled region (chance 21%). Furthermore, in 76% of the images (chance 43%), one or more of the top three salient locations fell on an outlined object, with performance leveling off after six predicted locations. The bottom-up attention model has neither notion of object nor notion of semantic relevance. Hence, our results indicate that selecting interesting objects in a scene is largely constrained by low-level visual properties rather than solely determined by higher cognitive processes.

Keywords: attention, awareness, sensory integration, objects, scene understanding

Citation: 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.

Introduction

Being able to identify interesting regions or objects in our cluttered visual environment is key to animal survival, be it to locate possible prey, mates, predators, navigation landmarks, tools, or food. Yet, very little is known of the computational neural mechanisms that underlie the behavioral selection of interesting objects in our visual world. Is it that we first have to attend to and select a number of candidate visual locations, then recognize the identity as well as a number of properties of each candidate, and finally evaluate these against current behavioral goals, intentions, and preferences, so as to decide whether that object was interesting or not (Navalpakkam & Itti, 2005; Rensink, 2000)? Here we show that the first phase of such seemingly complicated and time-consuming putative process—attentional selection based on intrinsic visual saliency—already is a strong predictor of which regions in digital photographs were labeled by human observers as potentially interesting objects.

Focal visual attention has long been known to be a necessary first step in locating potentially interesting elements in a scene (Itti & Koch, 2001; James, 1890). Indeed, unless attention is first directed toward a particular scene element, much of its attributes and even possibly its very existence will remain unnoticed, as has been vividly demonstrated by studies of change blindness and inattentional blindness (Mack & Rock, 1998; O'Regan, Rensink, & Clark, 1999). Hence, we hypothesized that regions or objects which human observers would find more interesting should also attract attention, that is, be visually salient (Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985). In this paper, we define *interesting* objects or image regions as those which, among all items present in a digital photograph, people choose to label when given a fairly unconstrained image annotation task (details below). The assumption that people would choose to label interesting objects comes simply from the fact that there is some motivation for people to label one region (whether being an object or not) over another.

Early work interested in characterizing what may attract attention toward potentially interesting objects in scenes has suggested that changes in illumination on the retina is a particularly effective cue (Franconeri, Hollingworth, & Simons, 2005; Jonides & Yantis, 1988; Yantis & Jonides, 1996). Indeed, abrupt luminance changes are typically observed when a new object appears in the scene; hence, detecting such low-level physical changes using luminance-tuned visual neurons would often quite effectively guide attention toward interesting novel objects (Kahneman, Treisman, & Gibbs, 1992). Other research suggests that sudden changes in color are also effective in attracting attention (Snowden, 2002; Turatto & Galfano, 2001), although this has been more largely debated (Folk & Annett, 1994; Franconeri & Simons, 2003; Jonides &

doi: 10.1167/8.3.3 Received April 20, 2007; published March 7, 2008 ISSN 1534-7362 © ARVO

Yantis, 1988; Theeuwes, 1995). Other influences also are tied to the behavioral task which an observer may be engaged in, for example, a search task (Quinlan & Humphreys, 1987; Treisman & Sato, 1990; Wolfe, Cave, & Franzel, 1989). Indeed, the effectiveness of simple bottom-up information, like color and illumination, in attracting attention can be modulated by task influences to yield complex search patterns (Desimone & Duncan, 1995; Evans & Treisman, 2005; Theeuwes, 1994; Treisman & Sato, 1990; Underwood & Foulsham, 2006; Wolfe, 1994). However, the relative strength of contributions from bottom-up information (e.g., salience) versus top-down information (e.g., relevance to a task) in determining what people find interesting remains largely unknown (Henderson, 2003). Possibly, when no specific search target, no search task, and no particular time or other constraint are specified to an observer, bottom-up information might play a predominant role in guiding attention toward potential generically interesting targets (Itti, 2005). Under such conditions (e.g., under free viewing), bottom-up information could provide a strong indication of what people might find interesting in a given scene.

We used a computational model to compute saliency maps in digital photographs and to test the extent to which saliency at a given image location indicates how interesting that location may be to human observers. Previous human eye-tracking studies have shown that saliency is a strong predictor of attention and gaze allocation during free viewing, both in static images (Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005; Underwood & Foulsham, 2006) and in natural video stimuli (Itti, 2005). However, it has not been shown, under natural viewing conditions, whether a visual location that is attracting the gaze is also being judged as interesting. Intuitively, it is possible that attentional selection based on low-level visual saliency may not be a good indicator of which scene elements are in the end judged subjectively interesting, as saliency may yield too many false positives. That is, observers may be attracted to salient locations and examine them, but may end up discarding an overwhelming majority of them as uninteresting, relying instead on different mechanisms to isolate more subjectively interesting locations. Testing the latter hypothesis requires ground-truth data of which locations in images may be more interesting to human observers.

Russell, Torralba, Murphy, and Freeman (2005) created a research tool called *LabelMe* for people to annotate objects in scenes (Figure 1). The scenes are submitted by various contributors and depict many indoor and outdoor locations. For instance, images include outdoor scenes of cities around the world, office buildings, parking lots, indoor offices and houses, country scenes, and many more. Examples of scenes and associated object outlines from the database are shown in Figure 1 and in (Russell et al., 2005). As can be seen, the labeled objects in the scene range from being in plain view and well lit to being partly

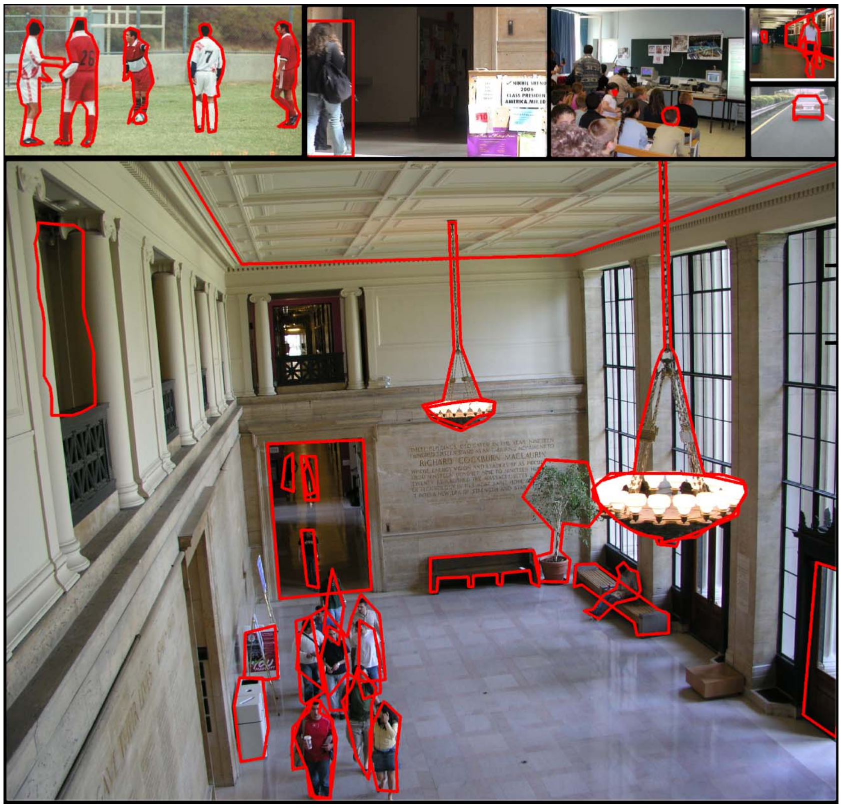

Figure 1. Example of scenes and associated objects outlines.

occluded and low contrast or distorted. Anyone can contribute new images to this shared repository. One can also manually trace the outlines of scene elements using simple computer-based tracing tools. Finally, one can associate semantic labels with the outlined elements (e.g., a car, a dog). This database and associated tracing and labeling tools are freely available on the World Wide Web. The original purpose is to collect ground-truth data about objects in complex natural scenes and to train computational object recognition algorithms. Note that the only given criteria for labeling are to make "nice labels" (i.e., outlining whole objects somewhat precisely, as opposed to just drawing a rectangular bounding box), and contributors certainly are not instructed to find or to label objects which are more salient. However, It can be argued that a selection bias might exists to submit and to label objects that are inherently salient compared with the domain of all possible objects—whether such a saliency-driven bias indeed exists is the main research question posed by the present study. That is, by looking at a large number of scenes, we attempt to find this bias and whether it is contributed by top-down or bottom-up processes. In addition, the images were viewed on various computers with various display sizes and color properties. As a result, this data set provides a good indication of what people would find "generically interesting," in the absence of a particular task, in uncontrolled conditions, under no time pressure, and outside the laboratory; obviously, the flip side to this is that the data set is highly heterogeneous and possibly very noisy, which might mask any of the effects being investigated here. At the time of this writing, there are 74,454 objects annotated by several human annotators in 24,863 scenes. In this paper, we propose to use this massive collective data set as an approximation to deciphering which scene elements may be more generically interesting to a human observer: Presumably, the scene elements which were outlined by the annotators were more interesting than the many other scene elements which were not outlined. The high heterogeneity and inherently uncontrolled nature of this data set regarding image types, resolution, annotators, image display conditions, tracing tools used, etc., is such that any bias that might be robust enough to be detected in the present study could reasonably be expected to generalize to other data sets.

Our main contribution is to test, using a very large scale data set, whether scene elements that human observers find more interesting and choose to outline in natural scenes have distinguishing low-level visual properties that differentiate them from other scene elements. Since we define interesting objects as objects that people choose to label, we hypothesize that one such distinguishing visual property is low-level visual saliency, as approximated by saliency map models (Itti et al., 1998). The assumption that people would choose to label interesting objects comes simply from the fact that there is some motivation for people to label one region (whether being an object or not) over another. A region might be labeled because it is easy to do so because it is very unique in the scene (in terms of bottom-up, top-down, or both), or the annotator randomly chose this region for reasons unknown even to him or her. Therefore, in this study we attempt to determine whether which image regions end up being labeled can be predicted using a model of bottom-up saliency, so as to test whether some of the motivation to label particular regions was due to bottom-up influences. In the following sections, we describe the experimental protocol and large-scale simulations and data analysis. We summarize our findings in several ways, including counting how many outlined objects are selected by the model as attention scans the saliency map in order of decreasing saliency values. Our findings have important implications for the understanding of human vision as well as the design of machine vision system. In particular, they indicate that humans observers are neither entirely "free" nor entirely unpredictable in their choice of which region to label, as we show that they choose salient regions above all else in a statistically highly significant manner. We conclude that a simple low-level saliency map may serve as quite a strong approximation of what humans find interesting in visual scenes.

Methods

Subjects/participants

Not much is known about the participants except that the data set contained labels by seventy eight distinct individuals that account for 7,330 of the labels (30%), while the rest of the labels are made by anonymous users. It is difficult to precisely know the motivation of the people who labeled objects in this data set. However, from the nature of the *LabelMe* project, it could be inferred that most images are labeled by computer science researchers interested in vision. As a result, it could be further inferred that the data set would contain objects and scenes that would be difficult for various vision algorithms to recognize.

Apparatus

The *LabelMe* database at the time of testing consisted a total of 24,863 labeled scenes, of which 7,719 were single-shot static scenes and 17,144 where images from sequence (video) scenes. Image size varied from 130 × 120 to 4,423 × 3,504 pixels for the static scenes (M = 1,549.90, SD = 1,321.48 × M = 898.23, SD = 680.06) and 160 × 120 pixels to 1,024 × 768 (M = 681.28, SD = 156.90 × M = 458.72, SD = 106.09) for the sequence scenes. Within all these scenes, 74,454 objects were labeled, from 1 to 87 per image (M = 8.90, SD = 11.30).

Each labeled object occupied between <0.01% and 99.82% of the total image area (M = 7.32, SD = 9.95), and the union of all labeled objects in each image occupied between <0.01% and 99.99% of the total image area (M = 20.82, SD = 0.12). On average, about 20% of each image was labeled. Sometimes object outlines overlapped within an image (e.g., a desk was outlined, and also a computer on that desk), which slightly complicated our analysis (see below). Lastly, the order in which people chose to label the objects is captured in the data set. This is later used as a measure of how salient the first labeled objects were. Examples of scenes and associated object outlines from the database are shown in Figure 1.

Design

To determine whether human observers labeled salient objects in the *LabelMe* data set, a saliency map was computed according to the algorithm proposed by Itti et al. (Itti & Koch, 2000; Itti et al., 1998). The saliency map was then inspected algorithmically in several ways to reveal whether or not it had preferentially highlighted the labeled objects over other objects. The results were then compared to chance to indicate how difficult was the task of selecting labeled objects in the data set.

The saliency map algorithm used in this study is inspired from biological systems, where an image is decomposed into several different feature maps at several spatial scales. In the experiments performed below, the feature map domains consisted of intensity, color opponency (red–green, blue–yellow), and four orientations (0, 45, 90, 135). Within each feature map, a spatial competition weighs the values of detectors in a data-driven fashion based on their uniqueness in that map. That is, the more different the response of a given local detector is from its neighbors, the higher the weight assigned to that detector's output. The 42 feature maps (seven features at six spatial scales) are then combined into a saliency map, which indicates the saliency of each location in the image. A winner-take-all neural network is then used to select the location of highest saliency value and to set a region of interest toward that location. Once the location has been selected and attended to, an inhibition of return (IOR) mechanism is used to inhibit that location in the saliency map, such that the region of interest (ROI) will then shift to the next most salient location. Implementation details of this model have been described previously (Itti & Koch, 2000; Itti et al., 1998) and the algorithm is freely distributed in source code at http://iLab.usc.edu/toolkit/.

Procedure

Each measure reported below was computed separately for static scenes and for sequence scenes as well as for both. In our analysis, we treated all images, including sequences, as independent frames (no motion cues). As a result, a bias toward similar frames in sequences can develop. That is, if the saliency map found a labeled region in one frame, it would presumably find that same region in the next frame (due to the similarities between frames). On the other hand, computing chance between frames would find different regions. Therefore, by examining the video sequences separately from the static images, any biases toward similar frames can be found.

Hit analysis

Once the computed saliency map was available, the location with the strongest activation was inspected to determine whether it fell inside a human-labeled object or not. Note that a hit was considered only when the center of the predicted location was inside a labeled object and not a region. This measure gave the hit rate, which indicated the saliency map's ability to locate a labeled object within the first predicted location. The hit rate was then compared to chance, which indicated the probability that a labeled object would be chosen given a random location. That is, if the scenes were fully labeled, then any point chosen at random would be a hit, but if the scenes were only partially labeled (which is the case for most scenes in the *LabelMe* data set), then choosing a point at random would have a low probability of hit. If one assumes a uniform probability of picking a random location anywhere in the image, chance probability of a hit is hence given by the ratio of the area of the union of all labeled regions to the area of the entire image. It is this ratio which we computed for every image and which we report. As a control, we confirmed that identical values (to within 0.1%) were obtained by picking 100 uniformly distributed random locations in each scene and testing at each location whether or not it belonged to a labeled object. In addition, a random map was created to obtain the chance results for the multiple location experiments. This process is explained later in the paper.

Hit analysis with knowledge

It can be argued that a selection bias may exist to label objects that are near the center of the image (the so called center-bias; Tatler et al., 2005). To account for this knowledge, a "bias image" of size 512 × 512 pixels was computed as the sum of all the filled labeled objects across all scenes (after rescaling every image to 512 × 512). The result of this image can be seen in Figure 2. As can be seen, a bias does exist for people to label objects that are more on the left/center side of the image. This knowledge was then used to bias both the random map and the saliency map. A simple technique to draw random

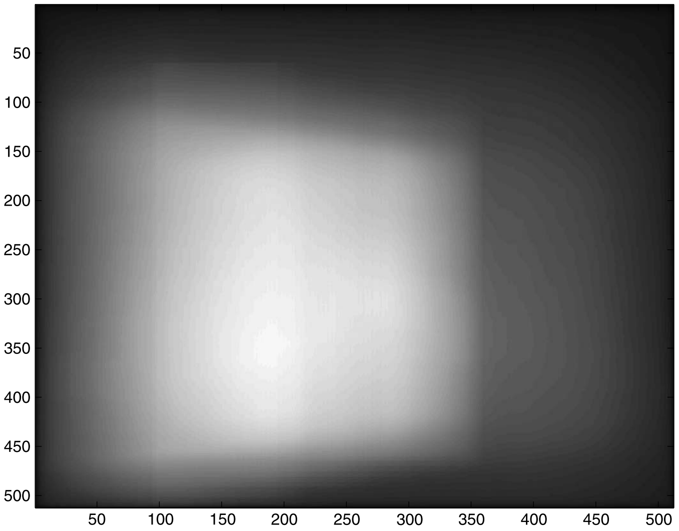

Figure 2. Bias image obtained by summing all pixels belonging to labeled objects across all scenes.

locations according to the distribution given by the bias image is as follows: After normalizing the computed bias image map to real values between 0 and 1, uniformly random samples were drawn in a 3-dimensional space [0..512[× [0..512[× [0.0..1.0[; a sample was then kept if its value in the third dimension was below the value given by the computed normalized bias image map at the location of the sample. To apply the bias image map to the saliency map, the result of the saliency map was simply weighted by this bias image map. The hit rate from these maps was computed separately and labeled as hit rate with knowledge.

Label order hit analysis

Given the order of labeling, the number of hits for a particular order was determined. That is, how often the first salient location fell within the first outlined object, second outlined object, etc. This gave a measure of whether the model can also predict what people would choose to label first. The underlying hypothesis is to test whether people would choose to label the most salient object first, then any other object.

Saliency analysis

To indicate the ability of the saliency map to determine labeled objects, several measures were evaluated. The ratio of the maximum saliency value within labeled objects (target) versus the maximum saliency value within unlabeled regions (background) was examined. Note that this ratio is related to the hit rate computed above. If this ratio is above 1.0 on a given image, then shifting the region of interest to the maximum saliency value would result in a hit. In addition, a similar ratio was examined, of average saliency within labeled versus background regions. The average saliency value of a labeled area and unlabeled area was computed by taking the ratio between the sum of saliency values within the region and the number of pixels in that region. Note that comparing the results of these analyses against chance is not fruitful. This is due to the fact that the random map does not contain regions. As a result, comparing the ratios between the maximum and the minimum (or averages) of the values in the labeled objects and background regions yields a value very close to one. Therefore, plotting these values on the same graph would only show a line at position one and would not be useful for retrieving any information.

Number of labeled objects found analysis

Results were also obtained to evaluate how well the next most salient location indicated a labeled object, after an inhibition of return was applied to the current most salient object. The inhibition of return was simply a Gaussian disk mask, which set all salient values underneath the mask toward zero, so that the next maximum salient location would have to be outside the disk. Note that the disk does not tile the image. This is due to the fact that the saliency map does not highlight every region in the image (some areas have zero saliency) due to some of the global competition for salience operations. As a result, these locations would never be found by the model. In addition, note that the IOR disks can also overlap. This is due to the fact that even though the region under the disk was inhibited, the value on the parameter of the disk might still be the next most salient location. As a result, choosing that location and using an IOR would result in half of the old IOR to be inhibited again, which would cause the disks to overlap.

To account for various sizes of objects, two disk sizes were examined, one with a radius of 1/4 the image width and the other with a radius of 1/16 the image width. This corresponds roughly to the mean object size (1/16) and 2 standard deviations from the mean (1/4). The process was then to compute the saliency map, attend to the location of maximum salience and determine whether it was inside a labeled object or not, apply the IOR disk of the particular radius at that location, and choose the next most salient location to issue the next shift in the region of interest. This process was repeated until 100 shifts of ROI were made, or when the saliency map had an extremely low maximum value (0.01 when the saliency map was normalized to values between 0.00 and 10.00), which was found to be the point at which subsequent shifts of ROI would not produce new attended locations. The number of shifts of ROI until a labeled object was first hit was used as a measurement, as well as the

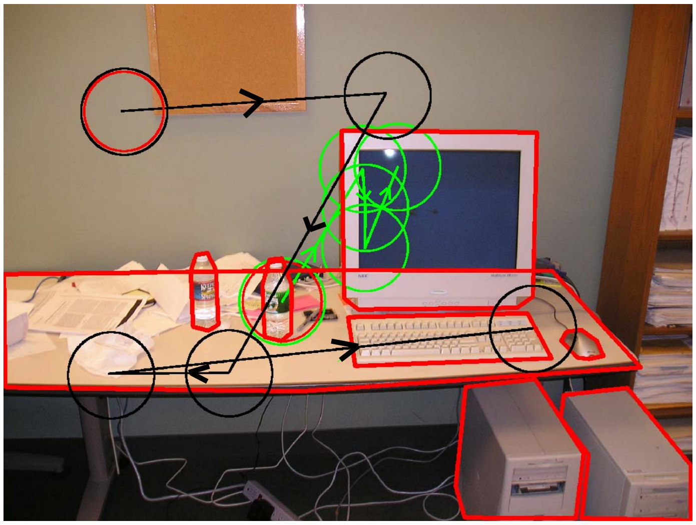

Figure 3. Example of the IOR process. Black circles indicate locations chosen at random, while green indicates locations chosen from the saliency map. In both cases, the inner red circle indicates the first of those chosen locations. Note that the center of one IOR disk is sometimes within the radius of another. This is due to the Gaussian effect where the edges of the IOR are inhibited by a lesser amount.

percentage of labeled objects found versus the number of predicted locations within the whole scene.

To test for chance, the same process outlined above was repeated but with a random map instead of the saliency map. This was achieved by first creating an image with unique values from 0 to image size and shuffling the pixels in a randomized order. The maximum location in this randomized map was then chosen, checked for a hit, and inhibited with an IOR of size 1/4 or 1/16 the image size. This gave the chance results for both the number of objects found and the number of shifts of ROI needed to find the first labeled object. An example of the IOR process can be seen in Figure 3.

Statistical tests

To test the statistical significance of the hit rate results, the binomial test was used to obtain a z score. This test was used due to the binary nature of the data. Therefore, the z score would indicate the probability of a hit in a given trial. Since the number N of images is much grater then 10, the normal approximation to the binomial distribution was used.

$$z = \frac{X - p}{\sqrt{N'\,P'\,(1 - P)}}, \tag{1}$$

where X is the hit rate that the saliency map obtained, p is the hit rate obtained by chance, P is the probability of hitting an object in a given image, and N is the number of images.

The Welch's t test (Welch, 1947) was also used to test the statistical significance when the saliency map was used to detect all of the objects in a given scene. This test is similar in nature to the Student's t test, but it considers the fact that the two sets of data have different variances.

$$t = \frac{X_1 - X_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{2}$$

where X_1 and X_2 are the mean of the saliency map results and chance results, respectively, s_1^2 and s_2^2 are the variance of the saliency map results and chance, respectively, and $N_1 = N_2$ is the number of images, and where the degrees of freedom were calculated as

$$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1 - 1)} + \frac{s_2^4}{N_2^2(N_2 - 1)}}. \tag{3}$$

Results

Hit analysis

Figure 4 shows the percentage of hits that the saliency map was able to predict in the *LabelMe* data set. We

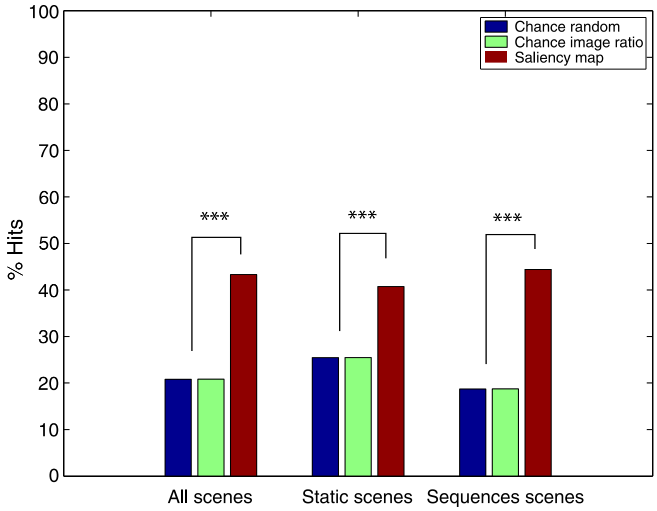

Figure 4. Percentage of hits in the *LabelMe* data set using the saliency map (red), chance based on 100 random points (blue), and chance based on the ratio between the label object area and the image size (green). Error bars on the chance values are too small to be visible. *Note:* ***z score > 30, p = 0.001.

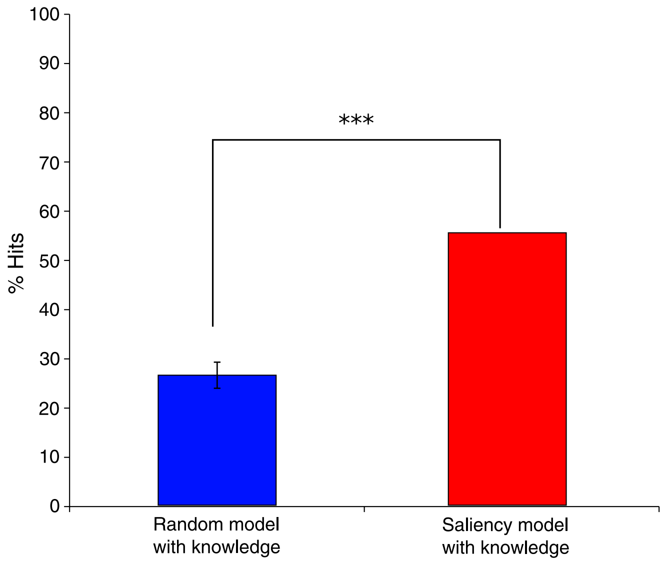

Figure 5. Percentage of hits in the *LabelMe* data set using the saliency map (red) and chance based on 100 random points (blue) with the knowledge of the biased image. *Note:* ****z* score > 30, *p* = 0.001.

found that the saliency map was highly significantly above chance at preferentially attending to labeled objects. Indeed, the hit rate, or percentage of images were the location of maximum saliency and hence the first attended location, was within a labeled object, was 43.29%, 40.07%, and 44.45% for all the scenes, the static scenes, and the sequences, respectively. These values were about twice above chance. The values of chance computed were $M = 20.82$, $SD = 0.08\%$, $M = 25.46$, $SD = 0.07\%$, and $M = 18.73$, $SD = 0.12\%$ for all, static, and sequences, respectively. A binomial test indicated that the hit rates for the saliency map were statistically significantly above chance with $z = 87.27$, $p = 0.001$, $z = 30.74$, $p = 0.001$, and $z = 86.32$, $p = 0.001$, for all, static, and sequences, respectively.

Hit analysis with knowledge

Figure 5 shows the percentage of hits which the saliency map was able to predict in the *LabelMe* data set with the knowledge of the overall object label bias image. We found that the saliency map was again highly significantly above chance at preferentially attending to labeled objects. Indeed, the hit rate, or percentage of images where the location of maximum saliency and hence the first attendant location, was within a labeled object, was 55.33% for all images, while chance was computed as $M = 26.40$, $SD = 2.65\%$. A binomial test indicated that the hit rates for the saliency map were statistically significantly above chance with $z = 105.80$, $p = 0.001$. It is worth noting that even if we compare the

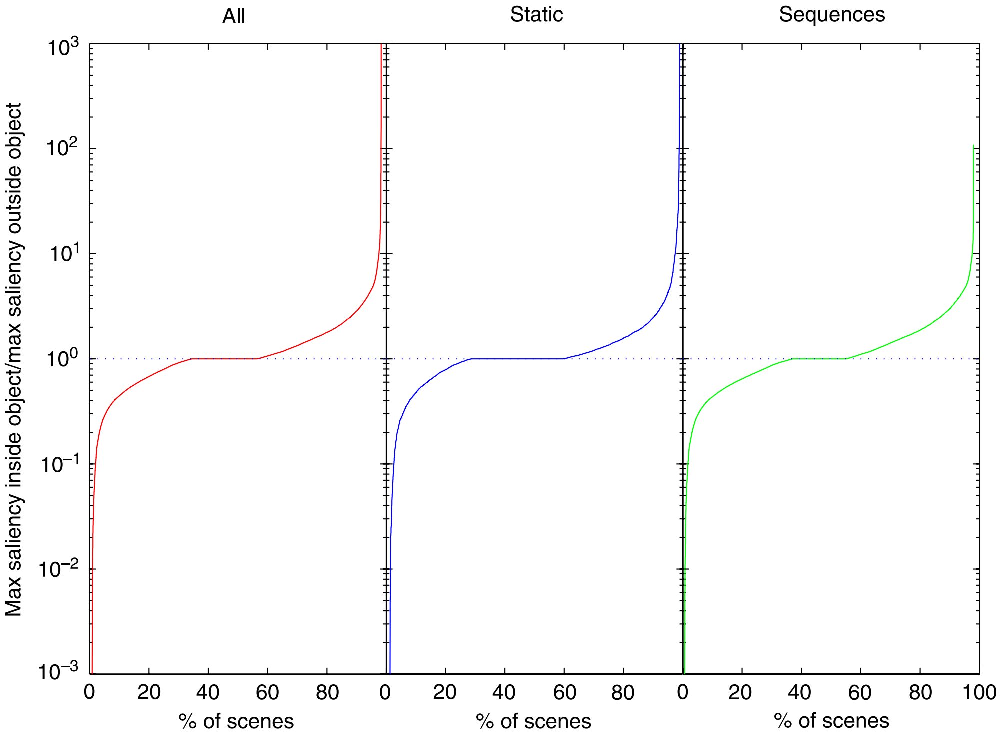

Figure 6. Ratio of the maximum saliency value within labeled regions (human-selected objects) to the maximum saliency value within unlabeled regions (background). This value is plotted for all labeled scenes (left), just the static scenes (center) and the sequence scenes (right). The doted line indicates the value of 1.0. In images with ratios above 1.0, the single most salient location over the entire image falls within a labeled region; this was the case for over 40% of all images.

unbiased version of the saliency map with the biased randomized model, the results are still significantly above chance. In particular, the saliency model still picks a labeled object 43.29% of the time versus 26.40% for biased-random, leading to a significant advantage for saliency (binomial test $z = 60.41$, $p = 0.001$).

In sum, if a system just picked a location at random, it would have an approximately 21% chance of hitting a labeled object; using the saliency map to guide the choice of location to that of maximum salience, however, increased the ability to detect a labeled object to 43%, or about twice the chance level. In addition, the knowledge of where people will often label increases these results to 55% versus a chance level of 26% (Figures 4 and 5).

Label order hit analysis

Figure 8 shows the percentage of the total hits as a function of the order in which the objects were labeled by the annotators. As can be seen, more hits where made to the first-labeled and second-labeled objects than to any other object. In particular, almost 30% of the hits where made to the first object, and almost 10% of the hits were to the second object. This indicates that people did tend to label the most salient object first.

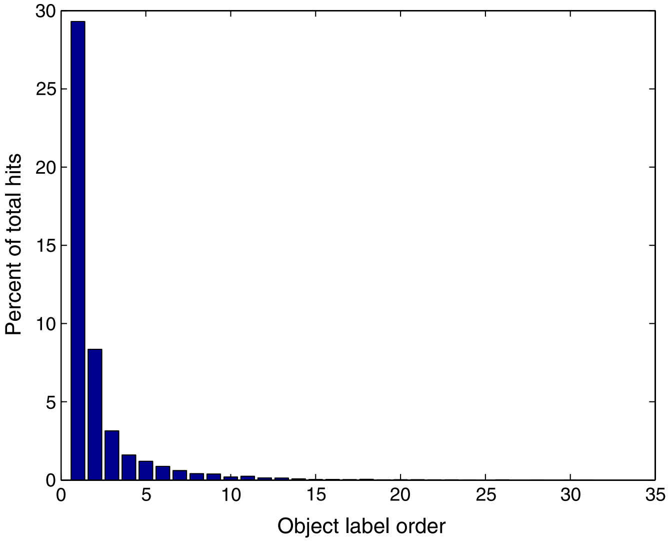

Figure 8. Percentage of hits out of the total hits versus the order in which objects were labeled by the annotators. This graph shows often the first salient location fell within the first-labeled object, second-labeled object, etc.

Saliency analysis

Figures 6 and 7 show the ratios of maximum and average saliency inside versus outside the labeled regions,

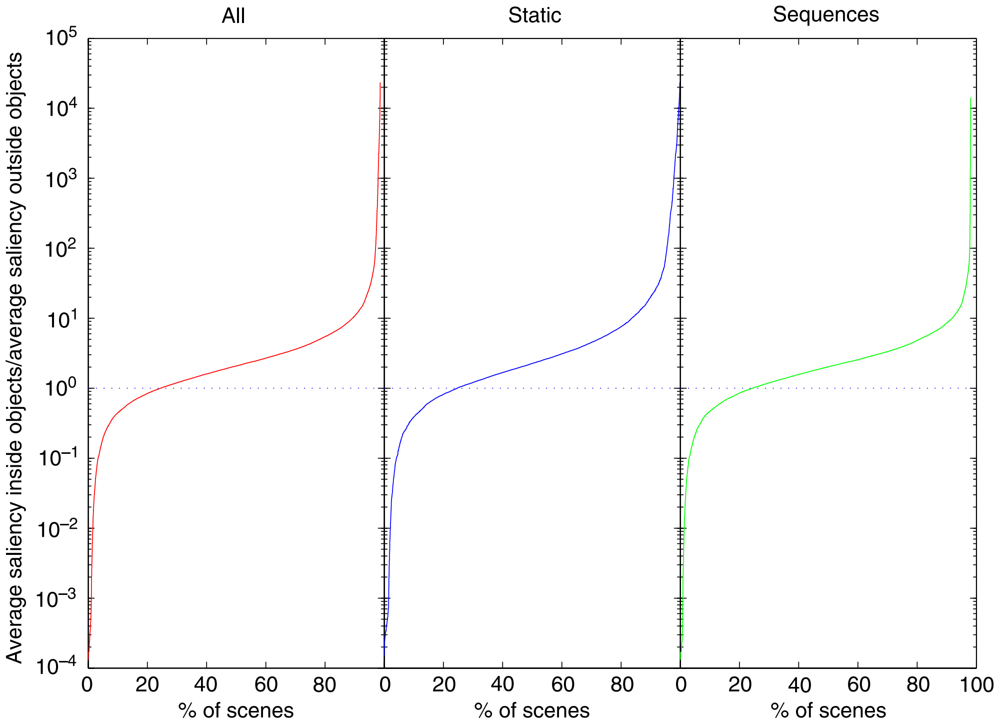

Figure 7. Ratio of the average saliency value within labeled regions to the average saliency value within unlabeled regions for all labeled scenes (left), static scenes (center), and sequence scenes (right). The doted line indicates the value of 1.0. In images with ratios above 1.0, the average salience within labeled objects is higher than the average salience of the background; this was the case for over 76% of all images.

as a function of percentage of scenes. The ratio of maximum inside versus outside saliency (i.e., maximum target saliency to maximum background saliency) was above unity for over 40% of the images, while the ratio of average inside versus outside saliency was above unity in more than 76% of the images (exact numbers vary slightly depending on image subset, see Figures 6 and 7).

In both figures, we observed similar results for static and sequence scenes, which was interesting as the *LabelMe* database contains substantially more sequence images than static images. Hence, our conclusions drawn from the entire data set are generally applicable to static scenes, sequence scenes, and both combined.

It is also interesting to examine the area under the curve when the ratio is above unity in Figure 6, which indicates that attending to the location of maximum saliency would result in a hit. This area is about 40% for static scenes and sequences scene as well as for both and corresponds to the data shown in Figure 4. In addition, Figure 6 shows a flat horizontal region very close to unity for about 20–30% of the images. This flat region indicates that in these images the saliency map was very close to achieving a hit but possibly fell short by one of more pixels from the outline. This shortcoming is mainly due to the IOR process and the way a region is considered a hit (one pixel accuracy). Therefore, if we look at the average saliency around that region and compare that with the outside area (Figure 7), we can see that over 76% of the images corresponded well with the saliency map. Therefore, if a small improvement to the method of choosing a hit is made (looking at a neighborhood of pixels for instance, or at multiple locations), then this would significantly raise the hit rate.

Number of labeled objects found analysis

Figure 9 shows the number of attended locations taken to reach a labeled object. These numbers were obtained for IOR sizes of 1/4 and 1/16 the image width. In a small fraction of the scenes (about 10% with an IOR of 1/16), a labeled object was never found. However, within two shifts of ROI, the saliency map was able to find a labeled region in more than 50% of the scenes and within three shifts in more than 71% of the scenes with an IOR of 1/16 (over 76% with an IOR of 1/4, with chance being 43%). In particular, the percentage of scenes in which one or more labeled objects were hit within three shifts of ROI with an IOR of 1/16 was 71%, 65%, and 73% for all the scenes, static scenes, and sequences, respectively. These values were significantly above chance. The values of chance

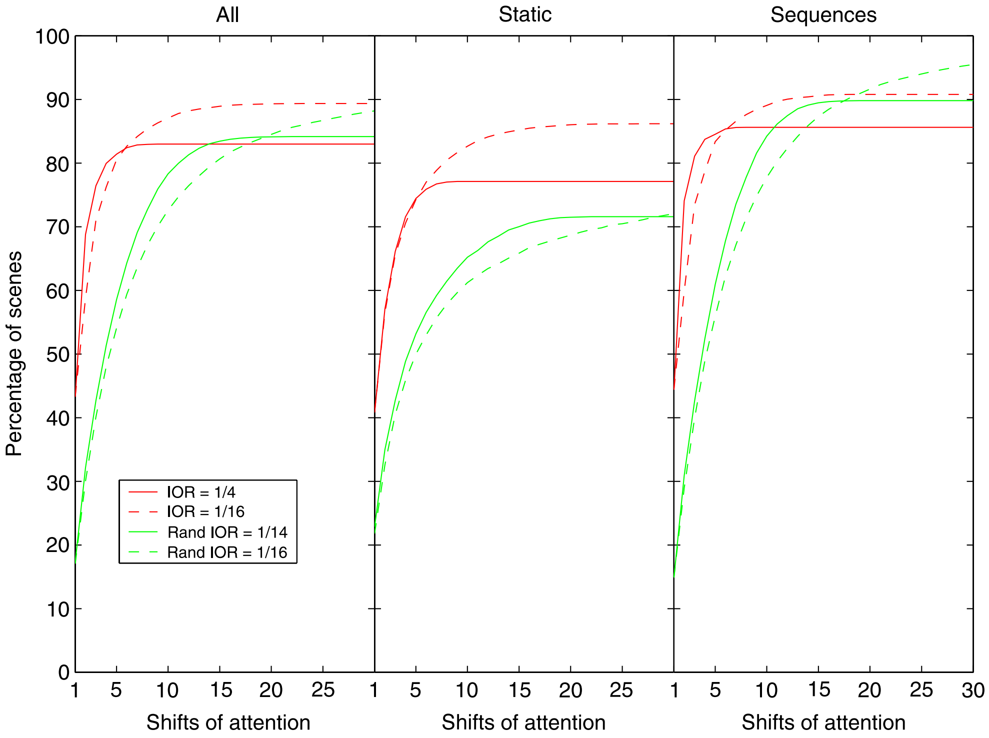

Figure 9. Number of attended locations taken to reach a labeled object versus the number of scenes for saliency-guided attention (red) and random attended locations (green). Only up to 25 predicted locations are shown here for display purposes (our simulations allowed for up to 100 locations). Results for two inhibition-of-return sizes are shown, with radius equal to 1/4 (solid line) the image width and 1/16 (dashed line) the image width. Results are plotted for all labeled scenes (left), just the static scenes (middle), and the sequences (right). Saliency-guided attention hit a labeled object quicker than random attention in most scenes, but the asymptotical gain in the number of hits when executing more shifts of ROI was also shallower (see text for further details).

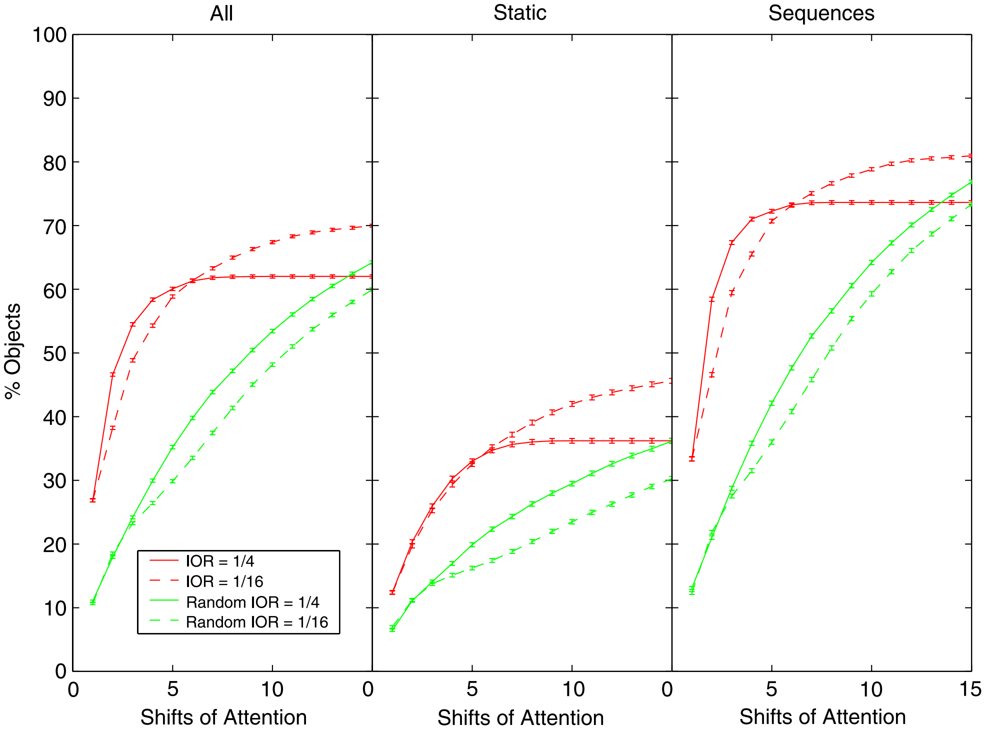

Figure 10. Percentage of objects found in the scenes versus the number of predicted locations (shifts of ROI) taken to find these objects for saliency-guided attention (red) and random attention (green). This was done using a disk-shaped inhibition of return with a radius equal to 1/4 (solid line) the image width or 1/16 (dashed line) the image width. Results are plotted for all the labeled scenes (left), just the static scenes (middle), and the sequences (right). Saliency-guided attention found a greater number of labeled objects quicker than random attention, but the asymptotical gain in the number of objects found when executing more attended locations was also shallower (see text for further details).

computed were 40%, 40%, and 39% for all, static, and sequences, respectively. A binomial test indicated that the hit rates for the saliency map were statistically significantly above chance with $z = 99.44$, $p = 0.001$, $z = 44.88$, $p = 0.001$, and $z = 89.65$, $p = 0.001$ for all, static, and sequences, respectively.

Figure 10 shows the percentages of objects found when the saliency map was allowed to make up to 100 predicted locations. The figure shows that within the first 5 predicted locations (with an IOR of 1/4) the saliency map found more than $M = 60.07\%$, $SD = 0.26\%$ of all labeled objects in all the scenes, more than $M = 72.25\%$, $SD = 0.29\%$ in the sequences scenes, and over $M = 33.01\%$, $SD = 0.38\%$ in the static scenes. The values of chance computed were $M = 35.21\%$, $SD = 0.26\%$, $M = 42.11\%$, $SD = 0.33\%$, and $M = 19.89\%$, $SD = 0.32\%$ for all, sequences, and static scenes, respectively. A Welch's t test indicated that these values are statistically significantly above chance with $t(49{,}723) = 67.99$, $p = 0.001$, $t(33{,}729) = 68.23$, $p = 0.001$, and $t(15{,}000) = 28.18$, $p = 0.001$ for all, static, and sequences, respectively.

Furthermore, within the first 5 predicted locations, the number of found objects was not completely dependent on the size of the IOR, although more objects where found at a faster rate (fewer shifts of ROI) when the IOR had a size of 1/4 the image width. However, after the 5th predicted location, the number of objects found became highly dependent on the IOR size. As a result, an IOR size of 1/16 was able to find more objects, increasing the object detection rate by 1%.

Example scenes

Figure 11 shows a sample of a scene and its saliency map. In this scene, the saliency map was able to correctly find a labeled object within the first predicted location (indicated by a green circle). Furthermore, the saliency map found the labeled fire hydrant, a small portion of the mailbox, and the top window as salient regions as well.

Figure 12 shows an example for when the saliency map resulted in a miss from the first predicted location. However, after triggering the IOR, the saliency map was able to correctly locate the labeled region on the second predicted location. This image is an example of a sequence, where the person was labeled multiple times in subsequent images. Therefore, if the saliency map finds the labeled region in one image, it will most likely find it in subsequent images as well.

Lastly, Figure 13 shows a compete miss due to a large sized IOR. From the first predicted location, the saliency map chose the center (unlabeled) computer as a salient

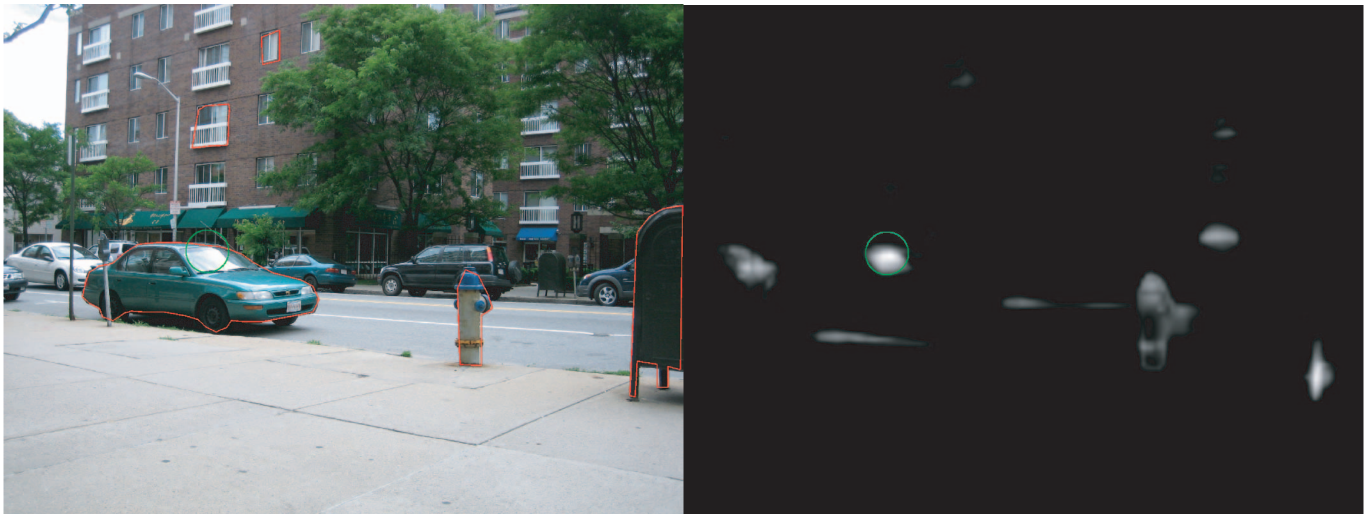

Figure 11. Example of a static scene and its saliency map with a labeled object found in the first saliency-guided shift of ROI (to the maximum saliency value over the entire image). Image size: 2,592 × 1,944 pixels. IOR radius (green circle): 162 pixels (1/16 of the image width).

region (indicated by the green dot). However, after triggering an IOR of 1/4 the size of the image width, most of the labeled regions was inhibited. This meant that the labeled regions would never be found again.

Discussion

In this study, we investigated whether the subjective choice of which objects in complex natural scenes may be judged more interesting, as defined by being worthy of being manually outlined, could be predicted by a simple model of low-level visual saliency. We found that indeed hot-spots in the saliency map tended to predict the locations of interesting objects, significantly above chance.

Although highly above chance level, the 43% hit rate when making a single attention shift to the location of

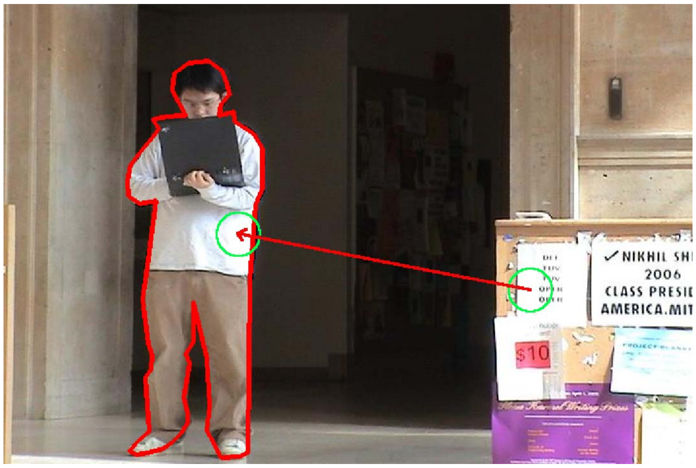

Figure 12. Example of a sequence scene where the first saliency-guided attention resulted in a miss, but the second was able to hit the labeled object. Image size: 720 × 480 pixels. IOR radius (green circle): 45 pixels (1/16 of the image width).

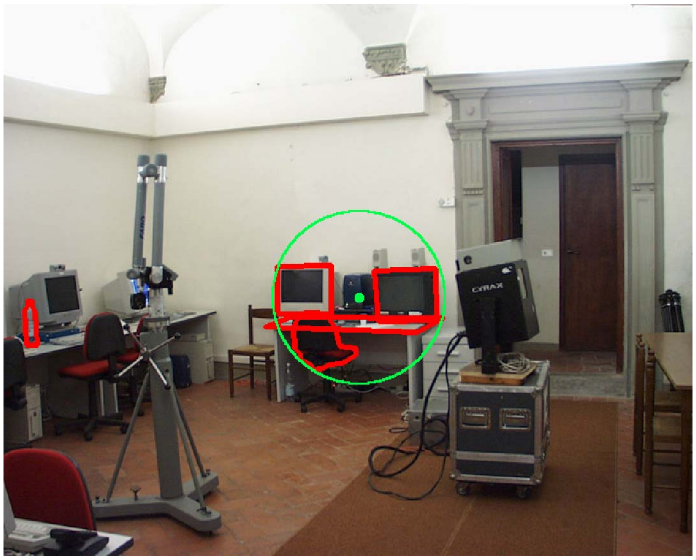

Figure 13. Example of a static scene where the first saliency-guided shift of ROI was a miss and it additionally inhibited four labeled objects due to a large IOR. Image size: 330 × 272 pixels. IOR radius (green circle): 160 (1/4 of the image width). Hence, these four objects will not be found since our implementation includes a perfect IOR with no decay for the purposes of this study. Images like this one explain why the asymptotes in Figures 9 and 10 do not necessarily tend towards 100% for saliency-guided predictions.

maximum saliency over the entire image is far from 100%, which would indicate that the most salient location was always considered so interesting as to deserve being labeled. This is well expected since the saliency map only computes bottom-up information and has no notion of what an object is. As a result, humans might use other strategies that involve top-down processing to label objects, along with bottom-up information. For example, is has been shown that context information can be a good predictor of were people would look when searching for specific objects (Oliva & Schyns, 1997; Oliva & Torralba, 2001; Torralba, Oliva, Castelhano, & Henderson, 2006). However, in the *LabelMe* data set, users have not been given a search task and are free to label anything they want. This could have partly removed the influence of contextual information. There still could have been some context information due to the fact that people would expect certain objects to exist in a particular scene, which might bias them toward these objects. In addition, other factors could have effected the selection of one object from another. For example, a less complex object might be labeled because it is easier to outline as opposed to a more complex object that might be more salient (e.g., a dull ball vs. a salient human), or an object might be labeled because of a more sophisticated measure of subjective interest (e.g., deriving from object recognition or higher-level mental processes). However, our further results looking at several successive shifts of ROI indicate that by just considering the top three most salient locations, one would hit at least one labeled object in 76% of all images. This is a remarkable figure given the simplicity of the saliency mechanism implemented here.

Although we have here focused on within-image selection bias (which objects were outlined), it is also possible that an image-level selection bias exists, in that subjects may have preferred to submit, for example, well-lit and in-focus images over darker and out-of-focus ones. It is useful to note that our measure of salience is normalized within each image in our analyses; that is, whether an image has overall low or high contrast, is crisp or blurry, or is noiseless or noisy, our saliency map will regardless rank-order locations within that image according to their relative salience. Likewise, annotators will select some of the objects in the image and annotate them. Here we have tested whether there is something that can be said about the objects that were selected compared to those that were not. It is an exciting question for future research to test whether some image-level selection bias might also exist in the *LabelMe* and in other data sets. This could be for example investigated by comparing the distribution of absolute peak salience values in all *LabelMe* images to those in a broader representative and unbiased sample of all images of the world. Obviously, a difficulty here is in gathering said representative sample of images (e.g., using cameras placed at randomly drawn GPS coordinates and mounted on randomly controlled motorized pan/tilt heads).

Some of the objects that were outlined but were not considered salient by the system were objects which can be considered having high-order saliency. High-order saliency simply means objects that are salient in a space which is not properly captured by the simple low-level visual features (orientations, brightness, color) implemented in our model. This space often contains objects which are not salient for their features but for their meaning. For example, some scenes contained labeled people, heads of people, or animals. These entities, however, did not have any salient feature according to our model, like a bright colored shirt. However, since people or animals have a meaning to other people beyond their features, they were labeled as objects of interest. Whether there is a simple neural correlate to such higher-order salience remains an open research question that pertains to deciphering which features may be able to effectively guide human attention (for a review on this topic, see Wolfe and Horowitz, 2004). In addition, objects which were out of place in the scenes were also often considered of interest and labeled, like a dog in an indoor party scene (for related findings, see Henderson & Hollingworth, 2003). Since the current saliency model does not consider these high-order features, it was not able to detect them. Future work should be considered into coding and detecting salient features in high-order space (see, e.g., Lee, Buxton, & Feng, 2005).

In the standard saliency model, an inhibition of return (IOR) is used to prevent the most salient location from being chosen again as the next predicted location. This is often implemented by suppressing the activation within a given neighborhood of the currently attended location. As a result, the next maximum location outside this neighborhood can be chosen. In the present work, the saliency map was inhibited based on a fixed-size disk with a radius of 1/4 and 1/16 the image width. This radius should approximately match the size of the objects that the saliency map is trying to locate. If one expects the saliency map to contain many small objects, then that value should be small, whereas when one expects mainly large objects, that value should be larger to speed up the search. Our results show that within three predicted locations, the saliency map yields over 76% hit rate. That is, within three predicted location, the saliency map finds a labeled object 76% of the time without having any notion of what an object is. Additionally, in Figure 10 we see that within the sixth predicted location, the saliency map starts to reach its limitations in finding objects. Furthermore, the system performs asymptotically better with sequences than with static scenes. Close examination of the scenes revealed that the sequence scenes typically contained labeled objects that were well separated (e.g., see Figure 12), while the static scenes contained labeled objects that more often were overlapping. Therefore, in the static scenes, after issuing an inhibition of return, the saliency map might have inhibited objects that were nearby (e.g., see

Figure 13), which then would never be found. This in part explains why asymptotical performance is better when the inhibition of return size is smaller (1/16 of the image width) since fewer such spurious inhibitions then occur. Ultimately, after choosing a location, a more sophisticated method should be used to identify the object and to inhibit only that object and not everything within a fixed radius of the attended location. This would be expected to result in more objects being found by the saliency-based approach since it would not accidentally inhibit objects. In addition, a dynamically sized IOR could be used to increase the speed of finding objects as well as the number of objects.

From the results above, one can determine that presence of low-level features along with some evaluation of their uniqueness in the visual field, as computed by the saliency model, are one of the reasons why human subjects would consider a visual location as interesting to look at. However, such elementary visual properties of scene locations only prime subjects toward attending to salient locations and toward subsequently extracting more information at these locations. There is still further processing required at these location to evaluate what these objects are and remember them, as demonstrated for example by the inattentional blindness phenomenon (Mack & Rock, 1998; Rees, Russell, Frith, & Driver, 1999; Simons & Chabris, 1999). This is also true for the *LabelMe* images evaluated in this paper. The various individuals who labeled the images did not have time constraints in their task and hence could first look at every location they liked and then decide which ones to label. Furthermore, their task was to label as many objects in the images as possible, which had nothing to do with saliency or attention. Nevertheless, our results demonstrate that as most people labeled only objects that they thought were of interest, this in turn led them to label objects that they had fixated as predicted simply by their salient properties.

We therefore conclude that the saliency map is a strong indicator of what people chose to label in complex natural scenes. In particular, the saliency map showed a 43% chance of finding a labeled object within the first predicted location and over 76% chance within the third predicted location. This means that even though choosing objects to label might seem like a “free” decision (that is, predominantly relying on higher-order mental abilities to evaluate and choose objects), humans are largely bound by bottom-up processing that influences their higher decision. However, top-down processing does play a role in which region we would choose to label, but only within the salient regions and not from the whole scene. This can be seen in Figure 11, where the user had a number of choices in labeling the scenes, but chose to label only salient locations (like the car). As a result, the saliency map can be used to provide a good indication of what people would deem worthy in a particular scene.

The conclusions derived from our study can be generalized to other data sets and tasks and can help solve some of the most challenging problems in computer vision and the visual sciences, in particular, the binding problem and the efficiency problem in object detection. Since the bottom-up saliency processing can yield to only a single region, the biding problem is simplified due to limited features (belonging to one object) that go to further processing. This then provides an automatic feature binding at low levels. For example, if a red square is deemed salient and is currently under the region of interest, then the edges of the square can be dynamically bound with the color of the square. The success of the feature bindings follows from the assumption that salient regions would contain multiple salient features (like edges and colors) that can be bound together for object recognition. In addition, efficient search can be performed on images since only a few regions needs to be considered. As a result, this information can be used for object detection and object recognition algorithms, robotics, navigation, and many other applications.

Acknowledgments

This work was supported by the Human Frontier Science Program (HFSP), the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), the National Geospatial-Intelligence Agency (NGA), and the Office of Naval Research (ONR).

Commercial relationships: none.

Corresponding author: Lior Elazary.
Email: elazary@usc.edu.
Address: Hedco Neuroscience Building Room 10, Mail Code 2520 3641 Watt Way Los Angeles, CA 90089-2520, USA.

References

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193–222. [PubMed]

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes; is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1476–1492. [PubMed]

Folk, C. L., & Annett, S. (1994). Do locally defined feature discontinuities capture attention? *Perception & Psychophysics, 56,* 277–287. [PubMed]

Franconeri, S. L., Hollingworth, A., & Simons, D. J. (2005). Do new objects capture attention? *Psychological Science, 16,* 275–281. [PubMed]

Franconeri, S. L., & Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception & Psychophysics, 65,* 999–1010. [PubMed] [Article]

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

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

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12,* 1093–1123.

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

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

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

James, W. (1890). *The principles of psychology*. Cambridge, MA: Harvard University Press.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics, 43,* 346–354. [PubMed]

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24,* 175–219. [PubMed]

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

Lee, K. W., Buxton, H., & Feng, J. (2005). Cue-guided search: A computational model of selective attention. *IEEE Transactions on Neural Networks, 16,* 910–924. [PubMed]

Mack, A., & Rock, I. (1998). *Inattentional blindness*. MIT Press.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45,* 205–231. [PubMed]

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34,* 72–107. [PubMed]

Oliva, A., & Torralba, A. (2001). Modeling the shape of the Scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes.' *Nature, 398,* 34. [PubMed]

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

Quinlan, P. T., & Humphreys, G. W. (1987). Visual search for targets defined by combination of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception & Psychophysics, 41,* 455–472. [PubMed]

Rees, G., Russell, C., Frith, C. D., & Driver, J. (1999). Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science, 286,* 2504–2507. [PubMed]

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

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2005). *LabelMe:* A database and Web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28,* 1059–1074. [PubMed]

Snowden, R. J. (2002). Visual attention to color: Parvocellular guidance of attentional resources? *Psychological Science, 13,* 180–184. [PubMed]

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

Theeuwes, J. (1994). Endogenous and exogenous control of visual selection. *Perception, 23,* 429–440. [PubMed]

Theeuwes, J. (1995). Abrupt luminance change pops out; abrupt color change does not. *Perception & Psychophysics, 57,* 637–644. [PubMed]

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

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 459–478. [PubMed]

Turatto, M., & Galfano, G. (2001). Attentional capture by color without any relevant attentional set. *Perception & Psychophysics, 63,* 286–297. [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]

Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika, 34,* 28–35.

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

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 419–433. [PubMed]

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews, Neuroscience, 5,* 495–501. [PubMed]

Yantis, S., & Jonides, J. (1996). Attentional capture by abrupt onsets: New perceptual objects or visual masking? *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1505–1513. [PubMed]