What do saliency models predict?

Kathryn Koehler
Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA, USA

Fei Guo
Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA, USA

Sheng Zhang
Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA, USA

Miguel P. Eckstein
Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA, USA

Saliency models have been frequently used to predict eye movements made during image viewing without a specified task (free viewing). Use of a single image set to systematically compare free viewing to other tasks has never been performed. We investigated the effect of task differences on the ability of three models of saliency to predict the performance of humans viewing a novel database of 800 natural images. We introduced a novel task where 100 observers made explicit perceptual judgments about the most salient image region. Other groups of observers performed a free viewing task, saliency search task, or cued object search task. Behavior on the popular free viewing task was not best predicted by standard saliency models. Instead, the models most accurately predicted the explicit saliency selections and eye movements made while performing saliency judgments. Observers' fixations varied similarly across images for the saliency and free viewing tasks, suggesting that these two tasks are related. The variability of observers' eye movements was modulated by the task (lowest for the object search task and greatest for the free viewing and saliency search tasks) as well as the clutter content of the images. Eye movement variability in saliency search and free viewing might be also limited by inherent variation of what observers consider salient. Our results contribute to understanding the tasks and behavioral measures for which saliency models are best suited as predictors of human behavior, the relationship across various perceptual tasks, and the factors contributing to observer variability in fixational eye movements.

Introduction

When humans view natural scenes without a particular task in mind, it has been proposed that their eyes are drawn to areas or objects that stand out amongst the background. These conspicuous regions are often referred to as salient, and the extent to which they guide human fixations has been a widely studied topic (Borji & Itti, 2013; Tatler, Hayhoe, Land, & Ballard, 2011). Researchers have recognized the utility of quantifying visual saliency and using that information to make behavioral predictions. Computational models of saliency take images as input and output a topographical map of how salient each area of that image is to a human observer. The output of these models have in fact been shown to be capable of predicting certain aspects of human eye movement selection when that behavior is driven by basic, bottom-up influences (e.g., Foulsham & Underwood, 2008) or an interaction of low-level image features (Peters, Iyer, Itti, & Koch, 2005).

How saliency is defined computationally varies widely across the literature. Among the various model approaches there has been agreement on the general notion of bottom-up saliency; however, the algorithms that implement this approach differ greatly by taking into account different degrees of local, global, or a combination of these image features (Borji & Itti, 2013). Most models of saliency computationally encode

Citation: Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision, 14*(3):14, 1–27, http://www.journalofvision.org/content/14/3/14, doi:10.1167/14.3.14.

doi: 10.1167/14.3.14 Received July 24, 2013; published March 11, 2014 ISSN 1534-7362

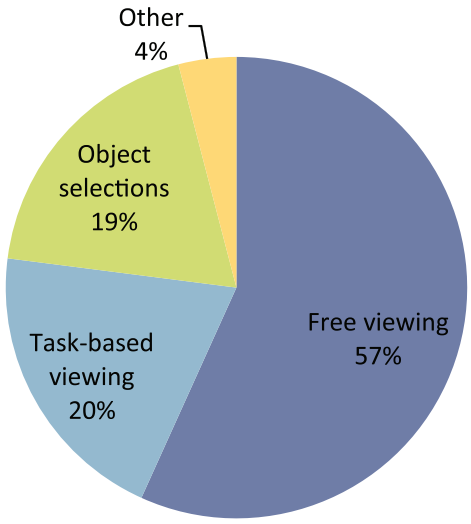

Figure 1. Types of tasks used to assess the predictions of saliency models in 74 papers. The list of papers included in this figure is not intended to be exhaustive, but generally representative of the popularity of various benchmark dataset types.

how different an area is from what surrounds it (e.g., Itti, Koch, & Niebur, 1998) based on a model of a biologically plausible set of features that mimic early visual processing. This definition formed the original conceptualization of a saliency map by Koch and Ullman (1985). Koch and Ullman's original saliency map was constructed from feature maps of basic visual elements (e.g., color, orientation, disparity, etc.). These were combined into a saliency map, upon which attention could be focused via a biologically plausible winner-take-all (WTA) process. Information from the attended area was then extracted to form a nontopographical central representation. The current WTA selection was then inhibited in order to move the eyes on to the next winner chosen from the saliency map.

The original model of visual saliency was developed to predict the allocation of covert attention (Koch & Ullman, 1985), although most studies have assessed the ability of the model to predict eye movement patterns while observers freely view images without a particular task (e.g., Bruce & Tsotsos, 2006; Itti & Baldi, 2005; Parkhurst, Law, & Niebur, 2002). Figure 1 shows the frequency with which four different types of tasks have been used in a representative sample of papers that have assessed the predictions of saliency models. A full list of the papers that are included in this figure can be found in the Appendix. The most common comparison used to gauge the predictive power of saliency models is fixations from free viewing tasks. The free viewing paradigm consists of participants viewing images (for 2–11 s) or short videos without a particular task in mind. The second most common comparison for saliency predictions is to human eye movements with more focused tasks (e.g., subsequent memory test, Tatler, Baddeley, & Gilchrist, 2005; description of social interaction between two people in the image, Birmingham, Bischof, & Kingstone, 2009; search and localization of an object, Itti & Koch, 2000; etc.). Another more recent benchmark is to compare saliency maps to images where objects have been selected or segmented by human observers or object recognition models. Newer models are often compared to databases of images (e.g., Liu, Sun, Zheng, Tang, & Shum, 2007), where foreground objects are enclosed in a rectangle by human observers, or where binary ground truth maps indicate whether or not each pixel in an image corresponds to an object (Achanta, Hemami, Estrada, & Susstrunk, 2009).

The utility of saliency models, as they relate to predicting human eye movement behavior, has been a topic of debate since their introduction. Competing models that incorporate information about a visual search target and its context, referred to as top-down information, have also had success in predicting visual search behavior (Beutter, Eckstein, & Stone, 2003; Najemnik & Geisler, 2005; Rao, Hayhoe, Zelinsky, & Ballard, 1996; S. Zhang & Eckstein, 2010). Even though entirely bottom-up driven models have been mostly rejected as fully accounting for human eye movements (Einhäuser, Rutishauser, & Koch, 2008; Smith & Mital, 2011; Tatler et al., 2011; Torralba, Oliva, Castelhano, & Henderson, 2006), the debate about the extent of the importance and contribution of bottom-up information still remains. Studies have demonstrated that adding bottom-up information into top-down models is detrimental to predicting eye movements of observers (e.g., Zelinsky, Zhang, Yu, Chen, & Samaras, 2005) and that fixations are more directed to objects than salient (nonobject) regions (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013). On the other hand, there are a number of studies that suggest that bottom-up information has an independent contribution to eye movement behavior (Kollmorgen, Nortmann, Schröder, & König, 2010) and that the task constraints determine the contributions of top-down and bottom-up processes (Bonev, Chuang, & Escolano, 2013).

Furthermore, due to varied methods of testing saliency map predictions, it is still not entirely clear what components of human behavior saliency predictions account for. Most notably, saliency models have not been assessed in their ability to predict a task in which the observer searches for the most salient region in the image. Our main goal is to further the understanding of behaviors predicted by saliency models by systematically evaluating the ability of various saliency models to predict human behavior

while engaged in four different tasks utilizing the same set of images. In particular, we expand on a new explicit judgment task (Koehler, Guo, Zhang, & Eckstein, 2011) and provide a dataset of location judgments of saliency (mouse click selections). Recent work by Borji, Sihite, and Itti (2013) provides a complementary assessment of explicit saliency in the form of boundaries around subjectively salient objects. We also include the typical free viewing task, an explicit saliency search task, and an object search task. In order to evaluate the generality of the task differences, three state-of-the-art saliency models (Bruce & Tsotsos, 2006; Itti & Koch, 2000; Zhang, Tong, Marks, Shan, & Cottrell, 2008) were used to predict human behavior, and performance was analyzed using three standard metrics. Our primary goal is to gain a better understanding of the tasks for which the models best predict human behavior and also to further understand the relationship among eye movements made during various perceptual tasks. The images and behavioral dataset will be provided for public use.

In addition, a secondary interest of the current work is to investigate the variability of human eye fixations across the different tasks. Tatler et al. (2011) have noted that free viewing tasks may prompt observers to view images in a highly variable, individual, and experimentally unknown way. Yet, there have been few comparisons of interobserver variability in eye movements across different visual tasks with real scenes. Thus, a secondary goal of the current work is to compare behavioral variability in observers' fixations and choices in the visual tasks investigated.

Tasks evaluated

The four tasks were chosen to emulate trends in existing data sets and to provide novel data sets that have not yet been compared to saliency models. In the first task, observers selected the object or region in a set of images that they considered to be most salient. This type of explicit saliency judgment task has never been compared to saliency model predictions.[1] The other three tasks were all conducted while participants' eye position was recorded. Different participants either completed a free viewing task, a task where they viewed an image and determined which half of an image (left or right) was most salient, or a task where they searched for a cued object and reported whether the object was present or absent.

Saliency models evaluated

There is a large selection of competing saliency models in the literature to test (Borji & Itti, 2013). Here, we concentrated on three different leading saliency models. They are all widely used as state-of-the-art benchmarks in novel model comparisons and represent three diverging attempts at operationalizing saliency. The Itti and Koch model (2000), which we will refer to as the IK model, was chosen due to its overwhelming popularity in the literature. The IK model is an updated version of the approach proposed in Koch and Ullman (1985). It has been highly influential and has served as a benchmark against which to compare subsequently developed models.

We also evaluated the Attention based on Information Maximization (AIM) model by Bruce and Tsotsos (2009). This model was developed in an effort to provide a computational architecture that was grounded in explaining biological organization. Rather than focusing solely on local surrounding areas to define saliency, a more global context was incorporated for each point in the map. They trained their model on a set of image patches analyzed using independent component analysis (ICA). They utilized ICA in order to ensure sparse encoding during later computing stages. Each point on an input image was then analyzed by the model to compute how much self-information it provided to the observer.

Finally, the Saliency Using Natural Statistics (SUN) model by Zhang et al. (2008) was analyzed. Like the AIM model, the computational definition of saliency at the root of this model emphasizes the probability of what is present at a particular location being at that location in an image. In other words, the model aims at computing the amount of surprise in each region of the image. The SUN model computes this probability from a Bayesian framework by estimating the point-wise mutual information of each point in an image. Output from all three models was used to calculate the top five most salient regions in a database of 800 images.

Metrics to compare saliency models to human behavior

The ability of these three models to predict human behavior was compared using three different metrics during analysis. A Region of Interest (ROI) metric was used to show the proportion of mouse selections (i.e., clicks) or fixations within a circular region surrounding locations of high saliency according to the models. A distance metric was used to show the average distance of each click or fixation to the top five salient locations as predicted by the models. Finally, the area under the receiver operating characteristic (ROC) curve was used following a procedure earlier established by Tatler et al. (2005). The Methods section provides further details about the metrics.

Figure 2. Example of images in database. Natural scenes are indoors and outdoors with a differing number of salient objects in each.

Methods

Human data

The data and stimuli from this project are publicly available at https://labs.psych.ucsb.edu/eckstein/miguel/research_pages/saliencydata.html.

Participants

All participants were undergraduate students (ages 18–23) at the University of California, Santa Barbara, who received course credit for participation. All participants had normal or corrected-to-normal vision. Informed written consent was collected from all participants. One hundred observers performed an explicit saliency judgment task, 22 observers performed a free viewing task, 20 observers performed a saliency search task, and 38 observers performed a cued object search task. Unequal sample size across the eye movement tasks resulted from inadequate eye tracking data or observer attrition during the experiment. Each observer in the cued object search task viewed only half of the images (the first or last 400). Time constraints prevented us from having each subject finish all 800 images in that condition.

Stimuli

Stimuli consisted of a database of 800 real scenes, photographed by researchers in the lab or collected from online search engines or databases (Russell, Torralba, Murphy, & Freeman, 2008), comprising both indoor and outdoor locations with a variety of sceneries and objects (see Figure 2 sample images). The images could be viewed in isolation, but were designed to contain lateral (left or right) contextual information for a tangible object. That is, each image was staged or chosen such that a contextually relevant object would be expected to exclusively appear either on the left or right portion of the image. In the object search task, prior to presentation of the real scene we presented a word (e.g., car) indicating the target to be searched for. Targets were present in half of the images. For example, an image depicting a kitchen with a stove located to the right might be paired with the object, "frying pan," present in the image. Alternatively, an

image showing a park with a tree on the left-hand side may be paired with the object, "bird," not present.

In the explicit saliency judgment task, the only task where eye-tracking data was not collected, viewing distance was not strictly enforced but remained approximately 40 cm. Images were displayed at a size so that they subtended 15° × 15° visual angle. Images were centrally displayed on a gray background. The names of cued objects (e.g., frying pan), if applicable, were centrally displayed in black text on a gray background.

Apparatus

Click data for the explicit saliency judgments were collected using an LCD monitor. Eye tracking data were recorded using a tower-mounted Eyelink 1000 system (SR Research Ltd., Mississauga, Ontario, Canada) monitoring gaze position at 250 Hz. Viewing distance was selected so that a pixel subtended 0.037° on a CRT monitor. Stimuli were displayed on an 800 × 600 pixel resolution monitor. Fixations were calibrated and validated using a nine-point grid system with a mean error of no more than 0.5° every 80 trials. Recalibration was also performed in the case of large head movements between regularly scheduled recalibration times. Saccades were classified as events where eye velocity was greater than 22°/s and eye acceleration exceeded 4000°/s^2. The first saccade per trial was considered to be the first fixation outside of a 2° radius around the initial fixation point or the first fixation within a 2° radius with latency greater than 120 ms. A latency of 120 ms corresponded to the lower 10.6% rank of all fixations made during the experiment. Initial fixation was monitored so that if participants moved their eyes more than 1° from the center of the fixation cross prior to presentation of the image, the trial would be restarted.

Procedure for explicit judgments of most salient region

Observers were instructed to view a picture on a computer monitor and click on the object or area in the image that was most salient to them. "Salient" was described to observers as something that stood out or caught their eye. The example of a red flower among a field of white daisies was provided to observers who asked for more clarification. Each observer completed 10 practice trials with images not included in the study. The experimental sessions consisted of 800 trials, each displaying one of the 800 different images. During each trial, a central fixation stimulus was displayed until the participant clicked the computer mouse. It remained for 500 ms, and then an image appeared. The image was displayed until the participant selected the most salient region via a mouse click. After the selection, a black crosshair marker indicated the location of the click and the participant was asked to confirm his or her selection. If accepted, another trial was initiated. If rejected, the marker was blanked and the observer was allowed to make another selection.

Procedure for eye-tracked tasks

For each of the tasks where eye movements were recorded (free viewing, saliency search, and cued object search), every trial began with an initial fixation cross randomly placed either centered, 13° left of center or 13° right of center. After fixating the cross, a trial was initiated by pressing the space bar. An eye tracker monitored whether or not observers maintained fixation during a randomly jittered interval (500–1500 ms) prior to the presentation of the image. If fixation was broken during that time, the trial restarted. The image remained visible for 2000 ms while eye position data was recorded. Each observer completed a total of 800 trials.

Observers completing the free viewing task were instructed to freely view the images. No further instructions were given, even upon request for task clarification. Observers in the saliency search condition were instructed to determine whether or not the most salient object or location in an image was on the left or right half of the image. "Salient" was defined in the same way as during the explicit judgment task. For this condition, after the image disappeared, observers pressed a button on-screen indicating whether the most salient region was on the right or left half of the image. Upon selection, the image reappeared until observers clicked where they believed to be the most salient region in the image, thus completing a trial. Finally, observers who performed the cued object search task were instructed to determine whether or not a target object was present in a displayed image. After successful sustained initial fixation, and prior to image appearance, an object name was shown for 1000 ms. The object was semantically consistent with the image and the object was present in 50% of the images. After the object word cue and image display (again, lasting 2000 ms), observers pressed a button on-screen indicating whether they believed the object was present or absent. After a selection was made, a subsequent trial began.

Model predictions

All 800 images were processed by each of three saliency model toolboxes for MatLab. A saliency map

was obtained for each image, and the pixel coordinate location of the top five maximum saliency values was recorded. The top five coordinates were calculated such that they were surrounded by a circular region that was two degrees of visual angle (54 pixels; 2°). Saliency maps for the IK model were calculated using the Matlab toolbox provided by Walther (2006; model explained in Walther & Koch, 2006). The AIM saliency maps were obtained from code available on a website by Bruce (2006). Finally, the SUN maps were obtained from code available on Zhang's (2008) website. It should be noted that the saliency toolbox we used has been shown to sometimes produce different results than the original implementation provided by Itti's lab in the iNVT toolbox (Borji, Sihite, & Itti, 2012). Images were input to the models at half of their original resolution. All other default settings were used with each model to produce saliency maps. For all models, a custom algorithm was used to select the top five salient regions by simply rank-ordering the saliency values and selecting the greatest values surrounded by nonoverlapping regions. The biologically inspired WTA method of selecting top regions was not used for the IK model because it resulted in poorer performance than when using the custom algorithm that ensures no overlap.

Metrics used to compute performance

ROI

This metric is an intuitive way of capturing how well the model predictions are encompassing human behavior. We determined the proportion of clicks or fixations that fell within the top five salient regions of each model. Each region was circular with a 2° radius, corresponding to the approximate size of the human fovea, centered on the model generated top five locations. We calculated the average proportion of clicks or fixations within each region (top five saliency regions) and also collapsed across all five regions.

Distance computations

We determined the average distance of a click or fixation to the center of the nearest region centroid among the top five saliency regions. After all clicks or fixations were analyzed, the average of those clicks or fixations closest to each region was computed. Unlike the ROI metric, the distance metric is a continuous measurement—clicks and fixations are not discretized as either belonging or not belonging to a region. Therefore the value of the distance metric reflects how tightly clustered a group of fixations are around a top salient location.

ROC curve calculations

The click or fixation data for each image, across all observers, was recorded into a nonsmoothed binary map and compared to a binary map indicating values above an incrementally increasing threshold on each saliency map. Ground truth maps were constructed by recording binary values indicating the presence of a click or fixation at each pixel coordinate in each image. In standard signal detection theory terms, the ground truth map represents the presence or absence of a signal. The saliency maps were then binarized into a "yes-model map" based on whether each coordinate location was above or below a certain threshold. A one on the yes-model map would indicate that the saliency value there was above threshold, and that the model expected a click or fixation to be at that location. A yes-model map was calculated for thresholds varying by 0.001 between 0 and 1 (the minimum and maximum values in all saliency maps). For each image, a 1 in the yes-model map with a 1 in the corresponding location in the ground truth map was considered a hit. Alternatively, a 1 in the yes-model map with a 0 in the corresponding location in the ground truth map was considered a false alarm. In this way, a point on an ROC curve was plotted for each threshold value, and one ROC curve was generated for each image. The area under the ROC curve of each image was calculated using nonparametric trapezoidal area estimation from point to point (DeLong, DeLong, & Clarke-Pearson, 1988). Tatler et al. (2005) demonstrated that this metric is useful because it is not overly influenced by the statistically inflated effect of small image variances, the non-normal distribution of natural image characteristics, or the monotonic transformation of saliency values based on arbitrary selections of model parameters.

Control conditions

We implemented two control conditions against which to compare human behavior in the experimental data set. First, we randomly sampled fixations with uniform probability across all image pixels. The random fixations were used to calculate a random control ROI and distance metric. However, previous studies have found that fixations are not uniformly distributed across images of real scenes (see Tatler, 2007, for a thorough discussion of this phenomenon, the central fixation bias). For this reason, we ran a separate control condition where all of the click or fixation results for each image were randomly permuted so that they would be paired for analysis with the saliency map prediction from another randomly sampled image. For example, the saliency map for Image 1 may be paired with the fixation data from Image 23. The permutation method allows us to identify gains in the ability of a model to predict

selections/fixations arising from observer biases to click or fixate toward the center of the image. We ran the random control once for each model. In the case of ROI analysis, we compared the results to the average total area subtended by all regions across images for each model. The area comparison was made because some regions are cut off by the edge of the image, potentially causing the random results to be different across models, depending on how frequently the regions are cut off by an edge. The permuted condition was run for each task and model, making a matched comparison for each cell in the factorial design structure of this experiment. Within each task, the same permutations and random coordinates were compared with each model, replicating the repeated measure structure within tasks.

Individual differences

Variance

We calculated the average variance across observers of the x- and y-coordinate locations for each fixation (first through sixth fixations). The final metric is the average of the x- and y-coordinate observer variances.

Kullback-Leibler divergence

To compare the distributions of observers' explicit judgments or fixations with each other within tasks, we also computed the Kullback-Leibler divergence (KL divergence). KL divergence assesses the similarity between two distributions, in this case the selections or fixations of one observer to those of all others. In order to compute this value, we first had to compile the explicit judgment or fixation distributions across observers for each image. The distributions, $P(X, Y)$ and $Q(X, Y)$, were estimated by dividing each image into equal sized bins ($1^\circ \times 1^\circ$) and summing the number of fixations or explicit selections in each, $F(X, Y)$, then normalizing by the total number of fixations or selections in the image. The KL divergence between the two distributions is defined as:

$$KL = \sum_{X,Y} \log\left(\frac{P(X, Y)}{Q(X, Y)}\right) P(X, Y) \quad (1)$$

In order to avoid undefined distribution ratios, we added a small constant ($c = 10^{-5}$) to each bin, as in Tatler et al. (2005). Therefore, the calculation of the fixation or selection distribution was calculated as:

$$P(X, Y) = \frac{F(X, Y) + c}{\sum_{X', Y'} (F(X', Y') + c)} \quad (2)$$

$P(X, Y)$ was taken to be the distribution of a single observer's selections or fixations for an image. $Q(X, Y)$ was calculated by using all other observers' selections or fixations for the same image. We calculated KL divergence for each observer against all other observers, and then averaged the results across observers and images. We analyzed both the divergence between the first six fixations made by each observer and each fixation one by one. When analyzing the explicit judgment or just a single fixation, there was only one bin in the $P(X, Y)$ distribution containing a value close to one (but not exactly one because of the addition of the constant, c). In these cases, we still summed across all bins because the $Q(X, Y)$ distribution contained multiple nonzero bins. A KL divergence of zero indicates that $P(X, Y)$ and $Q(X, Y)$ are identical distributions. The less similar the distributions are, the greater the value of KL divergence.

Statistics

ROI and distance, collapsed across region

We analyzed an unbalanced two-factor mixed ANOVA with model (IK, AIM, or SUN) as the within-subjects factor and task type (saliency click, saliency search, free viewing, object search) as the between-subjects factor. To ensure that ANOVA was appropriate in each case, the normality of data in each cell was checked by visually inspecting QQ-plots of the observed quantiles (drawn from the data) versus theoretical quantiles (randomly drawn from a normal distribution). After verifying that each sample was normal, homogeneity of variance was verified using a Harley F-max test with bootstrapped samples of equal size. Ten thousand repetitions were performed for each metric, and the proportion of those repetitions above the critical value was computed to be nonsignificant. Multiple Bonferonni corrected t tests were performed to compare each ANOVA sample mean to its corresponding permuted control condition and each random control to a corresponding permuted control condition. Post-hoc and repeated contrast tests in the general linear model analysis package of SPSS were used to determine differences in between-subjects and within-subjects factors, respectively.

ROI and distance, by region

We analyzed a balanced within-subject ANOVA with region number (with levels 1–5 for the ordered top-five saliency regions) as the relevant factor to compare the mean metric values for each region. Each model and task was analyzed with a separate ANOVA, totaling 12 ANOVAs per metric. A repeated contrast was computed to test the differences between means of adjacent regions, i.e., region 1 to 2, 2 to 3, 3 to 4, and 4 to 5.

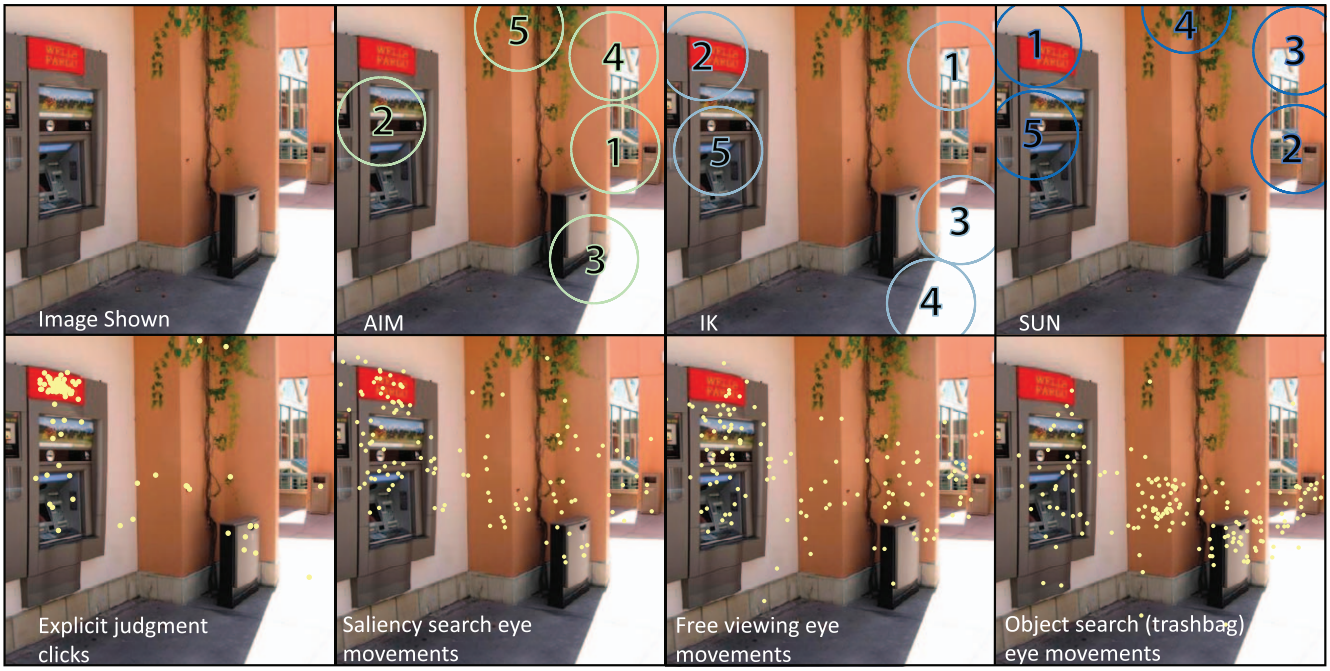

Figure 3. Top row: Example of top five salient regions output by AIM, IK, and SUN models respectively. Bottom row: Example of explicit judgment data, saliency search fixations, free viewing fixations, and object search fixations.

ROC

We analyzed a balanced two-factor mixed ANOVA with the same factors as the ROI and Distance metrics across regions, above. The same post-hoc and contrast tests as above were also computed. The ROC ANOVA was balanced, unlike the ROI and Distance metric designs because areas were computed across observers, per image, totaling 800 areas per cell.

Individual differences

To test the differences in fixation variance across tasks, we computed an F test for the equality of two variances for each image and each task combination (three combinations per image: free viewing and saliency viewing; free viewing and object search; saliency viewing and object search) for the first through sixth fixations. Therefore, there were a total of 18 fixation/task combination analyses, each comprised of 800 F-statistic calculations. For each of the 18 fixation/task combination analyses, we calculated the proportion of the 800 F statistics that were above a critical value.

Results

Figure 3 (top row) shows an example of the top five regions for the three models. Figure 3 also shows sample data for a single image for all four tasks. Figure 3 (bottom left panel) shows an example of click responses for the explicit judgment task. The median response time for 31 observers in the explicit judgment task was 1.16 ± 0.15 ms. Figure 3 also shows an example of every fixation made by each observer for an image in each of the eye-tracked tasks (bottom row, rightmost three panels; saliency search, free viewing, and cued object search tasks from left to right). There was a median of 6 ± 0.28 eye movements for the free viewing task, 6 ± 0.22 eye movements for the saliency viewing task, and 8 ± 0.22 eye movements in the object search task.

Comparison of the influence of tasks on models' ability to predict fixation/selections

ROI

We evaluated whether the behavioral predictions of all models varied across tasks. A clear trend can be seen in Figure 4, such that all three models predict behavior with variable success as the task changes, $F(3, 157) = 46.56$, $p < 0.001$. Bonferroni-corrected post-hoc tests revealed that the behavioral predictions generated by the models aligned most closely with human performance on tasks that probe explicit saliency judgments ($p < 0.001$). Furthermore, predictions for the saliency search task were marginally better than on the object search task ($p = 0.153$ after Bonferroni correction). The difference between predictions on the saliency and free viewing search tasks and free viewing and object search tasks were not significant. The proportion of human

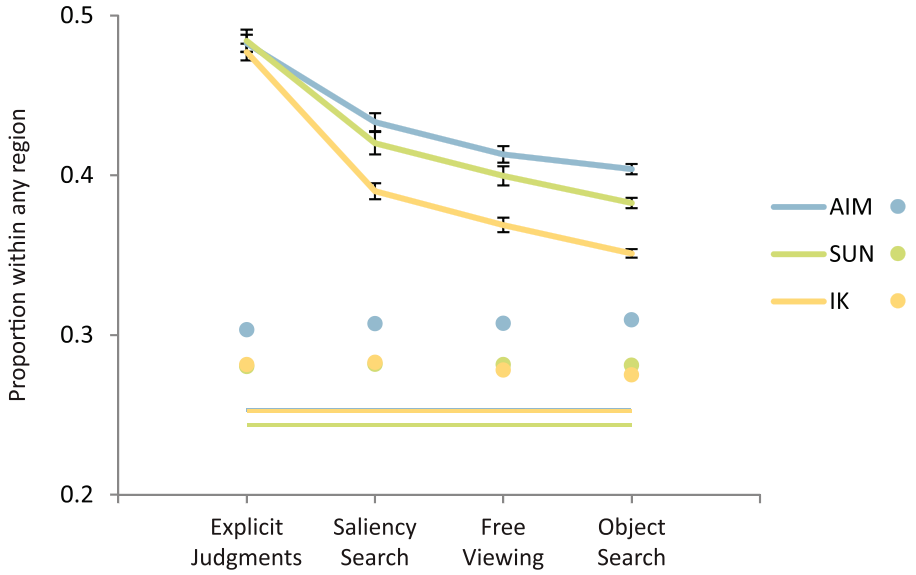

Figure 4. ROI results for each task (thick lines), with permuted (dots) and random controls (thin lines) shown. Each thick line represents the proportion of clicks or fixations within all of the top regions output by each model for the four tasks. Error bars represent *SEM*. Data points with connected lines do not represent continuous data, but are used to simplify the legend.

selections or eye movements within the regions predicted by each model was greater than that in permuted regions or randomly generated regions ($p < 0.001$ for each comparison). For each model, the random control was also compared to the permuted control for the click task. The models were significantly worse at predicting randomly generated clicks than permuted clicks ($p < 0.001$).[2] These results and findings, reported in subsequent sections, are summarized in Table 1.

We also wanted to determine if the highest saliency values in the maps were more likely to predict human fixation locations than lower saliency values. Figure 5 shows the proportion of selections or fixations within each region in ranked order (from highest to lowest top five saliency regions) by task for the experimental conditions. As can be seen in Figure 5, the proportion of selections or eye movements for the IK model within each region decreases from the first most salient to the fifth salient region. This suggests that areas of high value in the saliency maps are more likely to contain fixations or selections than areas of low saliency. This trend is the same across all tasks. Similar results are seen with the AIM and SUN models, as can be seen in Figure A1 in the Appendix. Repeated contrast analysis of the results from Figure 5, with ROI measures split up by region, revealed that each region was significantly different from the previous region for all models and tasks, $p < 0.05$.

Distance

We also explored the ability of the different saliency models to predict human eye movements for the various tasks using a distance metric. Figure 6 shows the average distance of each selection or fixation to the nearest region center for both the experimental (lines) and control conditions (points) by task.

The behavioral predictions of all models also varied between tasks, $F(3, 157) = 61.92$, $p < 0.001$. As with the ROI analysis, Bonferroni post-hoc tests of the distance analysis revealed that the behavioral predictions generated by the models aligned most closely with human performance on tasks that probe explicit saliency judgments. Similar to the ROI analysis, the models produced the best behavioral predictions on the

	Saliency search			Free viewing			Object search		
	ROI	Distance	ROC	ROI	Distance	ROC	ROI	Distance	ROC
Explicit judgments	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
Saliency search				0.649	<0.001*	0.072†	0.153	0.061†	<0.001*
Free viewing							1	0.149	0.038*

Table 1. *p* values for the post-hoc comparisons of model predictions between tasks, averaged across model, for each metric. *Note*: All values were Bonferroni corrected. *$p < 0.05$; †$p \sim 0.05$.

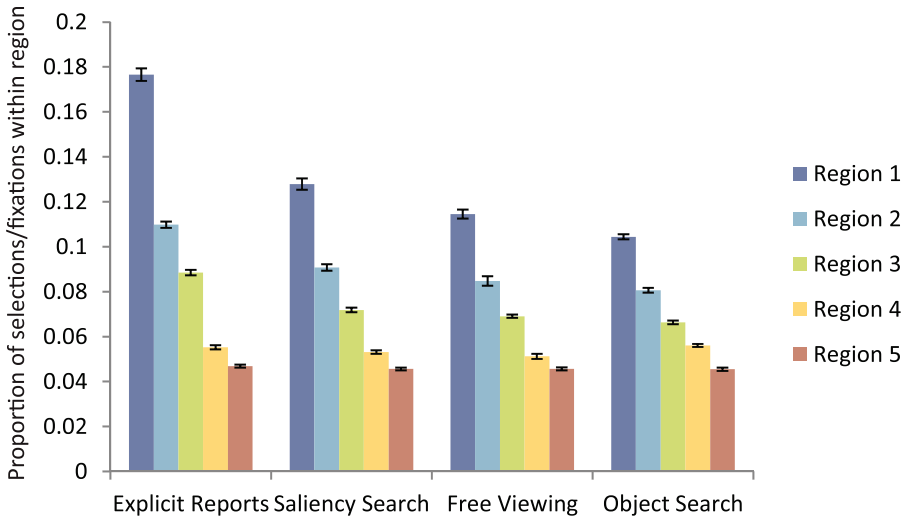

Figure 5. ROI results broken down by region for the IK model. Each bar represents the proportion of selections or fixations within that specific region, with regions ranked in order from most to fifth-most salient. Error bars represent *SEM*.

explicit saliency judgment task ($p < 0.001$). Unlike the ROI analysis, the saliency models were significantly better at predicting the saliency search task than the free viewing task ($p < 0.001$), but the difference between the saliency search and object search task was only marginally significant ($p = 0.061$). The difference between the free viewing and object search tasks was not significant.

As with the ROI metric, we found a decreasing ability to predict human behavior as saliency map

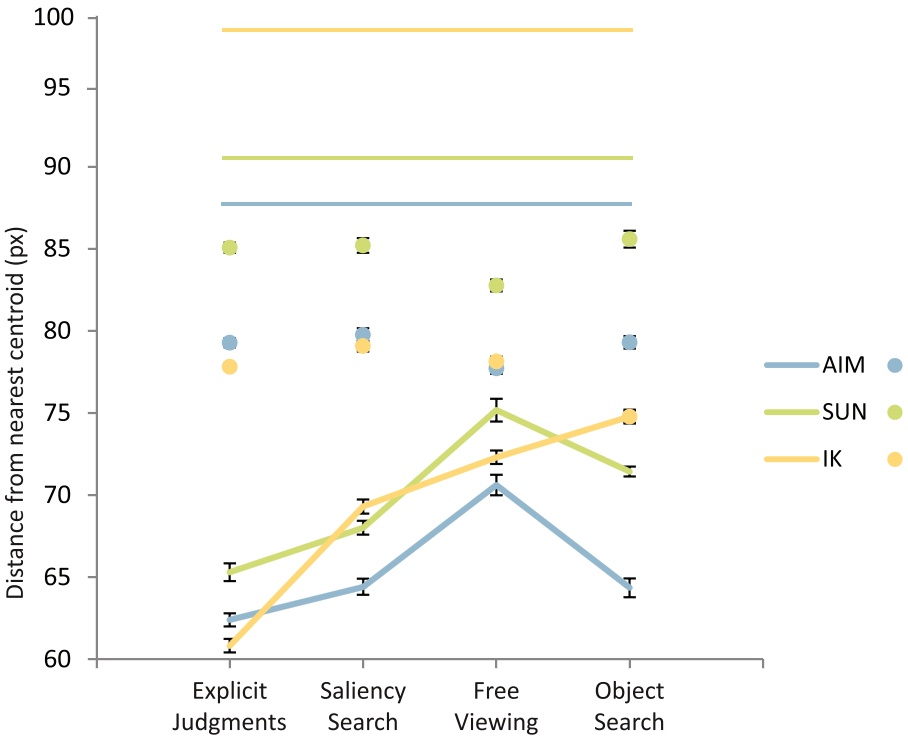

Figure 6. Distance results for each task (thick lines), with permuted (dots) and random controls (thin lines) shown. Each bar represents the distance of each selection or fixation to the nearest region for each model. Error bars represent *SEM*. The average distance of each selection or fixation to its nearest predicted region by each model was smaller than that in permuted regions or randomly generated regions ($p < 0.001$ for each comparison). The distance of salient regions from randomly generated clicks was much greater than that for permuted clicks ($p < 0.001$ for each model). Data points with connected lines do not represent continuous data, but are used to simplify the legend.

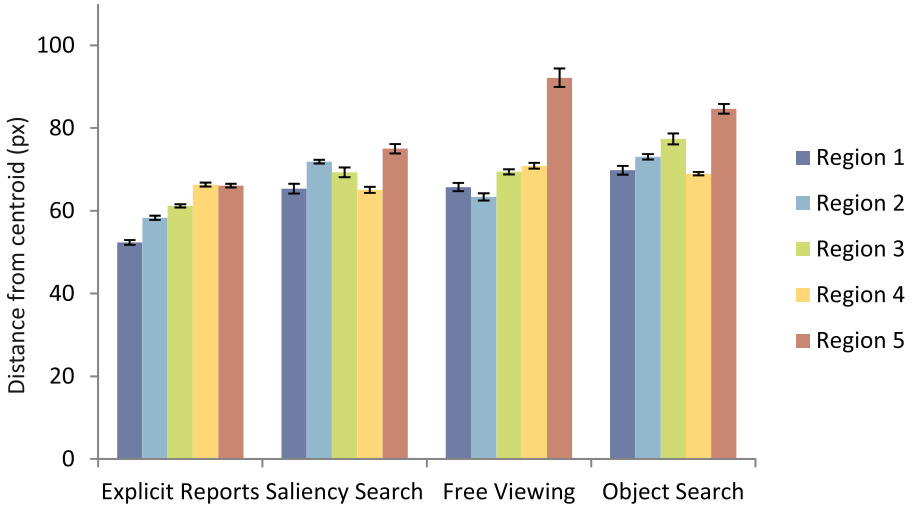

Figure 7. Distance results broken down by region for the IK model. Each bar represents the average distance of selections or fixations closest to the specified region, with regions ranked in order from most to fifth-most salient. Error bars represent *SEM*. Similar results are shown for the AIM and SUN models in Figure A2.

values decreased. Figure 7 shows the distance of each selection or fixation to its nearest top-region center in ranked order by task for the experimental conditions. In general, the average distance for the IK model regions to each selection or fixation increased from the first to the fifth generated region although there were some exceptions (Figure 7). Similar to the ROI analysis, this suggests that the highest saliency values more closely align with human behavior. Similar results were obtained with the AIM and SUN models, as can be seen in Figure A2 in the Appendix.

ROC

This section evaluates the ability of models to predict human fixations using the ROC metric across the four tasks. Figure 8 shows the average area under the ROC curve. Chance performance is indicated by the dashed line in the figure.

The behavioral predictions of all models also varied between tasks according to the ROC metric, $F(3, 3196) = 222.84$, $p < 0.001$. As with the distance analysis, Bonferroni-corrected post-hoc tests of the average area under the curve revealed that the behavioral predictions

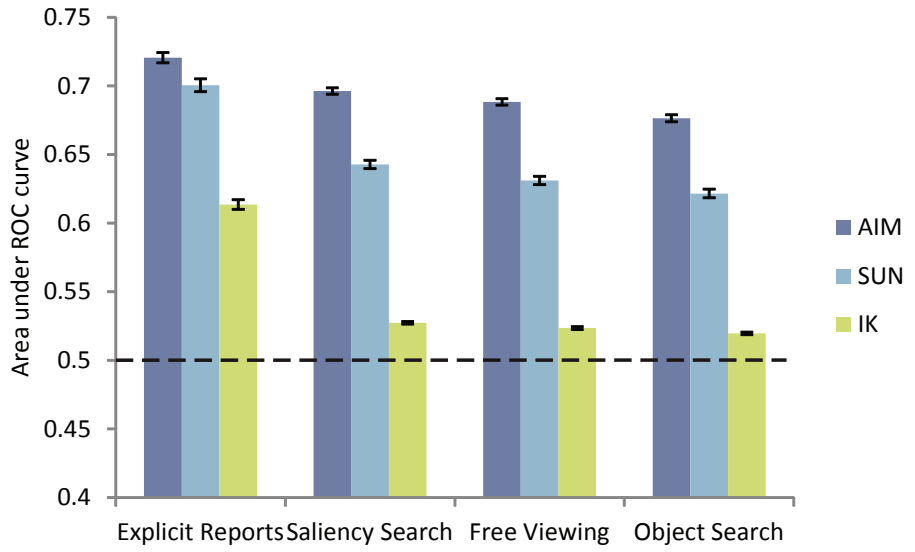

Figure 8. Area under the ROC curves, averaged across images by model and task. Error bars represent *SEM*, which was averaged across observers for each image.

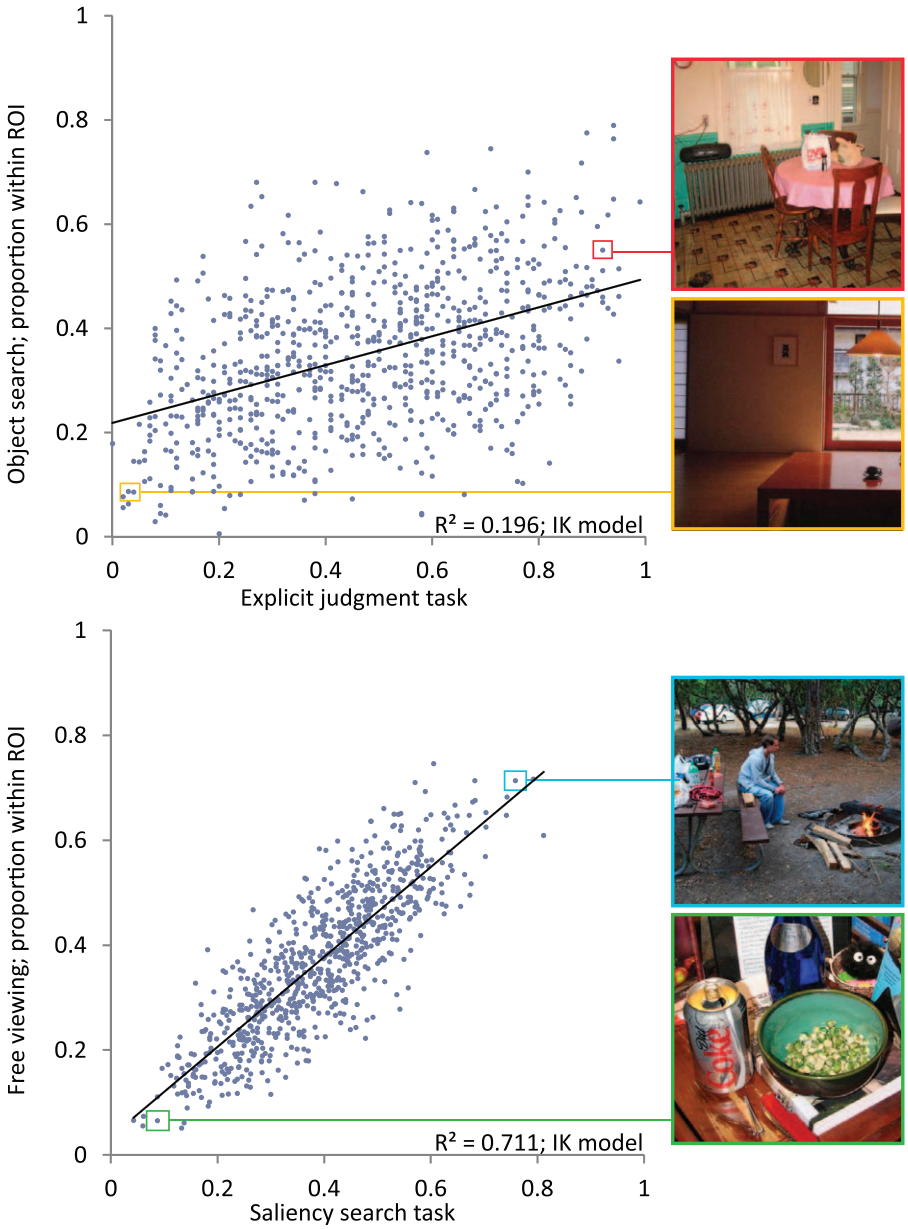

Figure 9. Correlation between the accuracy of the IK's models behavioral predictions of human behavior according to ROI metric on the explicit judgment versus object search tasks (top) and the saliency search and free viewing tasks (bottom). Images corresponding to points in the scatterplot are shown to illustrate representative examples of when the models are good predictors (red and blue) and poor (yellow and green) predictors of human behavior for both tasks.

generated by the models aligned most closely with human performance on tasks that probe explicit saliency judgments ($p < 0.001$). Similar to the ROI analysis, the behavioral predictions for the saliency search task were marginally better than those for the free viewing task ($p = 0.072$) and the behavioral predictions for the saliency search task were significantly better than for the object search task ($p < 0.001$). Finally, the ROC analysis was the only analysis to result in a significant difference in behavioral predictions between the free viewing and object search tasks, such that predictions were better for the free viewing condition ($p = 0.038$; see Table 1 for a summary of all results).

Correlation of model metrics across visual tasks

To further investigate task differences/similarities, we computed the correlation across images between the accuracy of each model's behavioral predictions (as measured by our three metrics) on two different tasks. The correlation assesses whether images that tend to produce fixations that are predicted well by the models

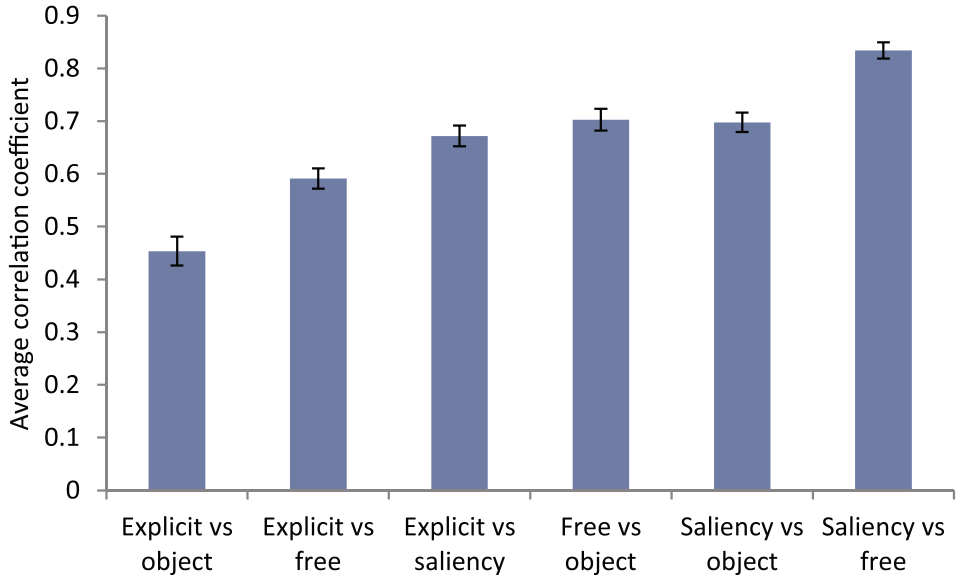

Figure 10. Average correlation coefficient, taken across metrics and models, for each task combination in the task correlation analyses. Error bars represent *SEM*.

for one task also produce fixations that are well predicted for the same image on another task, and similarly for images that produce poorly predicted fixations. We computed a total of 54 correlation coefficients (3 models × 3 metrics × 6 task combinations). Figure 9 shows two representative scatterplots with data taken from ROI analysis of the IK model. The top plot shows the somewhat weak relation between model predictions on the explicit judgment and object search tasks, $r(798) = 0.44$, $p < 0.001$, and the bottom plot shows the stronger relation between the saliency search and free viewing tasks, $r(798) = 0.84$, $p < 0.001$. The images corresponding to specific points in the scatterplot are shown to illustrate sample images with human behavior that was well-predicted or poorly predicted across tasks. In general, models are good predictors of human behavior for images with few or obvious stand-out objects (e.g., the bags in the red outlined images or the fire in the blue outlined image). Models poorly predict human behavior for images

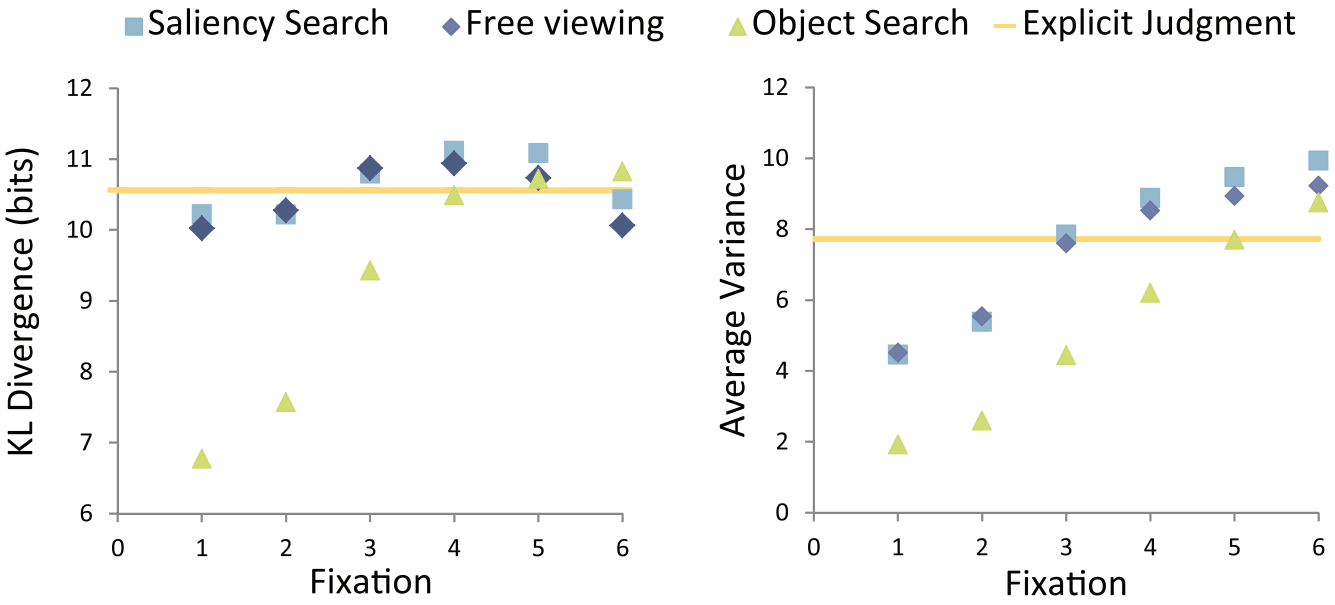

Figure 11. KL divergence (left) and variance in fixation location (right), averaged across observers and images for each task. The continuous yellow line represents the KL divergence and variance across observers for the single selection in the explicit judgment task. Error bar values, representing *SEM*, were too small to be visible in the figure, but ranged from 0.08–0.21 and 0.04–0.37 for KL divergence and variance, respectively.

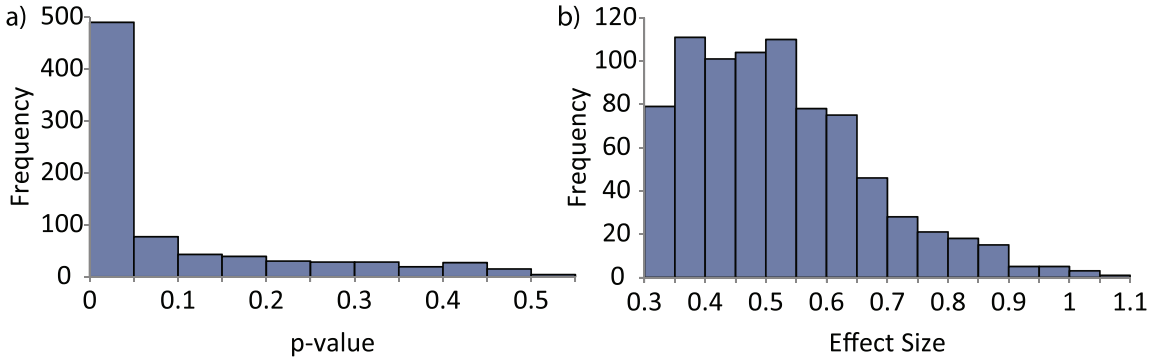

Figure 12. Histogram showing the frequency of the (a) *p* value and (b) effect size associated with each of the 800 *F* statistics calculated for the first fixation in the free viewing and object search comparison.

without any distinct salient objects, like the yellow outlined image, or with cluttered spaces, like the green outlined image.

For each pair of tasks we averaged the correlations across the three models and metrics (six correlations). Figure 10 shows the average correlation between pairs of tasks (all correlations were highly significant, $p < 0.001$). The strongest relation was between the free viewing and saliency search tasks and the weakest was between the object search and explicit judgment tasks.

Influence of tasks on differences across individuals

Figure 11 shows the average observer KL divergence (see Methods for details) and variance of fixation locations averaged across images for each task. KL divergence and variability in location increases for later fixations, and were lowest for the object search task.

Statistical analyses of the variance results are shown in Figures 12 and 13. Figure 12 shows a histogram across images of the *F*-test analysis results (testing for differences in fixation variances) for the first fixation when comparing the free viewing and object search tasks. The left portion of the figure shows the frequency of the significance values (*p* values) of each of the *F*-statistical tests computed for each of the 800 images. The right portion shows the frequency of the effect sizes (Cohen's *d*) associated with each *F* statistic. All of the effect sizes are greater than 0.3, with a median effect size of 0.5 (a medium strength effect). Similar histograms can be generated for each of the 17 other fixation/task combinations. A summary of the 18 *F*-test results are illustrated in Figure 13 which shows the proportion of *F* statistics below $p = 0.05$ for each fixation/task combination. From Figure 13, it is clear that there is a difference in fixation location variability between the free viewing and object search conditions and the object search versus saliency search conditions. This difference decreases as fixation number increases.

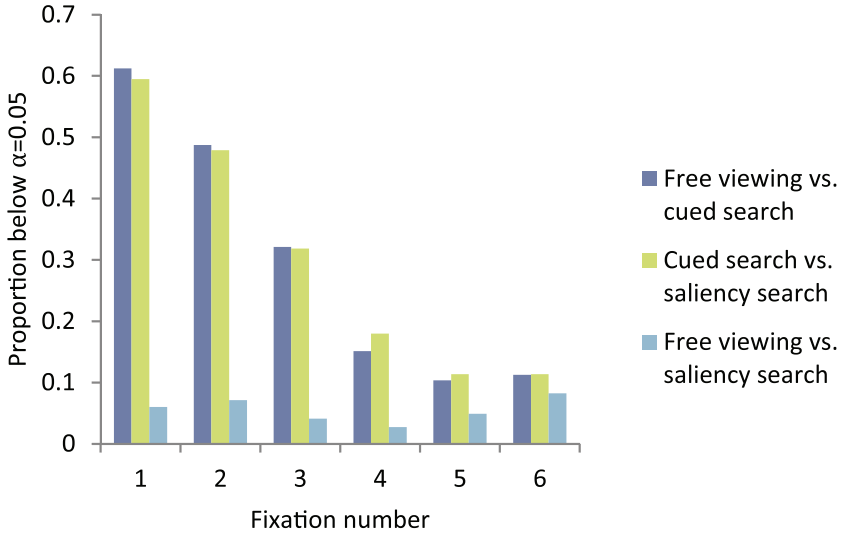

Figure 13. Proportion of *p* values below 0.05 for each fixation's *F*-test analysis comparing variances across tasks.

There is no difference in fixation variability between the free viewing and saliency search conditions. The KL divergence measure resulted in similar trends to the variance measure.

Of particular interest is to compare the variability in the eye movements in the saliency search task and free viewing task with the inherent variability in what observers judge to be the most salient region. We measured KL divergence and variance of observer selections made in the explicit judgment (continuous lines in Figure 11) as a measure of the inherent variability in what observers judge to be the most salient region in each image. Results (Figure 11) show that for the initial fixations (first and second) the observer variability (and to some extent the KL divergence) is smaller for the fixations than the explicit judgments. However, for later saccades, the fixation variability across observers is comparable and exceeded that of the explicit judgments.

Correlation of observer variance across tasks

To further assess the relationship between fixation variability across tasks we correlated observer variance in fixations across tasks. Figure 14 shows an example of two scatterplots illustrating the relationship between the variability of eye movements across observers on the free viewing and saliency search tasks. The top plot shows this relation when only analyzing the first fixation, $r(798) = 0.32$, $p < 0.001$, and the bottom plot incorporates the first six fixations, $r(798) = 0.76$, $p < 0.001$. Images corresponding to points in the scatterplot illustrate instances where there was high (yellow) and low (blue) variability between observers on both tasks. Images containing many objects tend to produce highly variable eye movements, whereas images with distinct salient objects, such as the red backpack in the blue outlined image in Figure 14, tend to produce minimally variable eye movements. Figure 15 shows the correlation coefficient between observer fixation variances of different task pairs. Results are shown for both the first fixation and first six fixations analyses for the fixation tasks (left), and for the first fixation compared to the explicit judgment (right). This comparison shows the largest correlation in observer fixation variance between the free viewing and saliency task, $r(798) = 0.76$, $p < 0.001$, further confirming that these two tasks are more related to each other than to an object search task. This result is consistent with comparing the explicit saliency judgment variability to fixation variability for the free viewing and saliency search tasks, as shown. All other correlations were also significant ($p < 0.001$, except first fixation free viewing versus object search and saliency search versus object search, $p = 0.005$ and $p = 0.019$, respectively, and for the explicit judgment versus cued object search, $p = 0.04$).

Comparison of ability of models to predict human behavior

ROI

In order to explore the differences in each models' ability to predict human behavior, we first compare the percentage of fixations/selections in all regions across all tasks. The three models varied significantly in their ability to predict human performance, $F(2, 314) = 81.57$, $p < 0.001$. According to the ROI metric, the AIM model was best able to predict human behavior across all tasks, performing better than both the IK and SUN models, $t(160) = 8.01$, $p < 0.001$ and $t(180) = 2.36$, $p = 0.019$, two-tailed, respectively. The SUN model also better predicted human behavior than the IK model, $t(160) = 5.68$, $p < 0.001$ (all t tests Bonferroni-corrected). Figure 4 shows the proportion of all selections or fixations within any of the top five salient regions for both the experimental (lines) and control conditions (points) by task. The proportion of selections or eye movements within an ROI decreased most drastically for the IK model across tasks, suggesting that behavioral predictions from this model are less reliable than other models on tasks that depart from explicit saliency judgments. The SUN model decreased slightly less across tasks than IK and the AIM model showed the smallest degradation in predictive performance as evidenced by a significant interaction, $F(6, 314) = 12.76$, $p < 0.001$.

Distance

Similar to the results of the ROI analysis, the three models varied significantly in their ability to predict human performance according to the distance metric as can be seen in Figure 6, $F(2, 314) = 170.62$, $p < 0.001$. Again, the AIM model was best able to predict human behavior across all tasks, performing better than both the IK and SUN models, $t(160) = 2.86$, $p < 0.001$ and $t(180) = 17.54$, $p < 0.001$, two-tailed, respectively. Unlike the results from the ROI analysis however, the IK model better predicted human behavior than the SUN model, $t(160) = 8.09$, $p < 0.01$ (all t tests Bonferroni-corrected). These results are summarized for comparison with the other metrics in Table 1.

ROC

Finally, similarly to the results of the previous two analyses, the three models varied significantly in their ability to predict human performance according to the

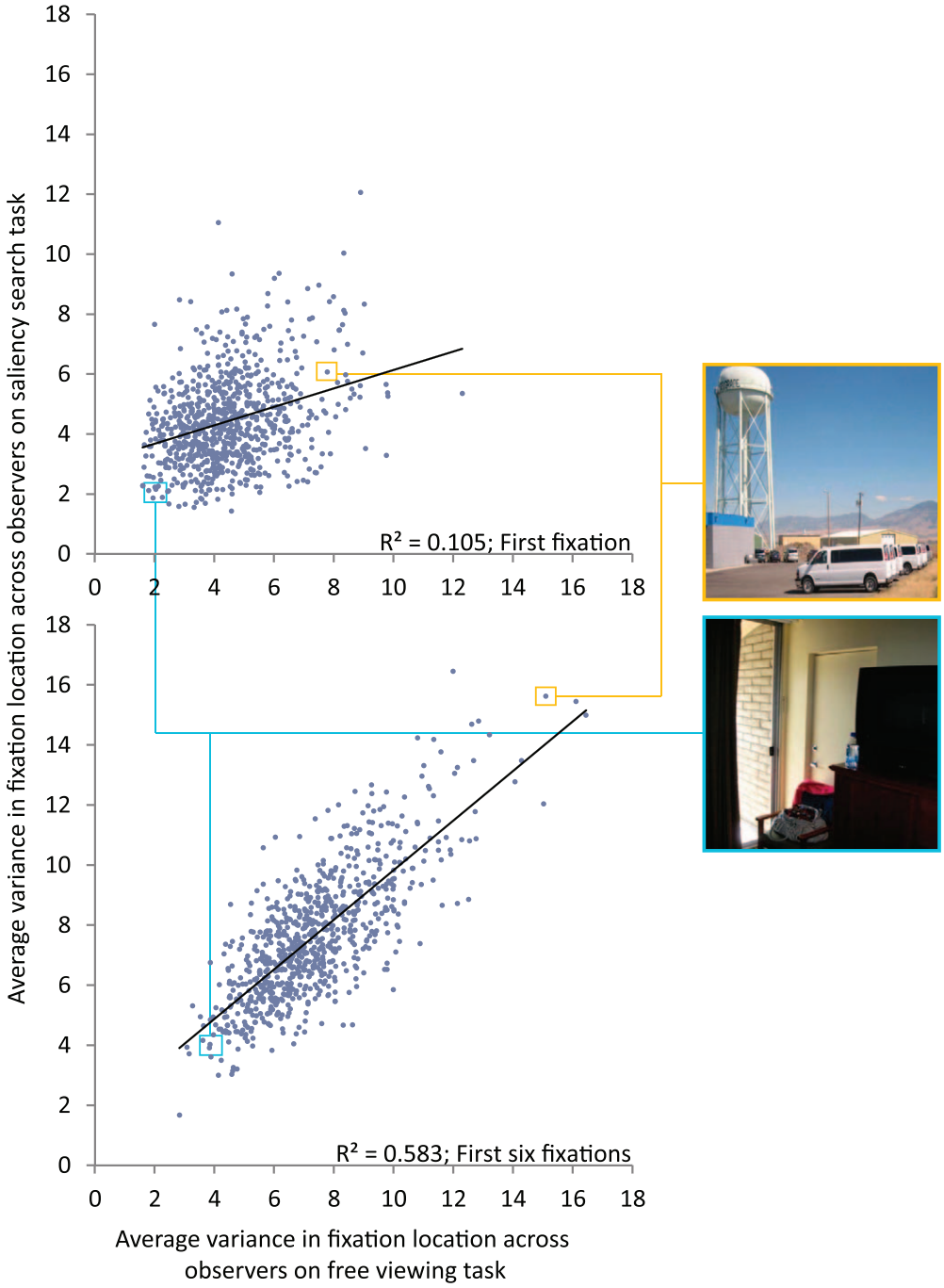

Figure 14. Scatterplots showing the relation between variability of fixation location across observers on the free viewing versus saliency search tasks. The top analysis shows variability only on the first fixation, and the bottom shows variability across the first six locations. Images corresponding to points in the scatterplot illustrate instances when there was high (yellow) and low (blue) variability across observers on both analyses.

ROC metric as can be seen in Figure 8, $F(2, 6{,}392) = 4{,}934.75$, $p < 0.001$. Again, the AIM model was best able to predict human behavior across all tasks, performing better than both the IK and SUN models, $t(1{,}399) = 98.85$, $p < 0.001$ and $t(1{,}399) = 35.92$, $p < 0.001$, respectively. Like the results from the ROI analysis, the SUN model better predicted human behavior than the IK model, $t(1{,}399) = 55.43$, $p < 0.001$ (all t tests Bonferroni-corrected). It is important to note that, due to its sparse output, the IK model does not lend itself well to ROC analysis. The ROC curves generated for most images are noticeably discretized; therefore, the area under the curves is often close to that expected by chance. Regardless of this, the IK model still predicted human behavior at an above chance level for the explicit judgment task.

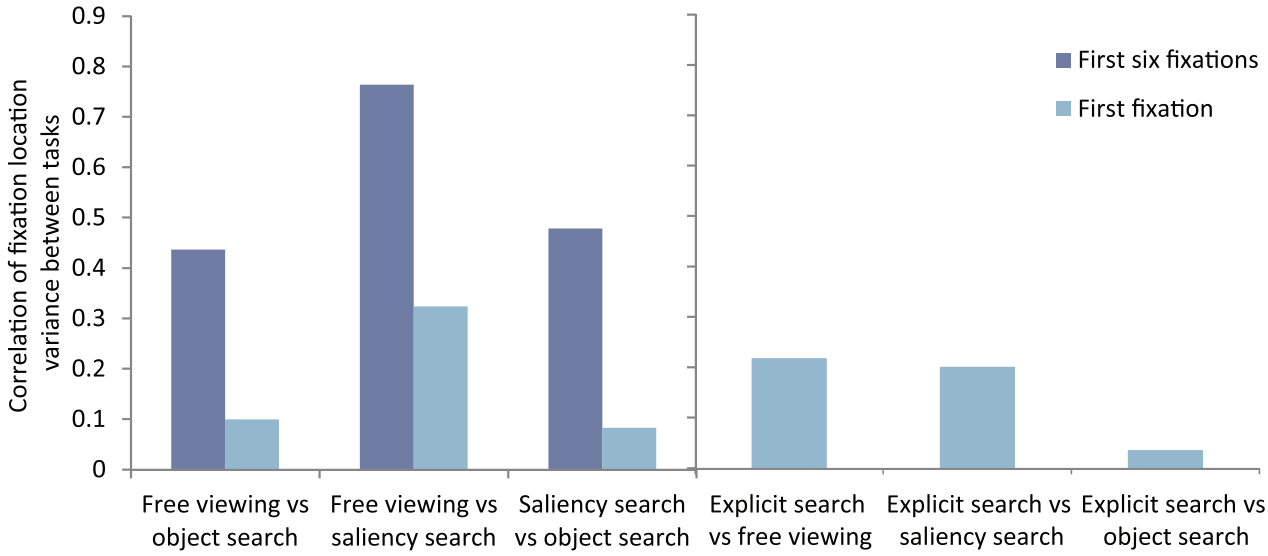

Figure 15. Correlation coefficients of observer variability of fixation location across images for different task combinations for eye movement tasks (left) and the explicit judgment task (right). A high correlation suggests that images that tend to have high observer variability for one task (free viewing) will also lead to high variability for another task (object search).

Relationship across metrics

We investigated the relationship of the metrics for each of the tasks. We computed the correlation for each model between the various metrics (ROI, ROC, distance) for the different tasks. A high correlation would indicate that images that are well predicted by the model with one metric would also be well predicted using a second metric. Figure 16 shows an example of the correlation between ROI and ROC metric results, $r(798) = 0.60$, $p < 0.001$, as well as the ROI and distance metric results, $r(798) = -0.91$, $p < 0.001$, for the IK model's predictions for the explicit judgment task. The results show that not surprisingly the ROI and distance metrics are highly correlated but the correlation between the ROI and the ROC metric is lower. Figure 17 shows the correlations across metrics for each task averaged across the three models (all correlations are significant, $p < 0.001$). The correlations confirm the trend seen in Figure 16, showing that the ROI and distance metrics are highly correlated for all tasks and that the correlation of these two metrics with the ROC metric is lower.

Discussion

Our main aim was to evaluate the ability of models of saliency to predict human behavior across different tasks using a similar image data set. Prior to discussing how saliency models predicted behavior across tasks, we briefly discuss comparisons across models of saliency (but see Borji et al., 2012, for a thorough assessment of 35 models).

Comparison across saliency models

According to all three utilized metrics, the AIM model (Bruce & Tsotsos, 2006) was best able to predict human behavior (judgments and eye movements) across all tasks (see Figures 4, 6, and 8). The different metric analyses resulted in a difference as to which model predicted human behavior second and third best. Overall, our results suggest that spatially global feature processing and statistical departure from a collection of images are more effective determinants in performance for the tasks used in this project. However, it is important to note that our implementation of the IK model was not well-suited for ROC analysis, as can be seen in Figure 8. Even when the toolbox was implemented with images at half-resolution, the resulting saliency maps were very sparse, making it unlikely that values away from the origin and top-right corner of the ROC curve would be populated. This underscores the importance of evaluating models with different metrics. ROI interaction results shown in Figure 4 suggest that the type of task largely affected the IK model. Specifically, the IK model was best at predicting explicit saliency judgments like AIM and SUN, but showed a greater decrease in performance than AIM and SUN in predicting eye movements for all eye-tracked tasks.

Analysis using the ROI metric of the top five saliency regions demonstrates an ordinal relationship between these salient regions and the ability of the models to

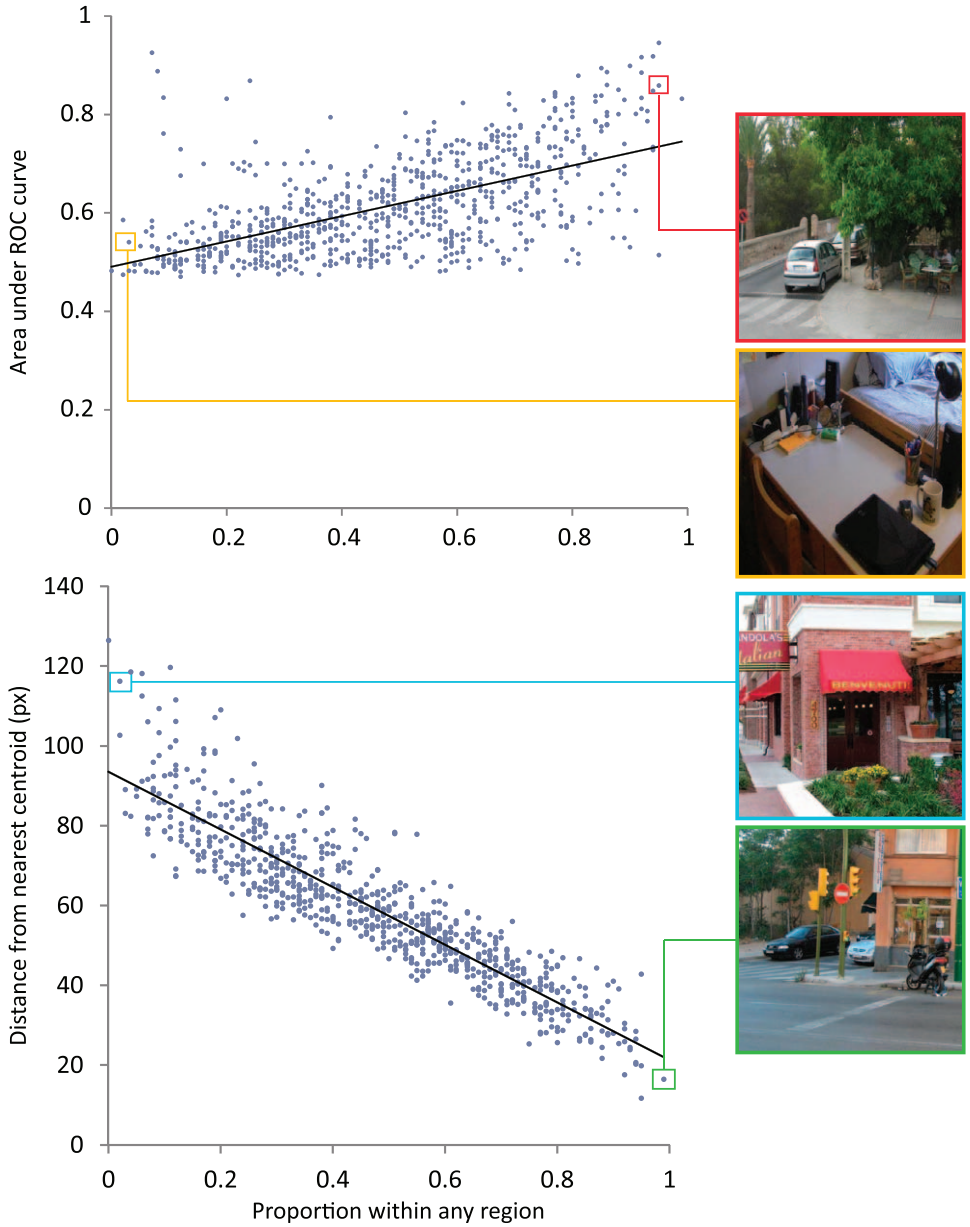

Figure 16. Scatterplots showing the correlation between the behavioral predictions of the IK model of selections made during the explicit judgments task according to the ROI versus ROC metric (top) and ROI versus distance metric (bottom). Images corresponding to points in the scatterplot are shown to illustrate examples of instances when the models perform well across metrics (red and green) and poorly across metrics (yellow and blue).

predict selections and eye movements (see Figures 5 and 7). For all models and among the five most salient regions, the top salient region had the highest accuracy at predicting human selections/eye movements. The accuracy progressively decreased from top-most to fifth-most salient region for all models.

What do saliency models predict?

Our overall results suggest that all saliency models were more accurate at predicting human judgments of explicit saliency than eye movements in the free viewing task. A clear trend can also be seen in model predictions for the eye movements in the saliency search versus free viewing and object search tasks, such that performance is better for the saliency search task (at least one metric showed this result, and another was marginally significant in each case). The ROC results also confirm that eye movements executed during the free viewing condition were different than those executed during object search tasks (Buswell, 1935; Underwood & Foulsham, 2006).

A typical underlying assumption in the literature is that when an observer is not engaged in a specified task

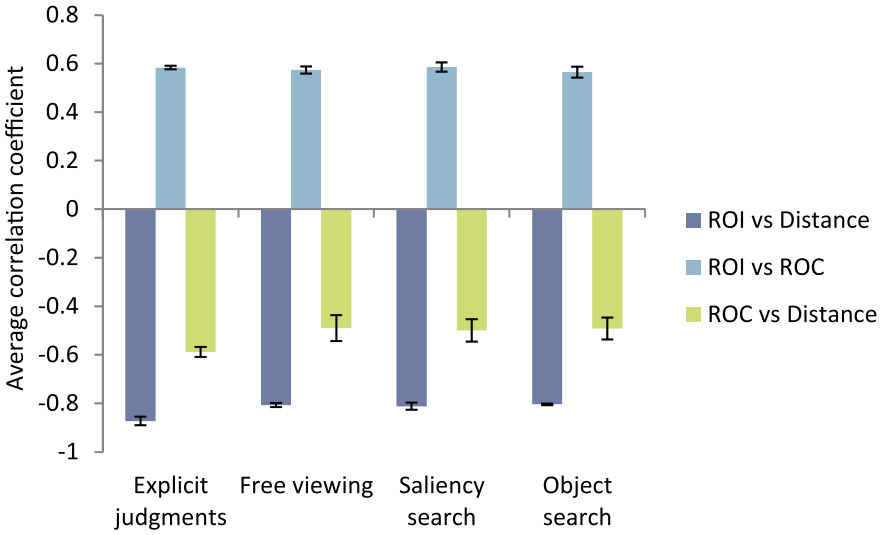

Figure 17. Average correlation coefficient taken across models for each metric pairing and each task. Error bars represent *SEM*.

his or her eye movements will be directed to areas of high saliency, i.e., prominent attention captors (Foulsham & Underwood, 2008; Parkhurst et al., 2002). Recently, however, the situations in which solely bottom-up saliency models meaningfully supply predictions have come under scrutiny. Critics advocate incorporating top-down processes into the calculation of the saliency or activation map (Beutter et al., 2003; Oliva, Torralba, Castelhano, & Henderson, 2003; Rao, Zelinsky, Hayhoe, & Ballard, 2002), or predicting fixations on the basis of reward maximization and uncertainty reduction (Najemnik & Geisler, 2005; Renninger, Coughlan, Verghese, & Malik, 2005; Tatler et al., 2011). The top-down component is consistent with a large body of behavioral studies showing that eye movements during search are guided by target-relevant features (Eckstein, Beutter, Pham, Shimozaki, & Stone, 2007; Findlay, 1997; Malcolm & Henderson, 2009, 2010; Tavassoli, van der Linde, Bovik, & Cormack, 2009) and scene context (Chen & Zelinsky, 2006; Eckstein, Drescher, & Shimozaki, 2006; Torralba et al., 2006).

Many of the models of eye movements during search are based on top-down components (Beutter et al., 2003; Najemnik & Geisler, 2005; Rao et al., 2002; Renninger et al., 2005; Zelinsky et al., 2005). Other models implement modifications of saliency models to include top-down components to more accurately predict gaze allocation (Navalpakkam & Itti, 2005; Oliva et al., 2003; Torralba et al., 2006). The need to include top-down processes in such models has also been interpreted to suggest that free viewing may not actually be task independent, considering that top-down processes are, by definition, guided by some sort of cognitive goal or task (Tatler et al., 2005; Tatler et al., 2011).

Those advocating that free viewing is well-described by bottom-up saliency models argue that when observers have no task in mind a default strategy by the visual system is to fixate on salient regions. One argument is that in the absence of a specific task, if one considers all possible tasks that a human might engage in after examining an image, fixating salient objects might be a functionally adaptive strategy given that objects tend to be salient (Eckstein, 2011).

Our results fall in between these two theoretical positions. First, the finding that saliency models were better able to predict behavior during saliency tasks (explicit judgments and eye movements) than free viewing eye movements can be interpreted to suggest that, during free viewing, observers are engaging (at least in a subset of the trials) in some goal directed task not involving fixating the most salient regions. These goal-directed tasks might be default tasks that are highly practiced and important, such as fixating faces (Cerf, Frady, & Koch, 2009), animals (Yun, Peng, Samaras, Zelinsky, & Berg, 2013), or text (Wang & Pomplun, 2012), even if these are not the most salient regions in the scene.

On the other hand, our results also show commonalities between the saliency search and the free viewing tasks (see Figures 10 and 11). Results from the ROI metric did not support a difference between the two tasks. Furthermore, variations in the ability of saliency models to predict human fixations across images were correlated more highly for the saliency search and free viewing tasks than with the object search task. Similarly, the fixational variance across observers varied in a correlated way across images for the free viewing and saliency search task. These results may suggest that free viewing is to some extent akin to moving the eyes to salient locations, or more specifically, among the four tasks tested, free viewing is most

similar to moving the eyes to salient locations. This lends credence to the use of saliency models as predictors of free eye movements; however, it must still be noted that their predictions of behavior on free viewing tasks is subpar as compared to saliency search and explicit judgment tasks (per the distance and ROC metrics). Another possible alternative explanation for the relationship between saliency search and free viewing tasks is that they are simply both mediated by a common strategy of fixating objects in the scenes (Einhäuser, Spain et al., 2008; Nuthmann & Henderson, 2010).

Variability of fixation patterns across observers

Analyses of the variability of fixations across observers, shown in Figure 11, revealed lower KL divergence and variance for the object search task and higher KL divergence and variance for the free viewing and saliency search task. The results are consistent with the idea that eye movements across observers are more similar when they are instructed to engage in a specific task such as searching for an object rather than a more open ended task such as free viewing.

Perhaps surprising is that the interobserver fixation variance is similar for the free viewing and saliency search task. We might have expected the saliency search task to lead to lower interobserver fixation variance given that observers are instructed with a top-down task. The results could be interpreted to reflect the commonalities between the saliency search task and the free viewing task. The finding could also suggest that there is a large degree of variability in what observers consider a salient image region and that even with the specific instructions observers might engage in search for saliency using somewhat dissimilar features. Comparison to the explicit saliency judgment provides a way to assess the observer variability in eye movements compared to the inherent variability in what observers consider the most salient region of an image. The variance and KL divergence for observers' explicit saliency judgment is larger than that of observers' first few eye movements in the saliency search and free viewing task. Around the third fixation, eye movements in the saliency search and free viewing task show comparable and larger variability than explicit saliency judgments.

The additional variability in later eye movements in the saliency search task when compared to explicit judgments might reflect that observers have completed the task and are engaging in idiosyncratic diverse tasks or simply reflect additional variability due to uncertainty in the spatial precision of fixational eye movements (Kowler & Blaser, 1995) when compared to mouse selections with longer times.

The finding that interobserver variability progressively increases with fixation number is consistent with a previous study (Tatler et al., 2005). A possible explanation for the lower variability of early fixations is that these are task-related and common to all observers. In contrast, later saccades might be executed after completion of the instructed task and reflect secondary idiosyncratic tasks engaged by each observer. Although this is a likely explanation for the object search task it is less likely for the saliency search and free viewing tasks. If the observer variability in the saliency and free viewing tasks was driven by the task to find the most salient object then we would expect the early fixation variability to be similar to that of the explicit saliency judgment. Yet, the observer variability in early fixations for the saliency search and free eye movement tasks was significantly lower than the explicit judgment. We speculate that the lower variability in these early fixations of the two tasks might be related to the bias to fixate initially towards the center of the images (center bias; Tatler, 2007). It should be noted that the viewing time in this experiment was relatively short (2 s) compared to other studies (e.g., Bruce & Tsotsos, 2006, 4 s; Tatler et al., 2005, 1–10 s), therefore the pattern of divergence in eye movement behavior between individuals may not be fully captured.

Effect of image content on human and model behavior

As was shown in Figures 14 and 16, the content of images can drastically affect the variability of fixations and ability of models to predict those fixations. Generally, based on qualitative visual inspection of the images, cluttered images produced more differences between observer fixation locations and poorer model prediction performance across all metrics. Images with few objects, or with stand-out objects or regions produced uniform fixation distributions across observers and better prediction performance across all metrics. These results underscore the importance of testing models against image databases that contain a variety of image types. Cluttered images will be inherently more difficult for models because there is a high degree of observer variability that needs to be accounted for with such images. A simpler benchmark would be to use images that result in similar eye movements among observers, as many do, but this will preclude an assessment of whether a model can account for the full range of human behavior.

Comparison of metrics

There has also been some debate regarding the most accurate metric to quantify various models' predictive

power. Researchers have utilized a variety of metrics, such as chance adjusted saliency (Parkhurst et al., 2002), area under an ROC curve (Tatler et al., 2005), KL-divergence (L. Zhang et al., 2008), and target detection (Itti et al., 1998) to quantify model ability to predict human eye movements. Our results indicate that patterns of results differ slightly depending on which metric is used. There are also some subtle differences depending on what type of task human observers were completing. If the differences in mean performance on task by model are compared across metrics, further insight can be gained as to the weaknesses and strengths of each metric. The most prominent difference in performance across metric was seen in the case of the IK model results according to the ROC metric. The ROC performance was computed to be quite low because the output saliency maps are sparse and close to binary. This drastically limits the number of unique points that can be plotted on the ROC curve, resulting in low areas under the curve. This serves as an excellent example to consider the nature of a models' output when selecting a metric by which to assess it. Similarly, models that only output maps do not lend themselves well to being compared to object-selection databases.

Conclusions

The main contribution of this work is a detailed analysis of the types of tasks for which human behavior is best predicted by saliency models. We show that three state-of-the-art saliency models were best able to predict explicit judgments of saliency and less able to predict human eye movements on other more commonly tested tasks: free eye movements, saliency search, and object search. The results support the notion that the fixations during free viewing tasks are not as well predicted by saliency models as explicit saliency judgments. However, our results suggest that the free viewing and saliency search eye movements share some commonalities when compared to an object search task. In terms of the observer variability in fixational eye movements, our results suggest that variability is modulated by the degree to which observers are performing the same task, the variability across observers in the underlying representation of most salient regions, and the specific content of the each image. Finally, the database of observer responses used in this project should be useful for future projects, especially since it consists of images depicting various numbers of salient (i.e., foreground) objects, and can be easily compared to models that output saliency maps.

Keywords: saliency, attention, eye movements, visual search, real scenes

Acknowledgments

We would like to thank Steve Mack, Chad Carlson, Joanna Williams, Vatche Baboyan, Stella Keynigstheyn, and Shantal Ben-Aderet for their assistance with data collection. In addition, we thank Ben Tatler and an anonymous reviewer for their helpful comments and recommendations during the review process. Support for this research was provided by the National Institute of Health (R21 EY023097), National Science Foundation (NSF-0819582), and the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Portions of this work were previously presented at the Vision Sciences Society Annual Meeting (Koehler et al., 2011).

Commercial relationships: none.
Corresponding author: Kathryn L. Koehler.
Email: kathryn.koehler@psych.ucsb.edu.
Address: Vision and Image Understanding Laboratory, Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, USA.

Footnotes

[1]This coordinate output is unlike the database of human selected rectangular enclosures of salient objects constructed by Liu et al. (2007) because the explicit judgment of saliency in the present paper could correspond to nonobjects and/or parts of objects.

[2]The proportion of clicks within all regions reported from the random control condition varies by model. It does not match the proportion of total image area subtended by all regions (0.28). This is because, as mentioned before, some of the region boundaries extend past the edge of the images. The average proportion of the total image area subtended by all regions was 0.25.

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition,* Miami, FL, 2009 (pp. 1597–1604).

Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using

stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(4), 693–708.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, *9*(5):19, 1–15, http://www.journalofvision.org/content/9/5/19, doi:10.1167/9.5.19. [PubMed] [Article]

Beutter, B. R., Eckstein, M. P., & Stone, L. S. (2003). Saccadic and perceptual performance in visual search tasks. I. Contrast detection and discrimination. *Journal of the Optical Society of America, A: Optics, Image Science, & Vision*, *20*(7), 1341–1355.

Bian, P., & Zhang, L. (2009). Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *Advances in Neuro-Information Processing* (pp. 251–258). Berlin: Springer.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*(24), 2992–3000.

Bonev, B., Chuang, L. L., & Escolano, F. (2013). How do image complexity, task demands and looking biases influence human gaze behavior? *Pattern Recognition Letters*, *34*(7), 723–730, doi:10.1016/j.patrec.2012.05.007.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6180177.

Borji, A., Sihite, D. N., & Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6253254.

Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, *91*, 62–77, doi:10.1016/j.visres.2013.07.016.

Bruce, N. (2006). Retrieved November 2011. from http://www-sop.inria.fr/members/Neil.Bruce/#SOURCECODE

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems*, *18*, 155.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://www.journalofvision.org/contents/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article]

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press. Retrieved from http://psych.wfu.edu/art_schirillo/articles/Buswell,%201935.pdf

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12):10, 1–15, http://www.journalofvision.org/contents/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems*, *20*. Cambridge, MA: MIT Press.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*(24), 4118–4133, doi:10.1016/j.visres.2006.08.008.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*(3), 837–845.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5):14, 1–36, http://www.journalofvision.org/content/11/5/14, doi:10.1167/11.5.14. [PubMed] [Article]

Eckstein, M. P., Beutter, B. R., Pham, B. T., Shimozaki, S. S., & Stone, L. S. (2007). Similar neural representations of the target for saccades and perception during search. *Journal of Neuroscience*, *27*(6), 1266–1270, doi:10.1523/JNEUROSCI.3975-06.2007.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973–980, doi:10.1111/j.1467-9280.2006.01815.x.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2):2, 1–19, http://www.journalofvision.org/content/8/2/2, doi:10.1167/8.2.2.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1–26, http://www.journalofvision.org/content/8/14/18, doi:10.1167/8.14.18. [PubMed] [Article]

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, *13*(4):11, 1–20, http://www.journalofvision.org/content/13/4/11, doi:10.1167/13.4.11. [PubMed] [Article]

Fang, Y., Chen, Z., Lin, W., & Lin, C.-W. (2012). Saliency detection in the compressed domain for

adaptive image retargeting. *IEEE Transactions on Image Processing*, *21*(9), 3888–3901.

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research*, *37*(5), 617–631.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2):6, 1–17, http://www.journalofvision.org/content/8/2/6, doi:10.1167/8.2.6. [PubMed] [Article]

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, *8*(7):13, 1–18, http://www.journalofvision.org/content/8/7/13, doi:10.1167/8.7.13. [PubMed] [Article]

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image & Vision Computing*, *30*(1), 51–64.

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, *12*(6):17, 1–22, http://www.journalofvision.org/content/12/6/17, doi:10.1167/12.6.17. [PubMed] [Article]

Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1915–1926.

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, *19*(1), 185–198.

Harding, G., & Bloj, M. (2010). Real and predicted influence of image manipulations on eye movements during scene recognition. *Journal of Vision*, *10*(2):8, 1–17, http://www.journalofvision.org/content/10/2/8, doi:10.1167/10.2.8. [PubMed] [Article]

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, *19*, 545.

Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *34*(1), 194–201.

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition,* 17–22 June, 2007 (pp. 1–8).

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems* (pp. 681–688). Cambridge, MA: MIT Press.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304–1318.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 631–637). Cambridge, MA: MIT Press.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, *18*, 547.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506, doi:10.1016/S0042-6989(99)00163-7.

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*(11), 1254–1259.

Jacobson, N., & Nguyen, T. (2011). Video processing with scale-aware saliency: Application to frame rate up-conversion. *Acoustics, Speech and Signal Processing (ICASSP)*, 2011 IEEE International Conference, 22–27 May, 2011, doi:10.1109/ICASSP.2011.5946653.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision,* September 29, 2009 (pp. 2106–2113).

Jung, C., & Kim, C. (2012). A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. *IEEE Transactions on Image Processing*, *21*(3), 1272–1283.

Kim, C., & Milanfar, P. (2013). Visual saliency in noisy images. *Journal of Vision*, *13*(4):5, 1–14, http://www.journalofvision.org/content/13/4/5, doi:10.1167/13.4.5. [PubMed] [Article]

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–27.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. (2011). Assessing models of visual saliency against explicit saliency judgments from one hundred humans viewing eight hundred real scenes. *Journal of Vision*, *11*(11):165, http://www.journalofvision.org/content/11/11/165, doi:10.1167/11.11.165. [Abstract]

Kollmorgen, S., Nortmann, N., Schröder, S., & König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology*, *6*(5), e1000791, doi:10.1371/journal.pcbi.1000791.

Kootstra, G., & Schomaker, L. R. (2009). Prediction of human eye fixations using symmetry. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 56–61). Retrieved from http://csjarchive.cogsci.rpi.edu/proceedings/2009/index.html.

Kowler, E., & Blaser, E. (1995). The accuracy and precision of saccades to small and large targets. *Vision Research*, *35*(12), 1741–1754.

Lang, C., Liu, G., Yu, J., & Yan, S. (2012). Saliency detection by multitask sparsity pursuit. *IEEE Transactions on Image Processing*, *21*(3), 1327–1338.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, *47*(19), 2483–2498.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817.

Le Meur, O., Thoreau, D., Le Callet, P., & Barba, D. (2005). A spatio-temporal model of the selective human visual attention. In *IEEE International Conference on Image Processing,* 11–14 September, 2005 (Vol. *3*, pp. III–1188).

Lee, W.-F., Huang, T.-H., Yeh, S.-L., & Chen, H. H. (2011). Learning-based prediction of visual attention for video signals. *IEEE Transactions on Image Processing*, *20*(11), 3028–3038.

Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(4), 996–1010. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6243147.

Li, J., Tian, Y., Huang, T., & Gao, W. (2009). A dataset and evaluation methodology for visual saliency in video. In *IEEE International Conference on Multimedia and Expo,* 28 June–3 July, 2009 (pp. 442–445).

Li, J., Tian, Y., Huang, T., & Gao, W. (2010). Probabilistic multi-task learning for visual saliency estimation in video. *International Journal of Computer Vision*, *90*(2), 150–165.

Li, J., Xu, D., & Gao, W. (2011). Removing label ambiguity in learning-based visual saliency estimation. *IEEE Transactions on Image Processing, 21*(4), 1513–1525.

Li, Y., Zhou, Y., Yan, J., Niu, Z., and Yang, J. (2010). Visual saliency based on conditional entropy. In *Computer Vision–ACCV,* 2009 (pp. 246–257). Berlin Hiedelberg: Springer.

Lin, Y., Tang, Y., Fang, B., Shang, Z., Huang, Y., & Wang, S. (2013). A visual-attention model using earth mover's distance based saliency measurement and nonlinear feature combination. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(2), 314–328. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6205759.

Liu, T., Sun, J., Zheng, N. N., Tang, X., & Shum, H. Y. (2007). Learning to detect a salient object. In *IEEE Conference on Computer Vision and Pattern Recognition, 33*(2), 353–367.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: evidence from eye movements. *Journal of Vision*, *9*(11):8, 1–13, http://www.journalofvision.org/content/9/11/8, doi:10.1167/9.11.8. [PubMed] [Article]

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, *10*(2):4, 1–11, http://www.journalofvision.org/content/10/2/4, doi:10.1167/10.2.4. [PubMed] [Article]

Marat, S., Phuoc, T. H., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, *82*(3), 231–243.

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

Murray, N., Vanrell, M., Otazu, X., & Parraga, C. A. (2011). Saliency estimation using a non-parametric low-level vision model. In *IEEE Conference on Computer Vision and Pattern Recognition,* 20–25 June, 2011 (pp. 433–440).

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205–231.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8):20, 1–19, http://www.journalofvision.org/content/10/8/20, doi:10.1167/10.8.20. [PubMed] [Article]

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Proceedings of the International Conference on Image Processing, 1*, 253–256.

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision*, *13*(5):2, 1–21, http://www.journalofvision.org/content/13/5/2, doi:10.1167/13.5.2. [PubMed] [Article]

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123, doi:10.1016/S0042-6989(01)00250-4.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T. S. (2010). An eye fixation database for saliency detection in images. In K. Daniilidis, P. Maragos, & N. Paragois (Eds.), *Computer Vision–ECCV* (pp. 30–43). Berlin Heidelberg: Springer.

Rao, R. P. N., Hayhoe, M. M., Zelinsky, G. J., & Ballard, D. H. (1996). Modeling saccadic targeting in visual search. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8 (NIPS*95)*. Cambridge, MA: MIT Press.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*(11), 1447–1463.

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, *17*, 1121–1128.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14):16, 1–20, http://www.journalofvision.org/content/7/14/16, doi:10.1167/7.14.16. [PubMed] [Article]

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1–3), 157–173.

Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, *9*(12):15, 1–27, http://www.journalofvision.org/content/9/12/15, doi:10.1167/9.12.15. [PubMed] [Article]

Smith, T. J., & Mital, P. K. (2011). Watching the world go by: Attentional prioritization of social motion during dynamic scene viewing. *Journal of Vision*, *11*(11):478, http://www.journalofvision.org/content/11/11/478, doi:10.1167/11.11.478. [Abstract]

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [PubMed] [Article]

Tavakoli, H. R., Rahtu, E., and Heikkilä, J. (2011). Fast and efficient saliency detection using sparse sampling and kernel density estimation. In A. Heyden & F. Kahl (Eds.), *Image Analysis* (pp. 666–675). Berlin Heidelberg: Springer.

Tavassoli, A., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2009). Eye movements selective for spatial frequency and orientation during active visual search. *Vision Research*, *49*(2), 173–181, doi:10.1016/j.visres.2008.10.005.

Toet, A. (2011). Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2131–2146.

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*(4), 766.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, *59*(11), 1931–1949, doi:10.1080/17470210500416342.

Vazquez, E., Gevers, T., Lucassen, M., van de Weijer, J., & Baldrich, R. (2010). Saliency of color image derivatives: A comparison between computational models and human perception. *JOSA A*, *27*(3), 613–621.

Vig, E., Dorr, M., Martinetz, T., & Barth, E. (2012). Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(6), 1080–1091.

Walther, D. (2006). *The saliency toolbox*. Retrieved November 2011 from http://www.saliencytoolbox.net

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Wang, H.-C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, *12*(6):26, 1–17, http://www.journalofvision.org/content/12/6/26, doi:10.1167/12.6.26. [PubMed] [Article]

Xie, Y., Lu, H., & Yang, M. (2012). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing, 22*(5), 1689–1698. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6291786.

Yan, J., Liu, J., Li, Y., Niu, Z., & Liu, Y. (2010). Visual saliency detection via rank-sparsity decomposition. In *IEEE International Conference on Image Processing,* 26–29 September, 2010 (pp. 1089–1092).

Yun, K., Peng, Y., Samaras, D., Zelinsky, G. J., & Berg, T. (2013). Studying relationships between human gaze, description, and computer vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 23–28 June, 2013 (pp. 739–746).

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2005). The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems* (pp. 1569–1576) Cambridge, MA: MIT Press.

Zhang, L. (2008). Retrieved November, 2011 from http://cseweb.ucsd.edu/~l6zhang/.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article]

Zhang, S., & Eckstein, M. P. (2010). Evolution and optimality of similar neural mechanisms for perception and action during search. *PLoS Computational Biology*, *6*(9), doi:10.1371/journal.pcbi.1000930.

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [PubMed] [Article]

Zhao, Q., & Koch, C. (2012). Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *Journal of Vision*, *12*(6):22, 1–15, http://www.journalofvision.org/content/12/6/22, doi:10.1167/12.6.22. [PubMed] [Article]

Appendix

Papers that compared saliency models to human behavior

This list was compiled by performing a keyword search for salienc* in *Journal of Vision*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PNAS*, *Vision Research*, and *Journal of the Optical Society of America*. Additionally, we included any papers that were tested in the extensive comparison of 35 saliency models to various human datasets performed by Borji et al. (2012). Finally, we also included any papers that contained the databases used in papers from the list above and a comparison to some type of saliency model. Models or data sets taken from dissertations were not included. Seventeen unique data sets are used in the free viewing comparisons, seven unique sets in the task-based comparisons, eight unique sets in the object selections, and three unique sets in the other category. The list of citations below is divided by task type as in Figure 1. Some papers used multiple data sets and therefore compared a saliency model to more than one task type, in which case they are listed in more than one task section below.

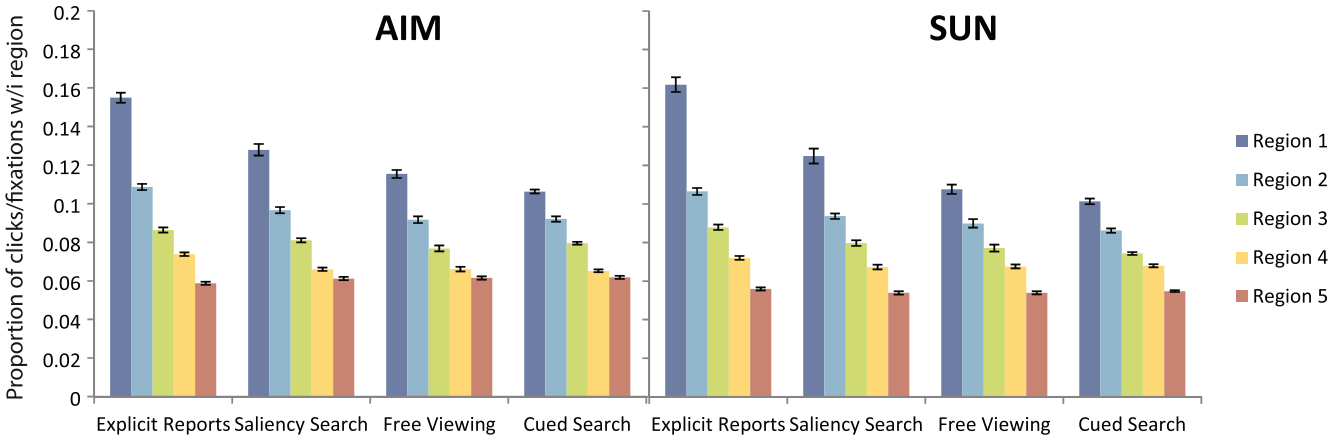

Figure A1. ROI results broken down by region for the AIM and SUN models, respectively. Each bar represents the proportion of clicks or fixations within that specific region, with regions ranked in order from most to fifth-most salient. Error bars represent *SEM*.

Figure A2. Distance results broken down by region for the AIM and SUN models. Each bar represents the average distance of selections or fixations closest to the specified region, with regions ranked in order from most to fifth-most salient. Error bars represent *SEM*.

Free viewing (42 papers)

(Berg, Boehnke, Marino, Munoz, & Itti, 2009; Bian & Zhang, 2009; Birmingham et al., 2009; Borji et al., 2012; Bruce & Tsotsos, 2009; Cerf, Harel, Einhäuser, & Koch, 2008; Cerf et al., 2009; Erdem & Erdem, 2013; Gao, Mahadevan, & Vasconcelos, 2008; Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012; Garcia-Diaz, Leborán, Fdez-Vidal, & Pardo, 2012; Goferman, Zelnik-Manor, & Tal, 2010; Guo & Zhang, 2010; Harel, Koch, & Perona, 2007; Hou, Harel, & Koch, 2012; Hou & Zhang, 2008; L. Itti & Baldi, 2005, 2006; Jacobson & Nguyen, 2011; Kim & Milanfar, 2013; Kootstra & Schomaker, 2009; Lang, Liu, Yu, & Yan, 2012; Le Meur, Le Callet, & Barba, 2007; Le Meur, Le Callet, Barba, & Thoreau, 2006; Le Meur, Thoreau, Le Callet, & Barba, 2005; Lee, Huang, Yeh, & Chen, 2011; J. Li, Tian, Huang, & Gao, 2010; J. Li, Levine, An, Xu, & He, 2013; Y. Li, Zhou, Yan, Niu, & Yang, 2010; Lin et al., 2013; Marat et al., 2009; Murray, Vanrell, Otazu, & Parraga, 2011; Parkhurst et al., 2002; Peters et al., 2005; Ramanathan, Katti, Sebe, Kankanhalli, & Chua, 2010; Seo & Milanfar, 2009; Tavakoli, Rahtu, & Heikkilä, 2011; Vig, Dorr, Martinetz, & Barth, 2012; Yan, Liu, Li, Niu, & Liu, 2010; L. Zhang et al., 2008; Zhao & Koch, 2011, 2012)

Object selections (14 papers)

(Achanta et al., 2009; Avraham & Lindenbaum, 2010; Fang, Chen, Lin, & Lin, 2012; Hou & Zhang, 2007; Itti & Koch, 2000; Jung & Kim, 2012; Lang et al., 2012; J. Li, Xu, & Gao, 2011; J. Li, Tian, Huang, & Gao, 2009; Y. Li et al., 2010; J. Li et al., 2013; Liu et al., 2007; Vazquez, Gevers, Lucassen, van de Weijer, & Baldrich, 2010; Xie, Lu, & Yang, 2012)

Task-based viewing (15 papers)

(Birmingham et al., 2009; Borji et al., 2012; Erdem & Erdem, 2013; Foulsham & Underwood, 2008; Goferman et al., 2010; Harding & Bloj, 2010; Itti, 2004; Judd, Ehinger, Durand, & Torralba, 2009; Lang et al., 2012; J. Li et al., 2011; Murray et al., 2011; Rothkopf, Ballard, & Hayhoe, 2007; Tatler et al., 2005; Zhao & Koch, 2011, 2012)

Other (3 papers)

(Masciocchi, Mihalas, Parkhurst, & Niebur, 2009; Toet, 2011; Vazquez et al., 2010)