Title

The Visual System Prioritizes High-Level Scene Properties for Attentional Selection

Author

Peacock, Candace Elise

Publication Date

2021

The Visual System Prioritizes High-Level Scene Properties for Attentional Selection

By

CANDACE ELISE PEACOCK
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

John M. Henderson, Chair

Joy J. Geng

Steven J. Luck

Committee in Charge

2021

Acknowledgements

Research

I would first like to thank my mentor, John Henderson. I am grateful that he took the care and time to develop me as a scientist, for his encouragement to pursue difficult but worthwhile projects, and for teaching me to trust my scientific intuition. I would not be where I am today without his mentorship. I would also like to thank my committee members, Steve Luck and Joy Geng, for helping me navigate graduate school and for encouraging me to be the best scientist I can be. I would also like to extend my thanks to Deb Cronin, Jessi Goold, and Gwen Rehrig, who are the best collaborators and friends I could ask for. They made Davis an encouraging environment and I appreciate their research and practical insights. I would also like to acknowledge Taylor Hayes for teaching me that a simple, intuitive analysis is the best analysis. Finally, I would like to thank Marian Berryhill for preparing me for graduate school. I would not be here without her wisdom and guidance.

Life

I am grateful to my partner, Rio, for his steadfast love, support, intelligence, and humor during my PhD. His strength and passion for life and knowledge inspires me. Thanks for always encouraging me to persevere. I am thankful to my mom for always reminding me of how interconnected all the human systems are, to my dad whose intellectual curiosity for the world inspires me, and to my sisters, Alyssa and Brianna, for always reminding me that things will work out. I am thankful for Rio’s family for their kindness, for nourishing my soul with nature, and for teaching me that my best is enough. To all the rest of my friends (Riley, Kevin, Kate, and beyond): thanks for being constant reminders of how lucky I am to have friends like you.

Table of Contents

Abstract

As we sample the world via shifts in gaze, the visual system filters out irrelevant information to prioritize the most relevant visual input. However, there is debate regarding why one source of information is selected over another for attention. The literature has suggested that the physical salience of image features is a dominant guidance factor in scene perception. However, cognitive relevance theory suggests that it is actually scene meaning (our knowledge of the world) that guides attention. Because meaning and image salience are correlated but have been represented differently in the literature, however, it has previously been impossible to test whether meaning or image salience uniquely predict attention when they are represented in the same format. To test their unique contributions to attention, Chapters 2 and 3 tested whether attention, as operationalized by fixation densities, was more related to meaning maps, which capture the spatial distribution of semantic densities in real-world scenes, or to saliency maps, which capture the spatial distribution of physically conspicuous features in scenes. Chapter 2 used a task in which viewers were instructed to count bright patches in scenes or rate the overall brightness of scenes while their eye movements were recorded. This resulted in image salience being task-relevant and meaning being task-irrelevant. Despite its task-irrelevance, meaning predicted fixation densities uniquely whereas image salience did not. A caveat of Chapter 2, however, is that the task required that eye movements be directed to scene-dependent information, thereby conflating whether the task was truly meaning-independent. To remedy this, Chapter 3 employed a free viewing task that did not require participants to attend to meaning or salience. Here, it was found that even during free viewing, meaning continued to explain the overall and unique patterns of attention significantly better than image salience. Together, these findings suggest that the visual system selects meaningful information for attentional selection, as consistent with cognitive relevance theory. Finally, prior work has combined spatial constraint (knowledge of where objects are located in scenes) and image salience to predict where fixations are directed during visual search. Given that meaning uniquely predicts attention beyond image salience, however, Chapter 4 therefore tested whether combining spatial constraint and meaning also predicts eye movements during visual search. Here, meaning was represented as meaning maps and spatial constraint was represented as surface maps that represented the likely locations of target objects as continuous probabilities. The results showed that combining spatial constraint and meaning predicted eye movements better than spatial constraint or meaning alone. This suggested that the visual system selects meaningful regions that appear on surfaces related to visual search targets for fixation. These findings collectively demonstrate that the human visual system prioritizes scene regions that contain meaningful content based upon our knowledge of the world for attention. This has implications for cognitive relevance theory which describes how humans orient attention in the real world and may help inform technologies that reduce distractions.

Chapter 1: Introduction

As we sample information in the world via saccadic eye movements, the visual system must select and prioritize only the most relevant information for analysis at any one time. However, there is debate regarding why the visual system selects one source of information over another for attention. Understanding what regions of the world the visual system prioritizes and why it prioritizes them may help inform technologies that can reduce distraction. For instance, assisted driving might be able to identify and warn drivers about easy-to-miss hazards and virtual learning/work interfaces might better be able to emphasize important information and de-emphasize less important information using the results of this research. This dissertation will explore the influences of different scene properties on the guidance of visual attention in real-world scenes.

Scene Meaning and Image Salience

Image salience is defined as a physical property of a stimulus in which a region of low-level visual features (colors, intensities, orientations) is sufficiently different from its surroundings, potentially resulting in that region 'popping out' for attention (Itti & Koch, 2001; Treisman & Gelade, 1980; Wolfe et al., 1989). Maps of image saliency ('saliency maps') measure the spatial distribution of physically conspicuous regions in real-world images (Harel et al., 2006; Itti & Koch, 2001; Koch & Ullman, 1987). Image salience is thought to be a dominant factor guiding visual attention in real-world scenes (Anderson et al., 2015; Borji et al., 2013, 2014; Harel et al., 2006; Itti et al., 1998; Itti & Koch, 2001; Koch & Ullman, 1987; Parkhurst et al., 2002) and has been suggested to be behaviorally important because biological systems need to quickly detect threats in their environment (Itti & Koch, 2001).

The predictions made by saliency maps are bottom-up in nature—stimulus features, rather than higher-order cognitive processes, are the predicted drivers of attention. However, information that is relevant to the cognitive system, such as scene meaning (Antes, 1974; Castelhano & Henderson, 2007; Henderson, 2017; Mackworth & Morandi, 1967) and task (Ehinger et al., 2009; Einhäuser et al., 2008; Torralba et al., 2006), may influence attention above and beyond salience. Scene meaning is a guidance factor that describes what scene regions the cognitive system predicts will be informative (Mackworth & Morandi, 1967) or recognizable (Antes, 1974) based upon our prior knowledge of a scene's semantic content. Cognitive relevance theory suggests attention is directed (or "pushed") to information that is relevant to the cognitive system (e.g., meaningful scene regions) over information that is irrelevant (e.g., uninterpreted image features) (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003, 2017; Henderson et al., 2009; Henderson & Hollingworth, 1999; Tatler et al., 2011; Yarbus, 1967). This stands at odds with image salience theory which suggests that attention is "pulled" to visually salient scene regions without regard to our knowledge of a scene. It could be the case, then, that despite the emphasis on image salience in the literature, it is actually scene meaning that guides attention rather than image salience.

Nevertheless, the literature has typically modeled meaning (i.e., manipulating small image regions; Brockmole & Henderson, 2008; De Graef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009) in a different format than image salience has been modeled (i.e., saliency maps; Harel et al., 2006; Itti et al., 1998; Koch & Ullman, 1987). Meaning maps resolve this issue as they represent the spatial distribution of meaning across a scene in the same format as saliency maps represent the spatial distribution of image salience (Henderson & Hayes, 2017). In an original study, Henderson and Hayes (2017) found that the spatial distribution of eye fixations from aesthetic judgment and memorization tasks were more related to meaning than to image salience. When the intercorrelation between meaning and image salience was statistically controlled, meaning continued to account for variance whereas saliency did not. Although this initial study provided strong evidence that the unique variance in attention is captured by meaning but not image salience, the memorization and aesthetic judgment tasks used may have biased participants to attend meaningful scene regions, resulting in an undue advantage of meaning over saliency. Given this limitation, Chapters 2 and 3 test whether there is an advantage of meaning or image salience in tasks that do not bias participants' attention towards meaning.

Demonstrating an advantage of meaning over image salience in these contexts would suggest that viewers cannot help but attend to meaning (rather than image salience) regardless of the context/situation. This would allow future research to generate models of scene perception that incorporate scene meaning with other top-down forms of knowledge, such as task. It would also provide us a better understanding of why the visual system selects certain regions of the world for analysis over others which, in turn, could be used in real-world settings to minimize the influence of distractors on attention.

Spatial Constraint

If we observe an advantage of meaning over saliency on attention regardless of the task or context, then a next logical step is to test whether meaning interacts with other top-down forms of knowledge, such as spatial constraint, to guide attention. Spatial constraint describes our knowledge of the likely locations of objects in scenes. For example, our prior knowledge of paintings suggests they will appear in upper scene regions on walls, while garbage bins will most likely appear in lower scene regions on the floor. When saliency maps are combined with information about the spatial constraint of a search target object (salient regions with a high probability of containing search targets are upweighted relative to regions with a low probability of containing search targets), these maps predict visual search fixations significantly better than image salience alone (Ehinger et al., 2009; Torralba et al., 2006). Although attention to image salience is modulated by spatial constraint, it is unknown whether attention to meaning, which outperforms image salience in predicting fixation placement (Hayes & Henderson, 2019; Henderson et al., 2018, 2019; Henderson & Hayes, 2017, 2018; Peacock et al., 2019b, 2020; Rehrig et al., 2020), is also modulated by spatial constraint. Chapter 4 combines spatial constraint and meaning to test whether the visual system selects meaningful regions at target-relevant locations for attention.

Knowledge of the relationship between meaning and spatial constraint is important for informing theories of visual search that integrate our knowledge of the world with our current goals to predict where we will look and why. Because both task (Ehinger et al., 2009; Einhäuser et al., 2008; Pereira & Castelhano, 2019; Torralba et al., 2006) and semantics (Antes, 1974; Castelhano & Henderson, 2007; Henderson, 2017; Mackworth & Morandi, 1967) independently predict where viewers fixate, it seems important to understand how the visual system integrates these forms of top-down knowledge to predict where viewers attend. This, in turn, could inform technologies that highlight relevant information with regard to our current goals.

Summary

This dissertation asks what types of real-world scene information are prioritized for attention. Chapters 2 and 3 aim to answer whether meaningful or physically salient scene regions are prioritized for attention during tasks in which either image salience is relevant (Chapter 2) or during free viewing which introduces no requirement to attend to meaning or salience (Chapter

3). Chapter 4 asks whether attention prioritizes meaningful scene regions that appear in target-relevant locations. These studies will provide a framework to understand what scene properties the visual system prioritizes for attentional selection. By understanding why the visual system selects some scene properties at the expense of others, we can better understand how humans deploy attention in the real world.

Chapter 2: Meaning Guides Attention During Scene Viewing, Even When It Is Irrelevant

The following chapter consists of a manuscript that is published at

Attention, Perception, and Psychophysics.

Abstract

During real-world scene viewing, humans must prioritize scene regions for attention. What are the roles of low-level image salience and high-level semantic meaning on attentional prioritization? A previous study suggested that when salience and meaning are directly contrasted in scene memorization and preference tasks, attentional priority is assigned by meaning (Henderson & Hayes, 2017). Here we examined the role of meaning on attentional guidance using two tasks in which meaning is irrelevant and saliency is relevant: a brightness rating task and a brightness search task. Meaning was represented by meaning maps that captured the spatial distribution of semantic features. Meaning was contrasted with image salience represented by saliency maps. Critically, both maps were represented similarly, allowing us to directly compare how meaning and salience influenced the spatial distribution of attention as measured by fixation density maps. Our findings suggest that even in tasks for which meaning is irrelevant and salience is relevant, meaningful scene regions are prioritized for attention over salient scene regions. These results support theories in which scene semantics play a dominant role in attentional guidance in scenes.

Because we can only attend to a small portion of the visual information available to us, we have to select some regions of the visual scene for preferential analysis at the expense of others via attention. It is therefore important to understand the mechanisms by which we guide our attention through real-world scenes. A good deal of work on attentional guidance in scenes has focused on the idea that attention is driven by bottom-up, low-level image features such as color, luminance, and edge orientation that are combined into saliency maps (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Harel, Koch, & Perona, 2006; Itti & Koch, 2001). Saliency maps are appealing because they are computationally tractable and neurobiologically plausible (Henderson, 2007, 2017).

At the same time, there is strong evidence that visual attention is influenced by cognitive factors such as the semantic informativeness of objects and entities within a scene (Antes, 1974; Henderson, 2017; Henderson, Brockmole, Castelhano, & Mack, 2007; Mackworth & Morandi, 1967), along with the viewer's task and current goal (Buswell, 1935; Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Henderson, 2007, 2017; Henderson & Hollingworth, 1999; Navalpakkam & Itti, 2005, 2007; Rothkopf, Ballard, & Hayhoe, 2016; Tatler, Hayhoe, Land, & Ballard, 2011; Yarbus, 1967). Yet much of the research on attentional guidance has continued to focus solely on image salience. One reason for the popularity of image salience is that it is relatively straightforward to compute and represent. In contrast, it has been less clear how to generate and represent the spatial distribution of semantic features across a scene. To directly compare image salience to semantic informativeness, it is necessary to represent scene meaning in a format equivalent to that of image salience.

To address this issue, we have recently introduced the concept of meaning maps as a way to represent the spatial distribution of scene semantics (Henderson & Hayes, 2017). To generate meaning maps, Henderson and Hayes (2017) used crowd-sourced responses in which naïve participants rated the meaning of image patches from real-world scenes. Specifically, photographs of scenes were divided into a dense array of objectively defined circular overlapping patches at a coarse and a fine spatial scale. These patches were then shown to raters who rated how informative or recognizable each patch was (see also Antes, 1974; Mackworth & Morandi, 1967). Finally, meaning maps of each scene were created by interpolating the ratings at each spatial scale and averaging across the two scales.

Meaning maps provide a pixel-by-pixel prediction of semantic content across a scene just as saliency maps provide a pixel-by-pixel prediction of saliency across a scene. Since meaning maps are represented in the same format as saliency maps, their predictions for visual attention can be directly compared to saliency maps using the methods that have typically been used to compare the relationship between saliency maps and attention (Carmi & Itti, 2006; Itti, Koch, & Niebur, 1998; Parkhurst, Law, & Niebur, 2002; Torralba, Oliva, Castelhano, & Henderson, 2006). In this way, meaning maps and saliency maps together provide a way to compare how meaning and salience influence visual attention during real-world scene viewing.

Henderson and Hayes (2017) investigated the degree to which meaning maps and saliency maps predicted visual attention in real-world scenes during memorization and aesthetic judgment tasks. In that study, attention maps were created based on the locations of eye fixations. The results showed that meaning maps and saliency maps were highly correlated, and both were able to predict the spatial distribution of attention in scenes. Importantly, in both tasks meaning accounted for significantly more of the variance in attention than image salience. Further, when the variance due to salience was controlled, meaning accounted for significantly more of the remaining variance in attention, but when meaning was controlled, no additional variance in attention was accounted for by salience. These results held across the entire viewing time. Henderson and Hayes (2018) replicated this pattern of results using attention maps constructed from duration-weighted fixations, and Henderson, Hayes, Rehrig, and Ferreira (2018) showed that the results extended to scene description tasks. In total, the findings showed that meaning (rather than image salience) was the main driver of visual attention.

Although the data favoring meaning over image salience have been clear, it could be argued that the viewing tasks used to compare meaning and image salience were biased toward meaning. That is, it might be that memorization, aesthetic preference, and scene description tasks require the viewer to focus on the semantic features of scenes. If this is true, then it may be that the advantage for meaning over salience is restricted to viewing tasks that specifically require analysis of meaning. To address this hypothesis, the current study investigated whether attention continues to be guided by meaning during scene viewing even when saliency is relevant, and meaning is irrelevant to the viewer's task.

Specifically, in the present study we used two tasks that were designed to emphasize salience and eliminate the need for meaning in attentional guidance: a brightness rating task in which participants rated scenes for overall brightness, and a brightness search task in which participants counted the number of bright patches within scenes (Figure 2.1). Critically, these tasks were designed to make meaning task-irrelevant and salience task-relevant. If the use of meaning to guide attention is task-based, then the relationship between meaning and attention found in our earlier studies should no longer be observed in these conditions. On the other hand, if the use of meaning to guide attention during scene viewing is a fundamental property of the attention system, then we should continue to observe a relationship between meaning and attention even when only salience is relevant to the task.

Method

Eye-tracking

Participants. Thirty University of California, Davis undergraduate students with normal or corrected-to-normal vision participated in the experiment (25 females, average age = 20.84). All participants were naïve concerning the purpose of the experiment and provided verbal consent. The eye-movement data from each participant were filtered for excessive track losses due to blinks or loss of calibration. Following Henderson & Hayes (2017), we averaged the percent signal ([number of good samples / total number of samples] x 100) for each trial and participant using custom MATLAB code. The percent signal for each trial was then averaged for each subject and compared to an *a priori* 75% criterion for signal. Overall, all participants had greater than 75% signal resulting in no removed subjects.

Apparatus. Eye movements were recorded using an EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01° rms) sampling at 1000 Hz (SR Research, 2010b). Participants sat 85 cm away from a 21" monitor, so that scenes subtended approximately 26.5° x 20° of visual angle at 1024 x 768 pixels. Head movements were minimized by using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software (SR Research, 2010a).

Stimuli. Stimuli consisted of the 40 digitized photographs (1024 x 768 pixels) of indoor and outdoor real-world scenes. Scenes were luminance matched across the scene set by converting the RGB image of the scene to LAB space and scaling the luminance channel of all scenes from 0 to 1. All instruction, calibration, and response screens were luminance matched to the average luminance (*M* = 0.45) of the scenes.

Procedure. Each participant completed two scene-viewing conditions in a within-subject design: a brightness rating task and a brightness search task (Figure 2.1). During the brightness rating task, participants were instructed to rate the overall brightness of the scene on a scale from 1 to 6 (1 = very dark; 2 = dark; 3 = somewhat dark; 4 = somewhat bright; 5 = bright; 6 = very bright). During the brightness search task, participants were instructed to count the number of bright patches within the scene. Because the goal of this study was to assess whether we could eliminate the relationship between meaning and attention in tasks that did not require the use of meaning, we emphasized speed and accuracy during both tasks. Participants were given a maximum scene-viewing time of 12 s (as done in Henderson & Hayes, 2017), but had the option to terminate the scene and continue to the response screen earlier by pressing a key on a button box (RESPONSEPixx; VPixx Technologies, Saint-Bruno, CA). We included the early termination option so that could focus on task-relevant eye movement behavior. Following their button press or the maxiumun 12 s of scene presentation, participants were shown a response screen in which the number 0 was enclosed in a square (Figure 2.1). Then, participants used left and right buttons on the button box to respectively increase or decrease the value of the number until it matched their rating or patch count for that scene. They then pressed the center key to continue to the next scene.

Before starting the experiment, participants completed two practice trials in which they were familiarized with each condition and the button-box. After the practice trials, a 13-point calibration procedure was performed to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error of less than 0.99°. Presentation of each scene was preceded by a drift correction procedure, and the eye-tracker was recalibrated when the calibration was not accurate. The calibration was also repeated between task blocks.

The 40 scene stimuli were randomly divided into two scene sets (set A and set B), each composed of 20 scenes, and for each subject each set was assigned to one task. Task order and scene set assignment was fully counterbalanced across all participants. Additionally, scenes within each set were presented in a randomized order for each participant in each condition.

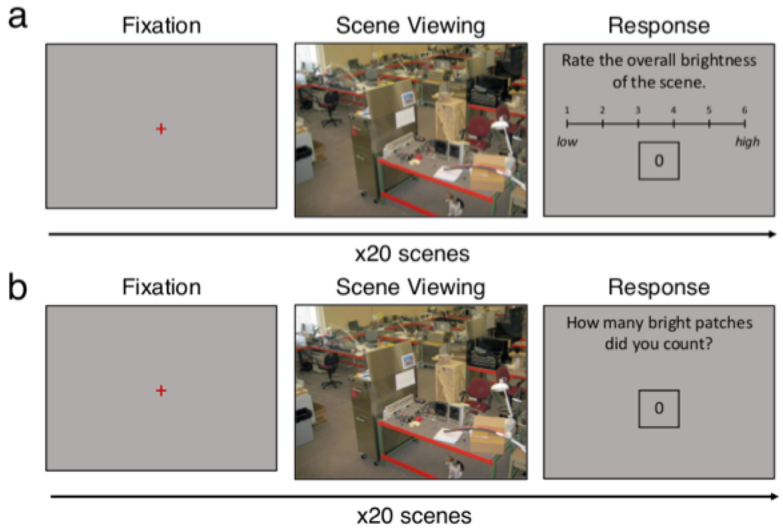


Figure 2.1. Trial structure for the two tasks. The trial structure for (a) the brightness rating task and (b) the brightness search task.

Analysis

All analyses were chosen *a priori* and based on our previous work (Henderson & Hayes, 2017, 2018; Henderson et al., 2018).

Data Segmentation and Outliers

Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s^2; SR Research, 2010b). Eye movement data were imported offline into Matlab using the EDFConverter tool. The first fixation on each scene, always located at the center of the display as a result of the pretrial fixation marker, was eliminated from analysis. Additionally, any fixations that were shorter than 50ms and longer than 1500ms were eliminated as outliers. This outlier removal process resulted in loss of 3.94% of the data.

Attention Maps

Attention maps were generated as described in Henderson and Hayes (2017). Briefly, a fixation frequency matrix based on the locations (*x,y* coordinates) of all fixations was generated across participants for each scene. A Gaussian low-pass filter with a circular boundary and a cutoff frequency of −6dB was applied to each matrix to account for foveal acuity and eyetracker error (Figure 2.2). The spatial extent of the low pass filter was 236 pixels in diameter.

Meaning Maps

Meaning maps were generated as per Henderson and Hayes (2017). Because the nature of our tasks resulted in peripheral fixations, we used both unbiased and center-biased meaning maps (Figure 2.2). Overall, the unbiased maps provided better predictive power than the center-biased maps. However, we included analyses from both because center-biased maps are standard in the literature and thus provide a basis for comparison with previous studies. The center-biased meaning maps were generated by applying a multiplicative center bias operation to the meaning maps using the same center bias present in the saliency maps.

Subjects. Scene patches were rated by 165 subjects on Amazon Mechanical Turk. Participants were recruited from the United States, had a HIT (human intelligence task) approval rate of 99% and 500 HITs approved, and were only allowed to participate in the study once. Participants were paid $0.50 cents per assignment, and all participants provided informed consent.

Stimuli. Stimuli consisted of the 40 digitized photographs used in the current experiment. Each scene was decomposed into a series of partially overlapping and tiled circular patches at coarse and fine spatial scales. The full patch stimulus set consisted of 12,000 unique fine patches and 4,320 unique coarse patches for a total of 16,320 scene patches.

Procedure. Each participant rated 300 random scene patches extracted from the scenes. Participants were instructed to assess the meaningfulness of each patch based on how informative or recognizable they thought it was. During the instruction period, participants were provided with examples of two low-meaning and two high-meaning scene patches to make sure they understood the task. They then rated the meaningfulness of test patches on a six-point Likert scale (‘very low’, ‘low’, ‘somewhat low’, ‘somewhat high’, ‘high’, ‘very high’). Patches were presented in random order and without scene context, so ratings were based on context-independent judgments. Each unique patch was rated three times by three independent raters for a total of 48,960 ratings. However, owing to the high degree of overlap across patches, each fine patch contained rating information from 27 independent raters, and each coarse patch from 63 independent raters.

Meaning maps were generated from the ratings by averaging, smoothing and combining the fine and coarse maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The average fine and coarse rating maps were then smoothed using thin-plate spline interpolation based on the center of each patch (MATLAB ‘fit’ using the ‘thinplateinterp’ method). Finally, the smoothed fine and coarse maps were averaged to produce the meaning map for each scene.

Saliency Maps

Saliency maps for each scene were computed using the Graph-Based Visual Saliency (GBVS) toolbox with default settings (Harel et al., 2006). GBVS is a prominent saliency model that combines maps of low-level image features to create image-based saliency maps (Figure 2.2).

Center bias is a natural feature of the GBVS saliency maps. To compare to unbiased meaning maps, we also generated GBVS maps without center bias (Figure 2.2). These maps were created using a whitening method (Rahman & Bruce, 2015), a 2-step normalization approach in which each saliency map is normalized to have 0 mean and unit variance. After this, a second pixel-wise normalization is performed so each pixel location across all the saliency maps has 0 mean and unit variance.

Histogram matching. Following Henderson and Hayes (2017), meaning and saliency maps were normalized to a common scale using image histogram matching with the fixation density map for each scene serving as the reference image for the corresponding meaning and saliency maps. This was accomplished by using the Matlab function 'imhistmatch' from the Image Processing Toolbox.

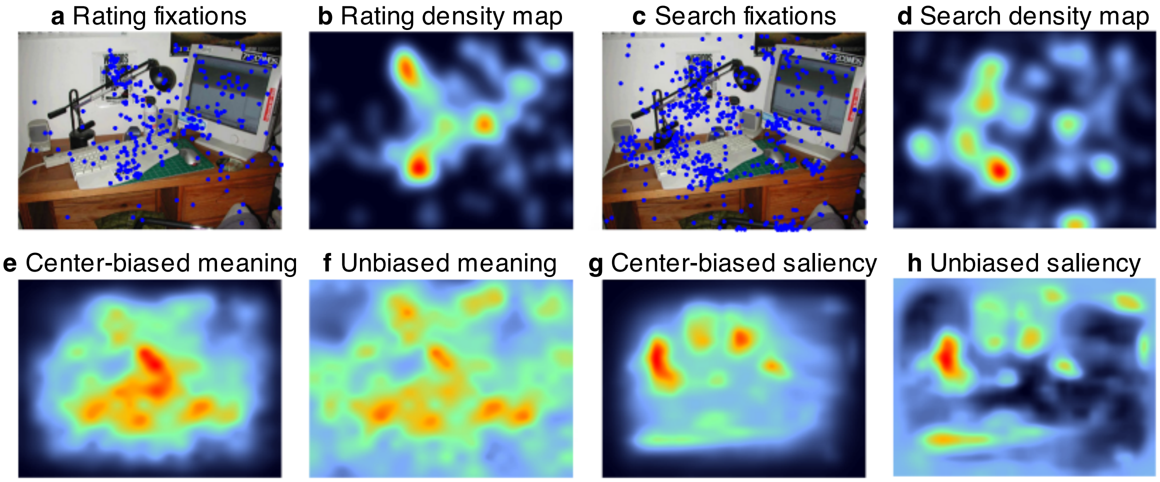


Figure 2.2. An example scene with the associated maps for each task. (a) is an example scene with fixation locations from all participants in the rating task aggregated and overlaid. (b) is the fixation density map representing the example scene and fixation locations for the rating task. (c) is the example scene with fixations from the search task overlaid and (d) is the fixation density map representing the example scene and fixation locations in the search task. (e) is the center-biased meaning map and (f) is the unbiased meaning map for the example scene. (g) is the center-biased saliency map and (h) is the unbiased saliency map for the example scene.

Results

Task Comparisons

Scene viewing. Because we gave participants the option to terminate each presentation trial early, we began by comparing the average scene-viewing (scene onset to response) time for each scene during each condition as well as the number of fixations per scene in each task (Figure 2.3). The average scene viewing time for the brightness rating task was 5262.55ms (*SD* = 3141.39) with 15.56 fixations (*SD* = 9.84), and for the brightness search task was 10726.52ms (*SD* = 2420.55) with 32.28 fixations (*SD* = 8.20). Because the distributions were not normal (Figure 2.3), Wilcoxon rank sum tests were conducted and showed that the scene viewing times and number of fixations were significantly different between the rating and search tasks: *Z*'s > 5.50; *p*'s < 0.001. These results showed that participants tended to view scenes during the rating task for shorter durations than the search task, with participants much more likely to use the entire 12s in the search compared to the rating task. The finding that the rating task produced significantly shorter viewing durations than the search task suggests that participants only viewed the scenes for the amount of time necessary to complete each task. Given that the viewing times and number of fixations were very different between the tasks, we treated the two tasks separately in the following analyses.

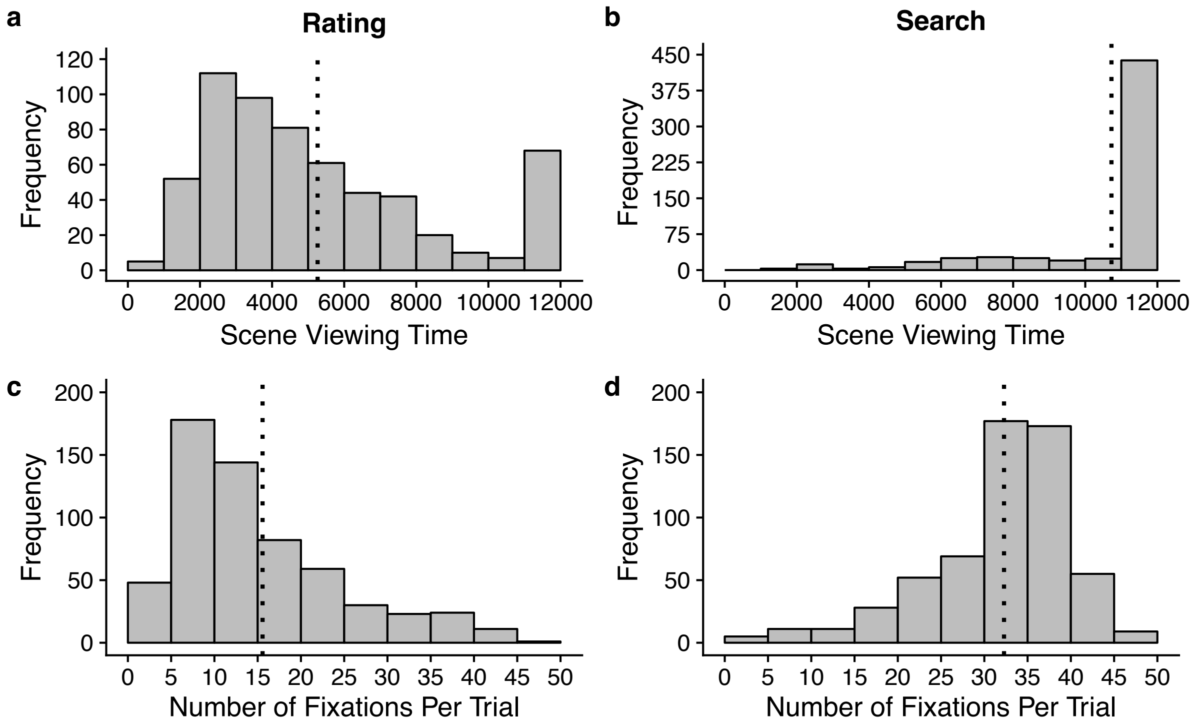


Figure 2.3. Scene-viewing times and number of fixations per trial for the brightness rating and brightness search tasks. Distributions are shown for the scene viewing times of (a) the brightness rating and (b) brightness search tasks, and the number of fixations per trial of (c) the brightness rating and (d) brightness search tasks. Black dotted vertical lines represent the mean for each task.

Response agreement. To verify that subjects were staying on task and attending to brightness during the study, we examined response agreement in the rating and search tasks. If subjects were on-task, then their responses should vary as a function of scene and be consistent within scenes. That is, subjects should generally agree in their judgements of brightness in the rating task and the number of bright regions in the search task. On the other hand, if subjects were simply attending to scene content rather than following instructions, then responses should be unsystematic across scenes and subjects. As can be seen in Figure 2.4, the former was true, suggesting that subjects were indeed following instructions.

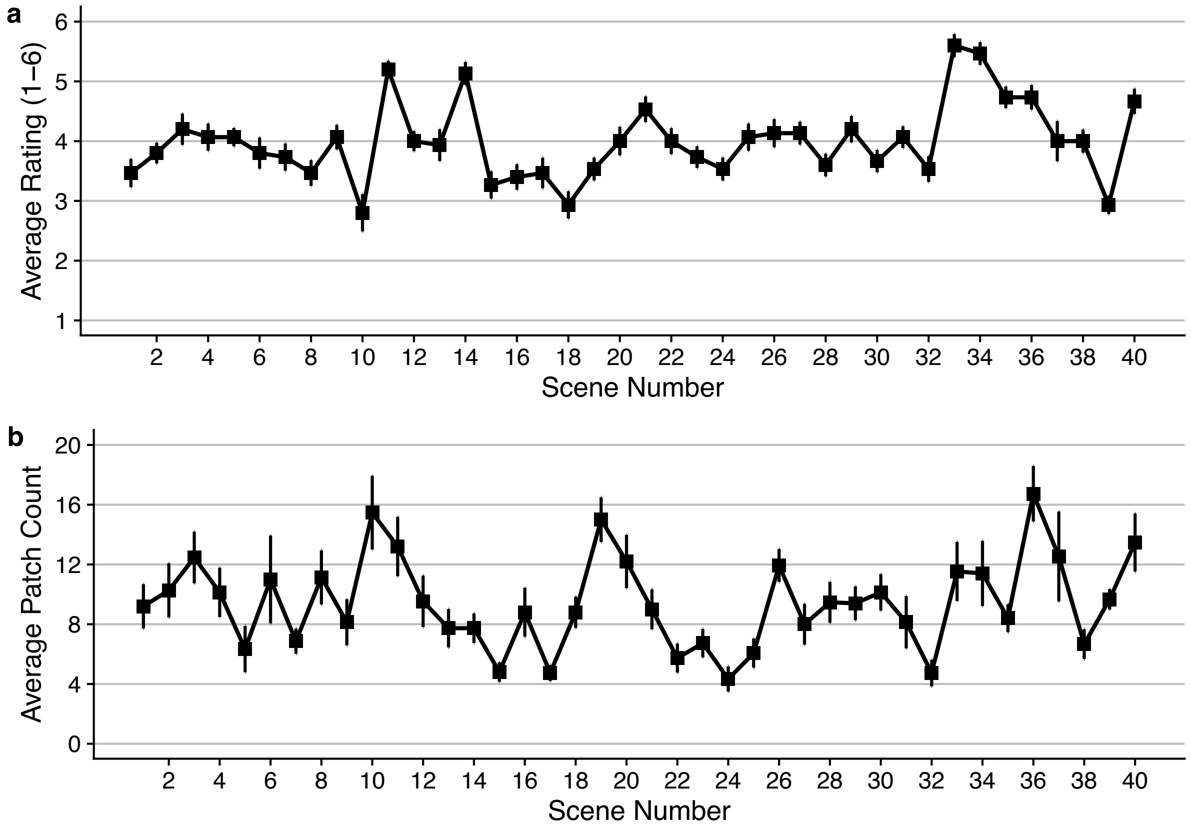


Figure 2.4. Response variability as a function of scene. The average participant response and standard error of responses as a function of scene for (a) the brightness rating task and (b) the patch count task.

Overall Scene Analyses

Following Henderson and Hayes (2017), we used squared linear and semi-partial correlations to quantify the degree to which meaning maps and saliency maps accounted for shared and unique variance in the attention maps. Specifically, we conducted two-tailed, two-sample t-tests for the correlations across scenes to statistically compare the relative ability of meaning and salience to predict attentional guidance.

For comparison to the literature, we tested how well traditional center-biased meaning and saliency maps could account for attention. In addition, because the center bias was substantially reduced in the brightness search task compared to the brightness rating task (Figure

2.5), we also conducted analyses using unbiased meaning and saliency maps that excluded a center bias.

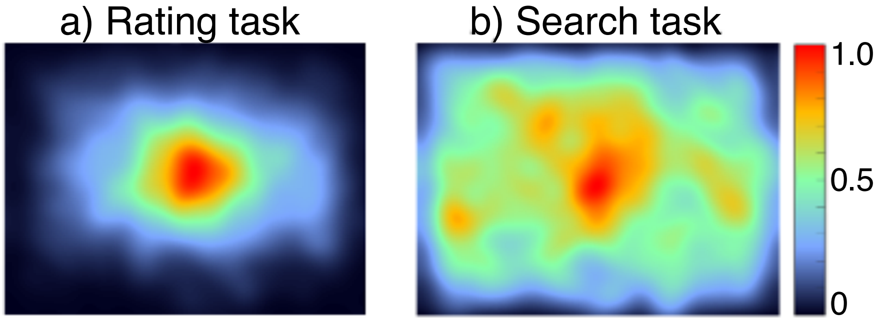


Figure 2.5. Center biased maps. Fixation density maps aggregated across subjects and scenes are shown for (a) the brightness rating task and (b) the brightness search task.

Brightness rating task. Using the center-biased maps, for squared linear correlations on average across all 40 scenes, meaning accounted for 55% of the variance in fixation density (M = 0.55, SD = 0.12) and salience accounted for 33% of the variance in fixation density (M = 0.33, SD = 0.14) (Figure 2.6). This difference between meaning and saliency maps was significant: $t(78)$= 7.31, $p < 0.001$, 95% CI = [0.16, 0.28]. Similarly, for squared semi-partial correlations, meaning accounted for 24% of the variance in fixation density (M = 0.24, SD = 0.13) controlling for salience, but salience accounted for only 3% of the variance in fixation density controlling for meaning (M = 0.03, SD = 0.03) (Figure 2.6). This difference was again significant: $t(78)$= 10.57, $p < 0.001$, 95% CI = [0.17, 0.25]. This pattern of results did not change when using the unbiased meaning and saliency maps (linear: $t(78)$ = 8.79, $p < 0.001$, 95% CI = [0.16, 0.25]; semi-partial: $t(78)$= 9.62, $p < 0.001$, 95% CI = [0.16, 0.25]) (Figure 2.6). These findings suggest that meaning played a dominant role in the guidance of attention even though meaning was irrelevant and salience was central to the brightness rating task.

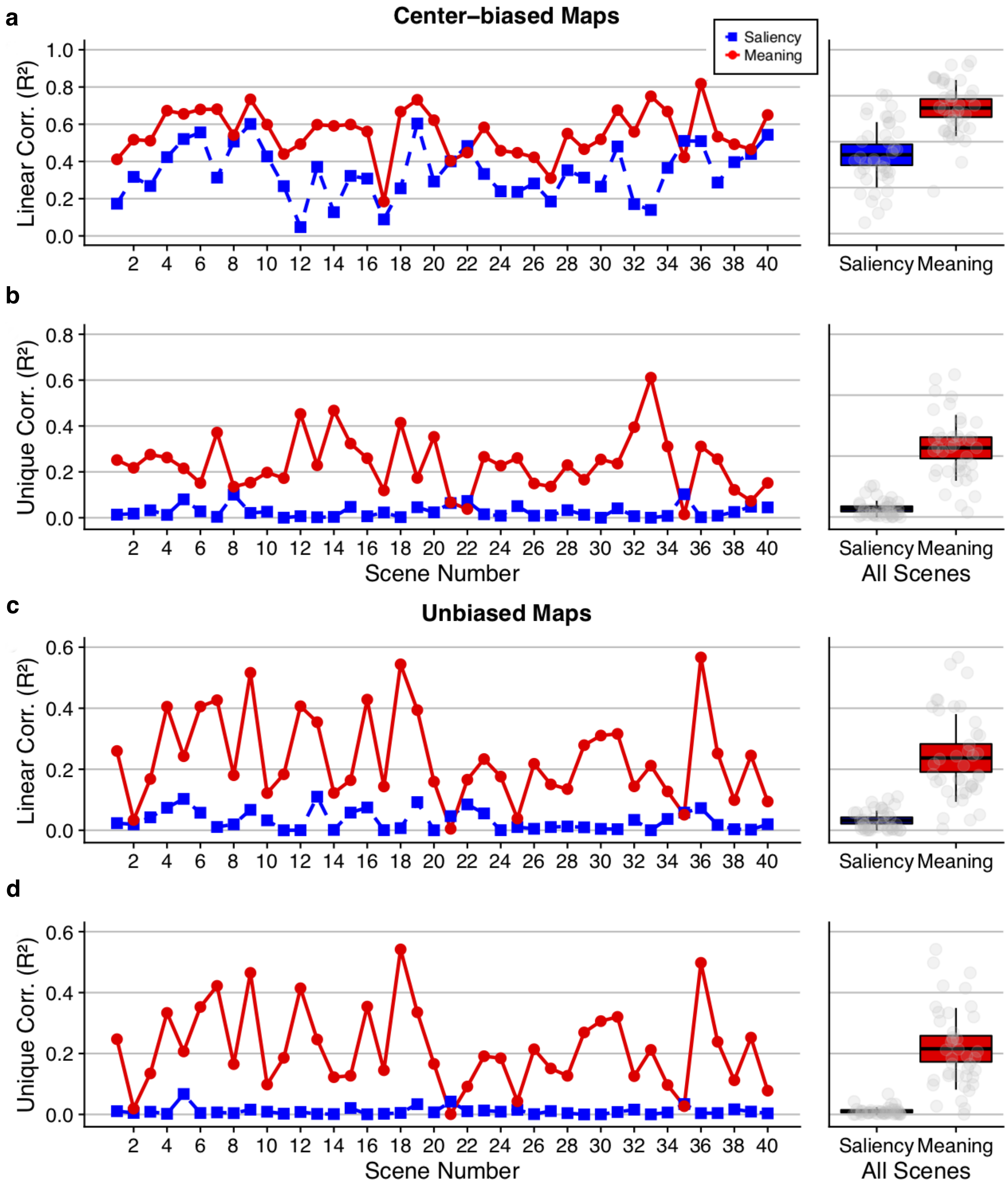


Figure 2.6. Squared linear correlation and semi-partial correlation by scene for the brightness rating task. Line plots show (a) the squared linear and (b) semi-partial correlations between the fixation density maps and meaning (red circles) and salience (blue squares) using center-biased meaning and saliency maps. Line plots also show (c) the squared linear and (d) semi-partial correlations using unbiased meaning and saliency maps. The scatter plots on the right show the grand mean (black horizontal line), 95% confidence intervals (colored boxes), and 1 standard deviation (black vertical line) for meaning and salience across all 40 scenes for each analysis.

Brightness search task. Using the center-biased maps, meaning accounted for 22% of the variance in fixation density (*M* = 0.22, *SD* = 0.13) and salience accounted for 24% of the variance in fixation density (*M* = 0.24, *SD* = 0.12) (Figure 2.7). This difference was not significant: $t(78)= -0.33$, $p = 0.74$, 95% CI = [–0.07, 0.05]. Similarly, for the semi-partial correlations, meaning accounted for 5% of the variance in fixation density controlling for salience (*M* = 0.05, *SD* = 0.07) and salience accounted for 6% of the variance in fixation density controlling for meaning (*M* = 0.06, *SD* = 0.07) (Figure 2.7). Again, this difference was not significant: $t(78)= -0.59$, $p = 0.56$, 95% CI = [–0.04, 0.02]. Importantly, however, this pattern of results changed when using the unbiased meaning and saliency maps (Figure 2.7). Using the unbiased maps, meaning accounted for 22% of the overall variance in attention (*M* = 0.22, *SD* = 0.11) whereas salience explained only 4% of the variance (*M* = 0.04, *SD* = 0.05) for the linear correlations, $t(78)= 6.42$, $p < 0.001$, 95% CI = [0.10, 0.18]. Similarly, for the semi-partial correlations, meaning accounted for 18% of the total variance in attention (*M* = 0.18, *SD* = 0.11) whereas salience explained only 1% of the variance (*M* = 0.04, *SD* = 0.04), $t(78)= 7.42$, $p < 0.001$, 95% CI = [0.10, 0.18]. These findings suggest that when the more distributed nature of attention away from scene centers and to scene peripheries in the brightness search task was taken into account, meaning influenced attentional guidance more than salience even though meaning was irrelevant and saliency was central to the task.

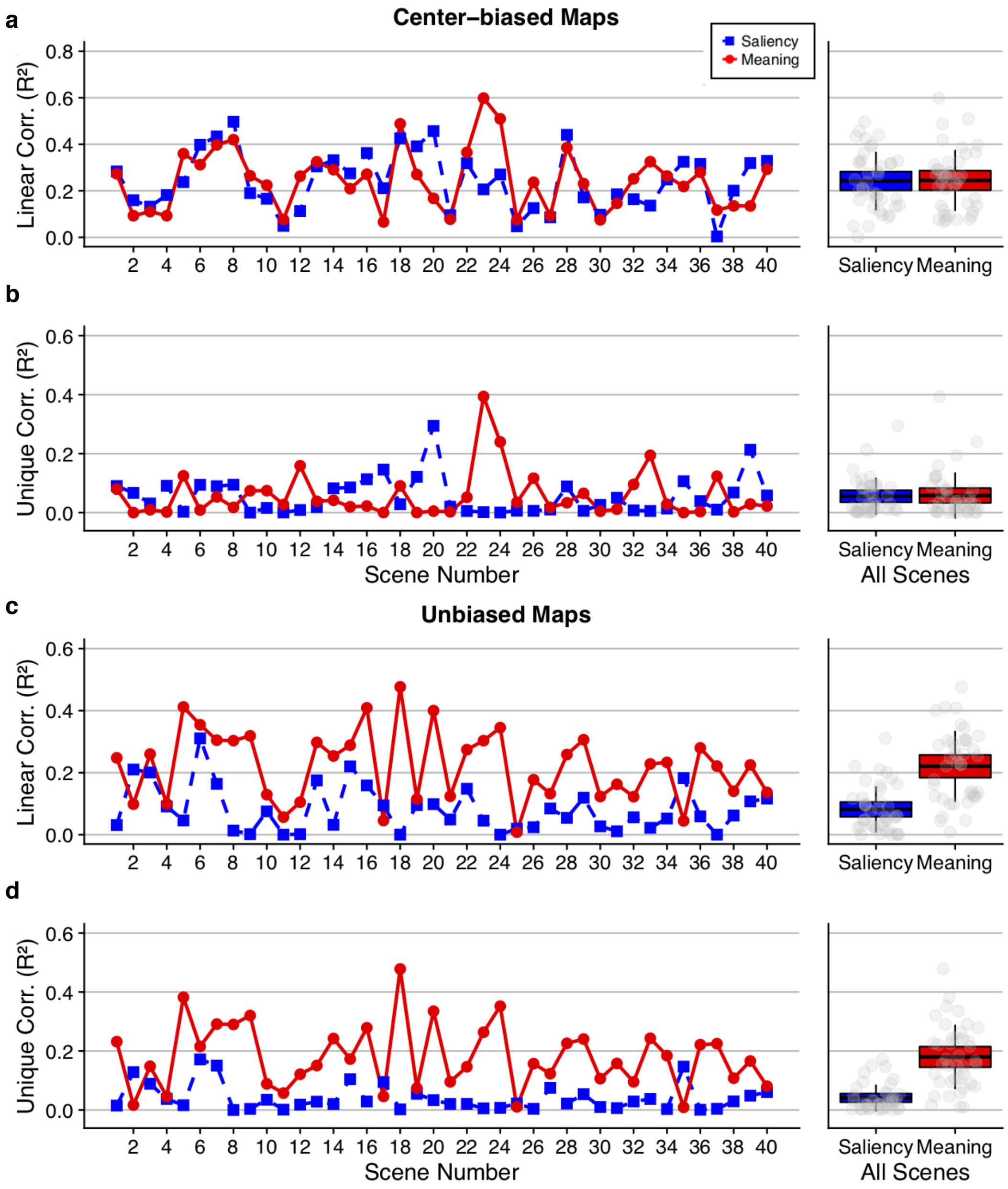


Figure 2.7. Squared linear correlation and semi-partial correlation by scene for the brightness search task. The line plots show (a) the linear and (b) semi-partial correlations between fixation density and meaning (red circles) and salience (blue squares) for the search task using the center-biased meaning and saliency maps. Line plots also show (c) the linear and (d) semi-partial correlations for the search task using the unbiased meaning and saliency maps. The scatter plots on the right show the corresponding grand mean (black line), 95% confidence intervals (colored box), and one standard deviation (black vertical line) for meaning and salience across all scenes.

Fixation by Fixation Analyses

Previously, it has been posited that attention during scene viewing might initially be guided by salience, but that as time progresses, meaning begins to play an increasing role

(Anderson, Donk, & Meeter, 2016; Anderson, Ort, Kruijne, Meeter, & Donk, 2015; Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999; Parkhurst et al., 2002). On the other hand, in two studies investigating the roles of meaning and salience in memorization and scene description tasks, we did not observe this change from guidance by salience to guidance by meaning (Henderson & Hayes, 2017; Henderson et al., 2018). Instead, meaning was found to guide attention from the first saccade. Because the current tasks were designed to make meaning irrelevant and salience central, they provide another opportunity to test this hypothesis.

We conducted a temporal time-step analysis in which a series of attention maps were generated from each sequential fixation (1st fixation, 2nd fixation, 3rd fixation, etc.) for each scene in each task. We then correlated each attention map for each fixation and scene using both the center-biased and unbiased meaning and saliency maps to calculate the squared linear and semi-partial correlations. Then the correlations for each scene and fixation were averaged across scenes to assess how meaning and image salience predicted attention on a fixation by fixation basis. The prediction of the salience first hypothesis is that the correlation between saliency and attention maps should be greater for earlier than later fixations, with salience dominating meaning in the earliest fixations.

Brightness rating task. Using the center-biased maps, meaning accounted for 34%, 23%, and 17% of the variance in the first 3 fixations whereas salience accounted for 8%, 12%, and 11% of the variance in the first 3 fixations, respectively, for the linear correlations (Figure 2.8). Two-sample, two-tailed t-tests compared meaning and salience for all 8 initial fixations using p-values corrected for multiple comparisons using a false discovery rate (FDR) correction (Benjamini & Hochberg, 1995). Overall, this confirmed the advantage for meaning over salience for all 8 fixations (all FDR *p*s < 0.05). Similarly, for the semi-partial correlations, meaning accounted for 28%, 14%, and 9% of the variance in the first 3 fixations and salience accounted for 2%, 3%, and 3% of the variance in the first 3 fixations (Figure 2.8). Again, meaning predicted attention significantly better than salience for all 8 initial fixations (all FDR *p*s < 0.001). Using the unbiased maps, this overall pattern of results did not change (linear and semi-partial correlations: all FDR *p*s < 0.001) (Figure 2.8). These results do not support the hypothesis that the influence of meaning on attentional guidance was delayed to later fixations.

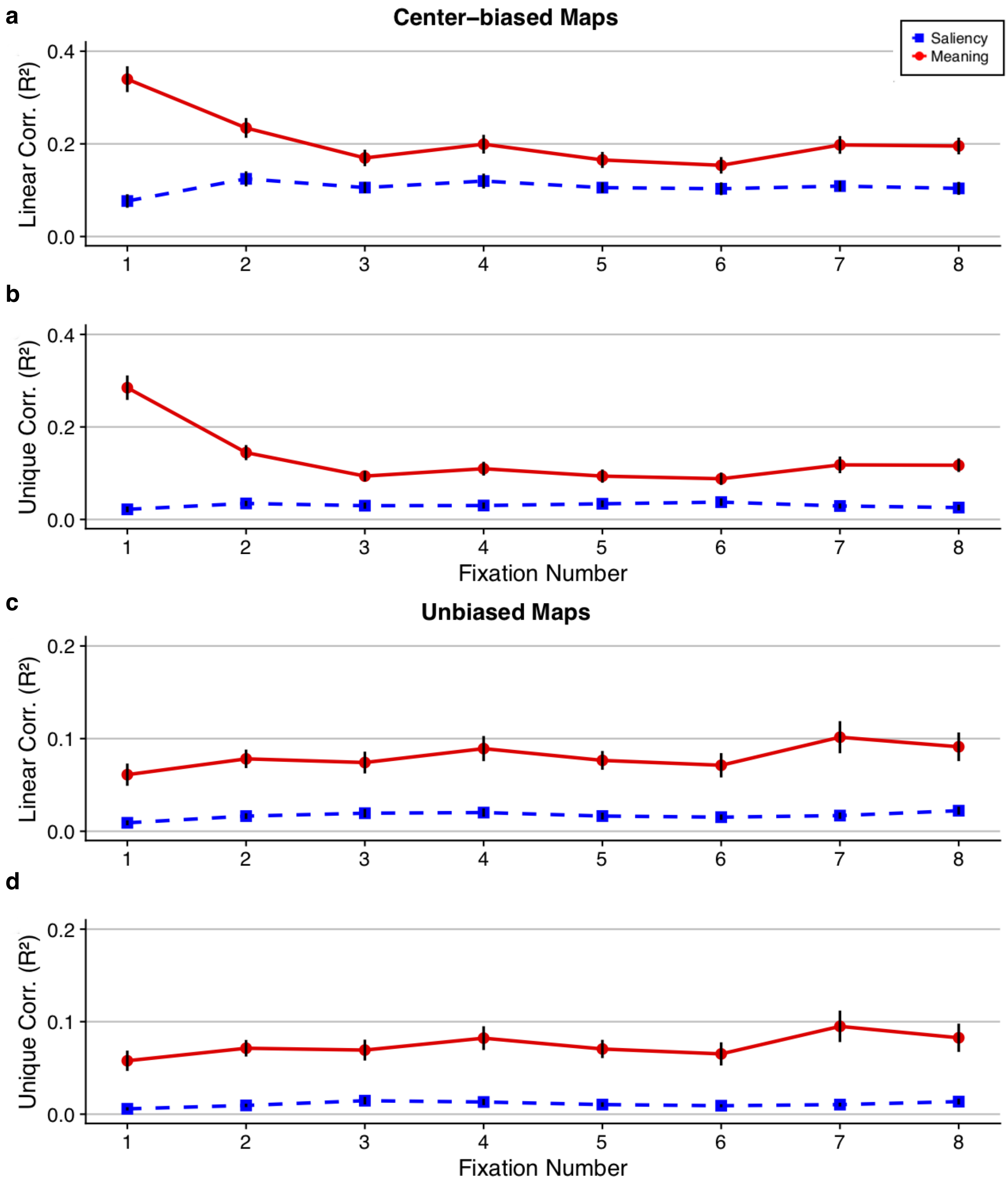


Figure 2.8. Fixation by fixation time-step analyses for the brightness rating task. The line plots show (a) the squared linear and (b) semi-partial correlations between fixation density and meaning (red circles) and salience (blue squares) as a function of fixation number collapsed across scenes for the rating task using the center-biased maps. Line plots also show (c) the squared linear and (d) semi-partial correlations between fixation density and meaning (red circles) and salience (blue squares) as a function of fixation order using the unbiased maps. Error bars represent the standard error of the mean.

Brightness search task. Using the center-biased maps, meaning accounted for 30%, 14%, and 7% of the variance in the first 3 fixations and salience accounted for 11%, 16%, and 14% in the first 3 fixations, respectively, for the linear correlations (Figure 2.9). Here, meaning produced an advantage over salience on the first fixation (FDR $p < 0.001$) but not fixations 2 through 8 (FDR $p > 0.05$). For the semi-partial correlations, meaning explained 22%, 8%, and 3% of the variance in the first 3 fixations and salience accounted for 3%, 10%, and 10% in the first 3 fixations (Figure 2.9). A significant advantage for meaning was observed on the first fixation (FDR $p < 0.001$) and for salience on the third fixation (FDR $p < 0.05$), with no other comparisons reaching significance (FDR *p*s > 0.05).

Using the unbiased meaning and saliency maps, the pattern of results changed. For the linear correlations, meaning accounted for 5%, 6%, and 4% of the variance and salience accounted for 1%, 4%, and 5% of the variance in attention in the first 3 fixations. Turning to the semi-partial correlations, meaning accounted for 5%, 6%, and 3% of the variance and salience accounted for 0.1%, 3%, and 4% of the variance in attention in the first 3 fixations. Meaning still produced an advantage over salience for the first fixation (linear and semi-partial FDR $p < 0.05$) with all other fixations nonsignificant (linear and semi-partial FDR $p > 0.05$). The advantage for saliency over meaning for the third fixation seen in the center-biased maps was not observed with the unbiased maps.

The fixation by fixation analyses were not consistent with the salience first hypothesis. In the analyses using both the center-biased and unbiased maps, meaning was more important than salience at the first fixation. Using the center-biased maps, salience was stronger in the third fixation. This result, however, was not true using the unbiased maps, suggesting that the advantage for saliency in the center-biased maps was driven by the center bias rather than saliency itself. Overall, the results are not consistent with the hypothesis that attentional guidance transitions from salience to meaning over time.

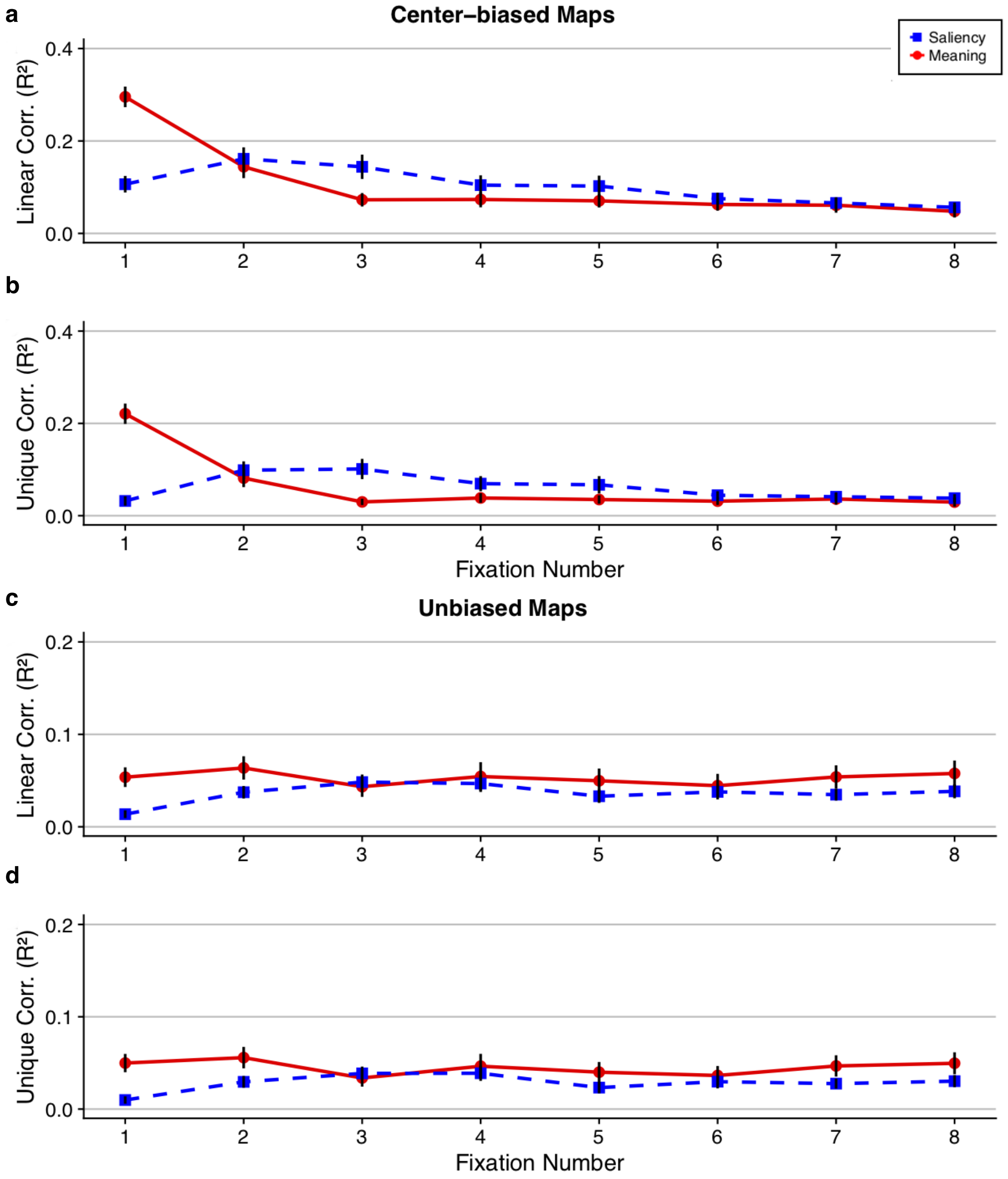


Figure 2.9. Fixation by fixation time-step analyses for the brightness search task. The line plots show (a) the squared linear and (b) semi-partial correlations between fixation density and meaning (red circles) and salience (blue squares) as a function of fixation number collapsed across scenes for the search task using the center-biased maps. Line plots also show (c) the squared linear and (d) semi-partial correlations between fixation density and meaning (red circles) and salience (blue squares) as a function of fixation order using the unbiased maps. Error bars represent the standard error of the mean.

Saccade Amplitude Analyses

In the analyses thus far, fixations following both shorter and longer saccades were included. It could be that meaning guides attention within local scene regions, whereas salience guides attention as it moves from one scene region to another. To test this hypothesis, we analyzed the role of meaning on attentional guidance as a function of saccade amplitude. If meaning plays a greater role for local (e.g., within-object) shifts of attention, then meaning should be more related to attentional selection following shorter saccades versus longer saccades. Such a pattern might be more likely in the case of the current study because meaning was not relevant to the tasks. To investigate this hypothesis, we assessed how meaning and salience related to attention following saccades of shorter to longer amplitudes (Figure 2.10). Specifically, saccade amplitudes were binned by decile, and fixation density maps were created for each saccade amplitude decile. Meaning and salience maps were then correlated with the fixation density maps for each decile. We conducted these analyses using both the center-biased and unbiased meaning and saliency maps. The saccade amplitude average for the rating task was 5.37° (*SD* = 3.41) and for the search task was 4.61° (*SD* = 3.51).

Brightness rating task. For the brightness rating task, using the center-biased maps, meaning produced an advantage over salience for saccade amplitude deciles 1 through 7 and 9 (FDR $p < 0.05$) but not deciles 8 and 10 (FDR $p > 0.05$). For the semi-partial correlations, meaning explained significantly more of the variance in fixation density than salience for all 10 saccade amplitude deciles (all FDR *p*s < 0.05). When using the unbiased meaning and saliency maps, this pattern of results became stronger as meaning produced an advantage over saliency across all deciles in both the linear and semi-partial correlations (FDR $p < 0.05$).

Brightness search task. For the brightness search task, using the center-biased maps, there were no significant differences between meaning and salience for any saccade amplitude deciles in either the linear or semi-partial correlations (all FDR *p*s > 0.05). When using the unbiased maps, on the other hand, this pattern of results changed as meaning produced an advantage over saliency for saccade amplitude deciles 1 through 9 (FDR $p < 0.05$) but not 10 (FDR $p > 0.05$).

Overall, it appears that meaning was used to guide attention for both short and long shifts of attention, though there was some evidence that this influence was reduced when the scene peripheries were removed from the analyses (i.e., with the center-biased maps) and for the longest shifts of attention.

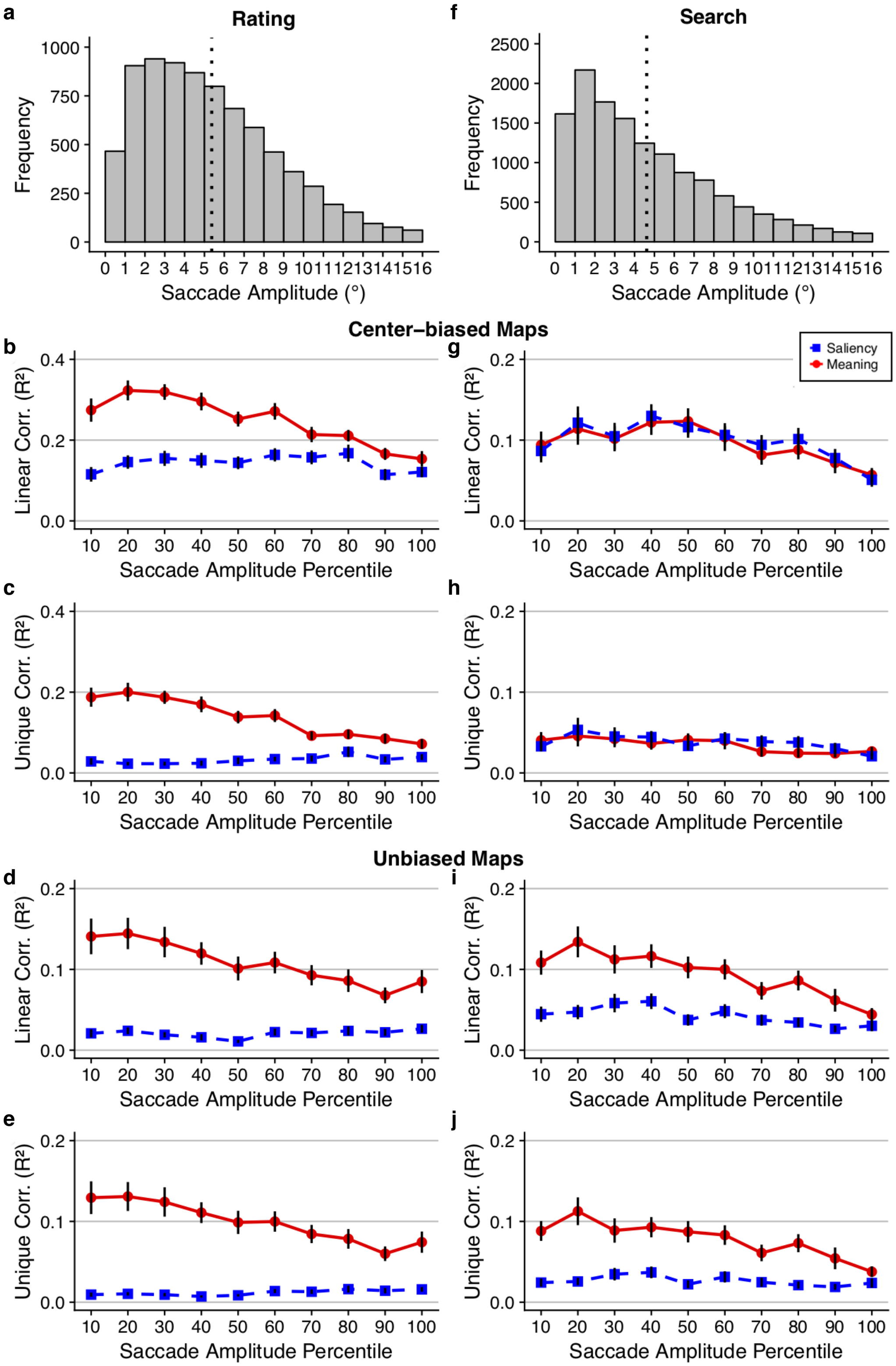


Figure 2.10. Squared linear correlation and squared semi-partial correlation as a function of saccade amplitude to fixation. The saccade amplitude results for the rating task are shown in the first column (a through e) in which (a) shows a histogram of saccade amplitude frequencies and average saccade amplitude (black dotted line), (b) and (d) show the squared linear and (c) and (e)

show the semi-partial correlations between meaning (red circles) and saliency (blue squares) and fixation density as a function of saccade amplitude percentiles prior to fixation for the center-biased maps (b and d) and the unbiased maps (c and e). The second column (f through j) shows the saccade amplitude results for the search condition in which (d) shows a histogram of saccade amplitude frequencies and the average saccade amplitude (black dotted line), (g) and (i) show the squared linear correlations and (h) and (j) show the semi-partial correlations between meaning (red circles) and saliency (blue squares) with fixation density as a function of saccade amplitude percentiles using the center-biased maps (g and h) and the unbiased maps (i and j). Data points are averaged across all 40 scenes at each decile. Error bars represent standard error of the mean.

General Discussion

Past research has emphasized image salience as a key basis for attentional selection during real-world scene viewing (Borji et al., 2014; Borji et al., 2013; Harel et al., 2006; Itti & Koch, 2001; Koch & Ullman, 1985; Parkhurst et al., 2002). Although this previous work has provided an important framework for understanding attentional guidance in scenes, it downplays the fact that attention is strongly guided by cognitive factors related to semantic features that are relevant to understanding the scene in the context of the task (Buswell, 1935; Hayhoe & Ballard, 2005; Hayhoe et al., 2003; Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009; Land & Hayhoe, 2001; Rothkopf et al., 2016; Yarbus, 1967). With the development of meaning maps, which capture the spatial distribution of semantic content in scenes in the same format that saliency maps capture the spatial distribution of image salience, it has become possible to directly compare the influence of meaning and image salience on attention in scenes (Henderson & Hayes, 2017).

In prior studies comparing meaning and image salience during scene viewing, meaning has better explained the spatial and temporal patterns of attention (Henderson & Hayes, 2017, 2018; Henderson et al., 2018). However, those studies used memorization, aesthetic judgement, and scene description viewing tasks, and it could be argued that those tasks were biased towards attentional guidance by meaning. The current study sought to determine whether the influence of meaning on attention would be eliminated in tasks that do not require any semantic analysis of the scenes. To test this hypothesis, we used two viewing tasks designed to eliminate the need for attending to meaning: a brightness rating task in which participants rated the overall brightness of scenes, and a brightness search task in which participants counted the number of bright areas in scenes.

For the brightness rating task, we found that meaning better explained the spatial distribution of attention than image salience. This result was observed both overall and when the correlation between meaning and image salience was statistically controlled, and held for early scene viewing, for short and long saccades, and using center-biased and unbiased meaning and saliency maps. For the brightness search task, using center-biased meaning and saliency maps, there were no differences between meaning and salience overall or when controlling for their correlation. However, the center-biased maps did not capture the fact that during the search task, the center bias in attention was greatly attenuated because attention was distributed much more uniformly over the scenes. Meaning and saliency maps with center bias over-weight scene centers and ignore scene peripheries, opposite to the attention maps actually observed. When the attention maps were analyzed using meaning and saliency maps that did not include center bias, the results were similar to those of the brightness rating task: meaning explained the variance in attention better than salience both overall and after statistically controlling for the correlation between meaning and salience. This pattern held for short and long saccades, and for the first saccade.

Overall, the results provide strong evidence that the meaning of a scene plays an important role in guiding attention through real-world scenes even when meaning is irrelevant and image salience is relevant to the task. Converging evidence across two viewing tasks that focused on an image property related to image salience showed that meaning accounted for more variance in attentional guidance than salience, and critically, that when the correlation between meaning and salience was controlled, only meaning accounted for significant unique variance. These results indicate that the guidance of attention by meaning is not restricted to viewing tasks that focus on encoding the meaning of the scene, strongly suggesting a fundamental role of meaning in attentional guidance in scenes.

Although the main pattern of results was clear and generally consistent across the two tasks, a few points are worth additional comment. First, our results suggest that tasks can differ in the degree to which center bias is present. Here, center bias was much greater when judging overall scene brightness than when searching for bright scene regions. These differences in center bias for the rating and search tasks likely occurred due to differences in the requirements of the tasks. The rating task simply required participants to rate the overall brightness of scenes, so there was no particular reason for viewers to direct attention away from the centers and to the peripheries of the scenes. In comparison, the search task required participants to count individual bright regions, many of which appeared away from the scene centers and in the peripheries. This resulted in fewer central fixations and more peripheral fixations in the brightness search task than the brightness rating task. Because there were more peripheral fixations in the search task, the center-biased meaning and saliency maps did not have the same predictive power to capture the relationship between meaning, salience, and attention as they did for the brightness rating task. Indeed, for this reason neither meaning nor saliency maps did a particularly good job of predicting attention when center-bias was included in the maps. However, when the center bias was removed from the two prediction maps, meaning maps were significantly better than saliency maps in accounting for attention.

The difference between the center-biased and unbiased maps was also evident in the analysis focusing on the earliest eye movements. According to the “salience first” hypothesis, we should have seen an initial bias of attention toward salient regions followed by a shift to meaningful regions. In our prior studies, we instead observed that meaning guided attention from the very first eye movement (Henderson & Hayes, 2018, Henderson et al., 2018). In the present study, when center-bias was included in the meaning and saliency maps in the brightness search task, meaning initially guided attention in the first eye movement, but there was a tendency for salience to take over for a few saccades before meaning again dominated. This pattern might offer some small support for salience first. However, as noted, viewers were much less likely to attend to scene centers and more likely to move their eyes to the edges of the scenes in the brightness search task. When the unbiased maps were used in the search task analysis, the trend from meaning to salience over the first few fixations was not observed. At best, then, there is a hint that when the viewer’s task is explicitly to find and count salient scene regions, they may be slightly more biased early on to attend to regions that are more salient. However, this result is weak at best given it appeared only in the third fixation and disappeared in the unbiased map analysis. Overall, even in a task that explicitly focused on salience and in which meaning was completely irrelevant, meaning played a stronger role in attentional guidance from the very beginning of viewing.

The type of meaning studied in the current work is what we refer to as context-free meaning, in that it is based on ratings of the recognizability and informativeness of isolated scene patches shown to raters independently of the scenes from which they are derived and independently of any task or goal besides the rating itself. Other types of meaning may be of interest in future studies. For example, we can consider contextualized meaning in which meaning is determined based on how important a scene patch is with respect to its global scene context. Additionally, the role of task may affect meaning as well. For example, meaning within a scene may change depending on a viewer's current tasks or goals. Because meaning can be defined in so many ways, it is necessary that we understand how these variants influence attentional guidance. The meaning map approach provides a method for pursuing these important questions.

Conclusion

We investigated the relative importance of meaning and image salience on attentional guidance in scenes using tasks that do not require semantic analysis and in which salience plays a critical role. Overall, the results strongly suggested that viewers can't help but attend to meaning (Greene & Fei-Fei, 2014). These findings are most consistent with cognitive control theories of scene viewing in which attentional priority is assigned to scene regions based on semantic properties rather than image properties.

References

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Psychonomic Bulletin and Review, 23*, 1794-1801. doi:10.3758/s13423-016-1035-4

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision, 15*(5), 1-22. doi:10.1167/15.5.9

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103*(1), 62-70. doi:10.1037/h0036799

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Royal Statistical Society B, 57*(1), 289-300.

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixaitons during free viewing. *Journal of Vision, 14*(13), 3-3.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*(1), 55-69. doi:10.1109/TIP.2012.2210727

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology and perception in art*. Oxford, England.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*(9), 898-914. doi:10.1.1.79.1508

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14-14. doi:10.1167/14.1.14

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188-194. doi:10.1016/j.tics.2005.02.009

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(6), 49-63. doi:10.1167/3.1.6

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16*(4), 219-222. doi:https://doi.org/10.1111/j.1467-8721.2007.00507.x

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences, 21*(1), 15-23. doi:http://dx.doi.org/10.1016/j.tics.2016.11.003

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. v. Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 537-562): Elsevier Ltd.

Henderson, J. M., Hayes, T. R. Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports, 8(13504).* doi: 10.1038/s41598-018-31894-5

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1-58). New York, NY, US: Psychology Press.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behavior, 1*, 743-747. doi:10.1038/s41562-017-0208-0

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision, 18*(6), 1-18.

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

Henderson, J. M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, *10*(5), 438-443.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin and Review, 16*(5), 850-856. doi:10.3758/PBR.16.5.850

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews, 2*(3), 1-11.

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).

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

Land, M. F., & Hayhoe, M. (2001). In what ways to eye movements contribute to everyday activities? *Vision Research, 41*(25-26), 3559-3565. doi:https://doi.org/10.1016/S0042-6989(01)00102-X

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics, 2*(11), 547-552.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205-231. doi:https://doi.org/10.1016/j.visres.2004.07.042

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*(4), 605-617. doi: https://doi.org/10.1016/j.neuron.2007.01.018

Nuthmann, A., & Henderson, J. M. (2010). Object based attentional selection in scene viewing. *Journal of Vision, 10(8): 20*, 1-19.

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

Rahman, S., & Bruce, N. (2015). Visual saliency prediction and evaluation across different perceptual tasks. *PlosOne*. doi:10.1371/journal.pone.0138053

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2016). Task and context determine where you look. *Journal of Vision, 7*(16), 1-20. doi:10.1167/7.14.16

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), 5-5.

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), 766-786. doi:10.1037/0033-295X.113.4.766

Yarbus, A. L. (1967). Eye movements during perception of complex objects *Eye Movements and Vision* (pp. 171-211). Boston, MA: Springer.

Chapter 3: The Role of Meaning in Attentional Guidance During Free-Viewing of Real-world Scenes

The following chapter consists of a manuscript that is published at *Acta Psychologica*.

Abstract

In real-world vision, humans prioritize the most relevant visual information at the expense of other information via attentional selection. The current study sought to understand the role of semantic features and image features on attentional selection during free viewing of real-world scenes. We compared the ability of meaning maps generated from ratings of isolated, context-free image patches and saliency maps generated from the Graph-Based Visual Saliency model to predict the spatial distribution of attention in scenes as measured by eye movements. Additionally, we introduce new contextualized meaning maps in which scene patches were rated based upon how informative or recognizable they were in the context of the scene from which they derived. We found that both context-free and contextualized meaning explained significantly more of the overall variance in the spatial distribution of attention than image salience. Furthermore, meaning explained early attention to a significantly greater extent than image salience, contrary to predictions of the 'saliency first' hypothesis. Finally, both context-free and contextualized meaning predicted attention equivalently. These results support theories in which meaning plays a dominant role in attentional guidance during free viewing of real-world scenes.

During real-world scene viewing, we are constantly inundated by visual information competing for our attention. It is therefore important to understand how we prioritize and guide attention to important objects and elements within a scene. However, the exact mechanism by which the human brain prioritizes one aspect of a visual scene over another for analysis remains unclear.

A substantial amount of research on attentional guidance in scenes has focused on image-based guidance models in which the salience of basic image features within a scene are used to control attentional guidance (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Harel, Koch, & Perona, 2006; Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1987; Parkhurst, Law, & Niebur, 2002). Image-based saliency models are popular because they are both computationally tractable and neurobiologically plausible (Henderson, 2007, 2017). At the same time, it is also well established that attentional guidance in scenes is influenced by semantic content (Henderson, 2007). For example, viewers attend to semantically informative scene regions (Antes, 1974; Buswell, 1935; Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Wu, Wick, & Pomplun, 2014; Yarbus, 1967), and to scene regions that are meaningful in the context of the current task (Castelhano, Mack, & Henderson, 2009; Einhäuser, Rutishauser, & Koch, 2008; Foulsham & Underwood, 2007; Hayhoe & Ballard, 2014; Neider & Zelinsky, 2006; Rothkopf, Ballard, & Hayhoe, 2007; Tatler, Hayhoe, Land, & Ballard, 2011; Torralba, Oliva, Castelhano, & Henderson, 2006; Turano, Geruschat, & Baker, 2003; Yarbus, 1967). We note that although there have been relevant attempts to integrate higher-level features into saliency maps (Chen & Zelinsky, 2019, Navalpakkam & Itti, 2005; Torralba, Oliva, Castelhano, & Henderson, 2006), these types of models continue to place much of the explanatory weight on the concept of salience, with cognitive representations serving only to modulate the influence of salience on attention.

It has been difficult to directly compare the influences of image salience and meaning on attentional guidance in scenes, because saliency maps represent the spatial distribution of salience across a scene in a way that has been challenging to reproduce for scene semantics. Given this challenge, studies of meaning-based guidance have typically focused on manipulations of one or at most a small number of specific scene regions or objects that do not allow a direct comparison of image salience and semantic informativeness across the entire scene (Brockmole & Henderson, 2008; De Graef, Christiaens, & d'Ydewalle, 1990; Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009).

To address this challenge, Henderson and Hayes (2017) introduced *meaning maps* as a semantic analog of saliency maps. Specifically, meaning maps were designed to capture the spatial distribution of semantic features in a scene in the same format that saliency maps use to capture the spatial distribution of image features. The key idea of a meaning map is that it represents the spatial distribution of semantic informativeness over a scene in the same format as a saliency map represents the spatial distribution of image salience. Inspired by two classic scene viewing studies (Antes, 1974; Mackworth & Morandi, 1967), meaning maps are created using crowd-sourced ratings given by large numbers of naïve subjects. These subjects rate the meaningfulness of individual scene patches taken from dense arrays of objectively defined circular overlapping patches at two spatial scales (Figure 3.1). Meaning maps are then constructed for each scene by averaging these ratings and smoothing the results (Figure 3.2). Meaning maps represent the spatial distribution of meaning across the scene, providing a means for directly comparing meaning and image salience and their relationships with attentional guidance. Research based on meaning maps has shown that meaning is a better predictor of attentional guidance than image salience across several active viewing tasks including scene memorization, aesthetic judgment (Henderson & Hayes, 2017, 2018), and scene description (Henderson, Hayes, Rehrig, & Ferreira, 2018).

Because previous viewing tasks (i.e., memorization, aesthetic judgment, scene description) comparing saliency maps and meaning maps may have drawn on semantic analysis, it is possible that they biased viewers to attend to meaning over image salience. In contrast, in many studies that have investigated image salience, the focus has been on the free viewing of scenes in which no specific task is imposed on viewers (Itti, Koch, & Niebur, 1998; Parkhurst et al., 2002). Furthermore, saliency models are typically benchmarked using free viewing (Bylinskii, Judd, Borji, Itti, Durand, Oliva, & Torralba, 2015; Itti et al., 1998; Parkhurst et al., 2002). One major goal of the current study was therefore to extend the investigation of meaning maps and saliency maps to free viewing in order to compare the influences of meaning and saliency on attention under benchmark viewing conditions. Specifically, we used a free viewing task in which participants freely viewed scenes with no experimenter-defined task. We hypothesized that if our past studies biased attention toward meaning by their viewing tasks, and if free viewing is by comparison meaning-neutral because it introduces no top-down task biases (Einhäuser, Rutishauser, & Koch, 2008; Parkhurst et al., 2002), then we should observe an advantage of saliency over meaning in the free viewing task. On the other hand, if the meaning advantage we have observed in prior studies is a general phenomenon, then we should continue to see it in the free viewing task.

The present study also provided us the opportunity to investigate a secondary question. In meaning map research to date, meaning maps were generated based on informativeness and recognizability ratings of isolated scene patches and thus were context-free (Henderson & Hayes, 2017, 2018). However, the meaning of an object or local element is often influenced by the scene context in which that element appears (Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978; Spotorno, Tatler, & Faure, 2013; Võ & Henderson, 2009). Therefore, it could be that meaning in the context of the scene (which we will refer to as contextualized meaning) is more related than context-free meaning to the distribution of attention in a scene. To examine this hypothesis, in the current study we generated new contextualized meaning maps and directly compared the relationships of our previous context-free meaning maps and the new contextualized meaning maps with the spatial distribution of attention during real-world scene viewing.

In summary, the current work sought to replicate and extend our prior research in two ways. First, we used a free viewing task in which participants viewed real-world scenes as they naturally would in their daily lives. The free viewing task does not introduce any particular requirement to attend to semantic features, and so provides an unbiased test of meaning versus image salience. Second, we introduced the concept of contextualized meaning maps in which scene patches were rated in the context of the scenes from which they came. Contextualized meaning maps were compared to the original context-free meaning maps to investigate whether contextualized meaning provides any additional advantage over context-free meaning in predicting attentional guidance.

Method

Eye-tracking

Participants. Thirty-two University of California, Davis, undergraduate students with normal to corrected-to-normal vision participated in the experiment (24 females, average age =

20.91). All participants were naïve to the purpose of the study and provided verbal consent. The eye movement data were inspected for excessive artifacts due to blinks or loss of calibration. Following Henderson and Hayes (2017), we averaged the percent signal ([number of good samples / total number of samples] x 100) for each trial and participant using custom MATLAB code. The percent signal for each trial was then averaged for each participant and compared to an *a priori* 75% criterion for signal. Outlier removal was then conducted by trial and participant. If a trial had less than 75% signal, it was excluded from analysis. Furthermore, if a participant's average percent signal was less than 75%, that entire participant was excluded from analysis. In total, no individual trials were excluded based on these criteria. Because two participants had lower than 75% signal, their data were excluded from analyses, resulting in a total of 30 participants/datasets analyzed. The number of participants used in the current study (N = 30) was derived from previous meaning map studies using 30 participants (Henderson et al., 2018; Peacock et al., 2019).

Apparatus. Eye movements were recorded using an EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01° rms) sampling at 1000 Hz (SR Research, 2010b). Participants sat 85 cm away from a 21" monitor, so that scenes subtended approximately 26.5° x 20° of visual angle at 1024x768 pixels. Head movements were minimized by using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software (SR Research, 2010a). Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s^2; SR Research, 2010b). Eye movement data were imported offline into Matlab using the EDFConverter tool. The first fixation, always located at the center of the display as a result of the pretrial fixation marker, was eliminated from analysis. Additionally, fixations that landed off the screen, and any fixations that were less than 50ms and greater than 1500ms were eliminated as outliers. Occasionally, saccade amplitudes are not segmented correctly by EyeLink's standard algorithm, resulting in large values. Given this, saccade amplitudes > 25° were also excluded. Fixations corresponding to these saccades were included as long as they met the other exclusion criteria. This outlier removal process resulted in loss of 5.84% of the data across all subjects.

Stimuli. Stimuli consisted of 20 digitized photographs (1024x768 pixels) of indoor and outdoor real-world scenes. Scenes were luminance matched across the scene set by converting the RGB image of the scene to LAB space and scaling the luminance channel of all scenes from 0 to 1. Luminance matching was done to ensure that there were no overly bright or dark scenes in the experiment and does not change the relative ranking of salience within a scene. All instruction, calibration, and response screens were luminance matched to the average luminance ($M = 0.45$) of the scenes.

Procedure. Before starting the experiment, participants completed two practice trials in which they were familiarized with the task. Here, participants were instructed that a real-world scene would appear on the screen for 8 seconds. During this time, they were instructed to view each scene, naturally, as they would in their daily lives. Given the free viewing nature of this task, we did not require participants to provide any responses.

After the practice trials, a 13-point calibration procedure was performed to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error of less than 0.99°. Presentation of each scene was preceded by a drift correction procedure, and the eyetracker was recalibrated when the calibration was not accurate.

Each participant viewed all 20 scene stimuli during the task. Scenes were presented in a randomized order for each participant.

Map Creation

Context-free meaning maps. For this study we used a subset of the meaning maps created by Henderson and Hayes (2017). To create those maps, scene-patch ratings were performed by 84 participants on Amazon Mechanical Turk. Participants were recruited from the United States, had a hit approval rate of 99% and 500 hits approved, and were allowed to participate in the study only once. Participants were paid $0.50 per assignment, and all participants provided informed consent. Rating stimuli were 20 digitized (1,024 × 768 pixels) photographs of real-world scenes depicting a variety of indoor and outdoor environments used in the eyetracking portion of the experiment. Each scene was decomposed into a series of partially overlapping (tiled) circular patches at two spatial scales (Figure 3.1). The full patch stimulus set consisted of 6,000 unique fine patches (87-pixel diameter) and 2,160 unique coarse patches (205-pixel diameter), for a total of 8,160 scene patches.

Each participant rated 300 random patches extracted from 20 scenes. Participants were instructed to assess the meaningfulness of each patch based on how informative or recognizable it was. They were first given examples of two low-meaning and two high-meaning scene patches, to make sure they understood the rating task, and then they rated the meaningfulness of scene patches on a 6-point Likert scale (very low, low, somewhat low, somewhat high, high, very high). Patches were presented in random order and without scene context, so ratings were based on context-free judgments. Each unique patch was rated three times by three independent raters for a total of 19,480 ratings. However, due to the high degree of overlap across patches, each patch contained rating information from 27 independent raters for each fine patch and 63 independent raters for each coarse patch. Meaning maps were generated from the ratings by averaging, smoothing, and then combining fine and coarse maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The average rating maps were then smoothed using thin-plate spline interpolation (fit using the thinplateinterp method in MATLAB; MathWorks, Natick, MA). Finally, the smoothed maps were combined using a simple average. This procedure was used to create a meaning map for each scene.

We previously estimated the optimal meaning-map grid density for each patch size by simulating the recovery of known image properties (i.e., luminance, edge density, and entropy as reported in Henderson and Hayes 2018). Here we briefly summarize this procedure with respect to luminance; application to other scene properties and procedural details can be found in the original report. The first step in the recovery simulation was to generate the ground-truth luminance image for each scene for a given patch size, which sets an upper limit on the luminance resolution that can be recovered. Then the patch-density grid (simulating patch ratings) was systematically varied from 50 to 1,000 patches (fine patches) and 40 to 200 (coarse patches), and recovery of the ground truth was performed for each potential grid. Using this method, simulated recovery of known scene properties suggested that the underlying known property could be recovered well (98% of the variance explained) using the fine and coarse spatial scales with patch overlap adopted for rating.

Finally, we added a center bias to the meaning maps. The tendency to fixate centrally is a behavioral phenomenon that occurs during scene viewing, and modeling this center bias is necessary to understand visual behavior (Clarke & Tatler, 2014). Given center bias in viewing, most saliency models contain center bias in their maps to enhance prediction accuracy, including the Graph-based Visual Saliency (GBVS) model used in this study (Harel et al., 2006). Since meaning maps do not naturally include a center bias, we added the GBVS center bias so that the centers of the saliency and meaning maps were equally weighted. Note that alternatively we could delete the center bias from GBVS maps, but removing center bias from GBVS changes the assumptions of that model. To create meaning maps with center-bias, we applied a multiplicative center bias operation to the meaning maps using the center bias present in the GBVS saliency maps. To do so, we inverted the 'invCenterBias.mat' (i.e., inverted the inverse) included in the GBVS package as an estimate of center bias. From here, we multiplied the resulting center bias and the raw meaning maps to create meaning maps with center bias.

Contextualized meaning maps. Contextualized meaning maps were generated using the identical method as the context-free meaning maps (Henderson and Hayes, 2017) with the following exceptions. For contextualized maps, we instructed participants to rate how 'meaningful' a patch was based on how informative or recognizable it was in the context of the larger scene (Figure 3.1). Additionally, for each rating, the patch was circled in green in the context scene. Other than these changes, the rating methods were identical. Importantly, the patches were identical to those used for context-free mapping, allowing direct comparison of the meaning mapping methods. In their raw form without center bias added, the resulting contextualized maps were significantly correlated with the context free maps, ($M = 0.67$, $SD = 0.09$): $t(19) = 34.69$, $p < 0.001$, 95% CI = [0.63, 0.71]. This correlation increased with center bias applied ($M = 0.88$, $SD = 0.05$): $t(19) = 76.59$, $p < 0.001$, 95% CI = [0.85, 0.90].

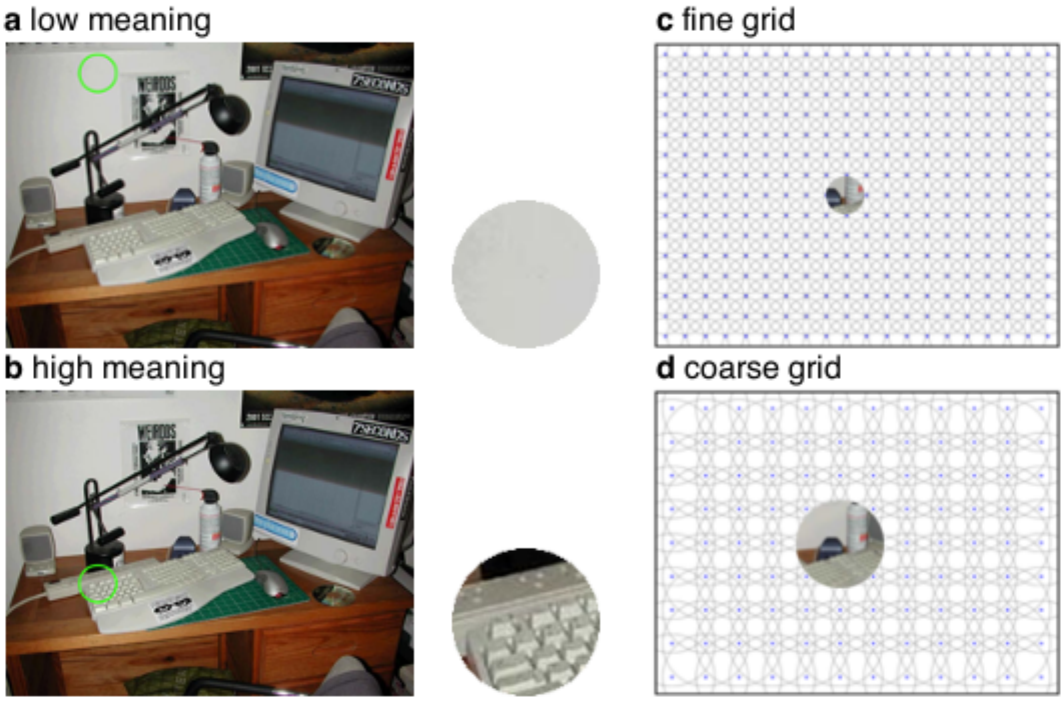


Figure 3.1. Rating Patch Examples. Examples of a low meaning fine patch (a) and a high meaning fine patch (b) shown alongside the scene from which each patch derived. Patch locations are circled in green in each scene. Example patches and their grids for a fine grid (c) and a coarse grid (d).

Saliency maps. Saliency maps for each scene were generated using the Graph-Based Visual Saliency (GBVS) toolbox with default settings (Harel et al., 2006). GBVS is a prominent saliency model that combines maps of low-level image features to create saliency maps (Figure 3.2).

Fixation density maps. Fixation density maps were generated from the eye movement data as described in Henderson and Hayes (2017). A fixation frequency matrix based on the locations (*x,y* coordinates) of all fixations was generated across participants for each scene. A Gaussian low-pass filter with a circular boundary and a cutoff frequency of −6dB (a window size of approximately 2° of visual angle) was applied to each matrix to account for foveal acuity and eyetracker error. The Gaussian low-pass function is from the MIT Saliency Benchmark code (https://github.com/cvzoya/saliency/blob/master/code_forMetrics/antonioGaussian.m).

Histogram matching. Following Henderson and Hayes (2017), meaning and saliency maps were normalized to a common scale using image histogram matching with the fixation density map for each scene serving as the reference image for the corresponding meaning and saliency maps. Image histogram matching is desirable because it normalizes an input image to a reference image, ensuring that the distribution of "power" in the two images is similar. In this study, we normalized both the saliency and meaning maps to the ground-truth fixation density maps so we could directly compare the meaning and saliency maps. Image histogram matching was accomplished by using the Matlab function 'imhistmatch' from the Image Processing Toolbox.

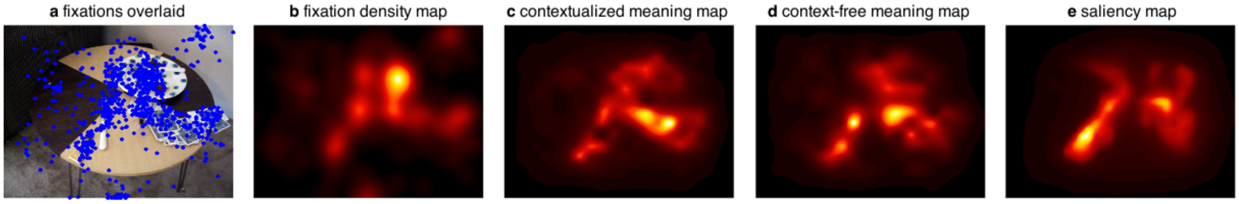


Figure 3.2. Map Examples. The panels show an example scene overlaid with fixation locations (a), the fixation density map (b), the contextualized (c) and context-free meaning maps (d), and the GBVS saliency map (e) for the example scene.

Results

Whole Scene Analyses

We used linear (i.e., Pearson) correlation (Bylinskii, Judd, Oliva, Torralba, & Durand, 2019) to test the degree to which the two prediction maps (meaning and saliency) accounted for the variance in the fixation density maps. There are many ways in which the prediction maps can be compared to the fixation density maps, and no method is perfect (Bylinskii et al., 2019). We chose linear correlation because it is sensitive to small differences in predictors, makes relatively few assumptions, is intuitive, can be visualized, generally balances the various positives and negatives of different analysis approaches, and allows us to tease apart variance due to salience and meaning (Bylinskii et al., 2019). It also provides a basis for comparing against our prior meaning map results.

To calculate the Pearson correlation, we used the CC.m function from the MIT saliency benchmark code set (https://github.com/cvzoya/saliency/blob/master/code_forMetrics/CC.m). The CC.m function has been used to evaluate the various metrics included in the MIT saliency benchmark (Bylinskii et al., 2019). The function works by first normalizing the to-be-correlated maps. It then converts the two-dimensional map arrays to one-dimensional vectors and correlates these vectors. The output of the function is then squared to calculate the shared variance explained by meaning and saliency. We used two-tailed, paired t-tests to statistically test the relative ability of the prediction maps (saliency, context-free meaning, and contextualized meaning) to predict the fixation density maps. We also report 95% confidence intervals (CI) that indicate the range of values that are 95% certain to contain the true mean of the population.

Because our primary research question concerned the ability of meaning and salience to independently account for variance in fixations, we used semi-partial correlations. Semi-partial correlations capture the amount of total variance in the fixation density maps that can be accounted for with the residuals from each of the predictors (meaning and salience) after removing the correlation between those predictors. In other words, the semi-partial correlations indicate the total variance in the fixation density maps that can be accounted for by the meaning-independent variance in salience and the salience-independent variance in meaning. Two-tailed one-sample t-tests were used to compare the unique variance in attention explained by each map type against zero. The same 95% CI accompany these results.

Context-free meaning vs. image salience. For the squared linear correlation, context-free meaning explained 39% of the variance in fixation density (*M* = 0.39, *SD* = 0.14) and image salience explained 24% of the variance (*M* = 0.24, *SD* = 0.14), *t*(19) = 7.08, *p* < 0.001, 95% CI = [0.10, 0.19] (Figure 3.3). For the semi-partial correlations, context-free meaning explained a unique 16% of the variance in fixation density controlling for salience (*M* = 0.16, *SD* = 0.07): *t*(19) = 9.52, *p* < 0.001, 95% CI = [0.13, 0.20], whereas salience uniquely explained only a unique 2% of the variance in fixation density controlling for meaning (*M* = 0.02, *SD* = 0.03): *t*(19) = 2.37, *p* = 0.03, 95% CI = [0.002, 0.03].

These results replicate and extend to a free viewing task the previous context-free meaning map results from memorization, aesthetic judgment, and scene description tasks (Henderson & Hayes, 2017; Henderson et al., 2018). Once again, meaning was a better predictor of the spatial distribution of attention than image salience.

Contextualized meaning vs. context-free meaning. For our secondary question, we investigated whether contextualized and context-free meaning maps would produce similar results. For the squared linear correlation, contextualized meaning explained 40% of the variance in fixation density (*M* = 0.40, *SD* = 0.14) and context-free meaning explained 39% of the variance in fixation density (*M* = 0.39, *SD* = 0.14), *t*(19) = 1.44, *p* = 0.17, 95% CI = [−0.007, 0.04] (Figure 3.3). When the variance explained by context-free meaning was statistically controlled, contextualized meaning uniquely explained 4% of the variance in fixation density (*M* = 0.04, *SD* = 0.03): *t*(19) = 5.74, *p* < 0.001, 95% CI = [0.03, 0.05]. When the variance explained by contextualized meaning was statistically controlled, context-free meaning uniquely explained 2% of the variance in fixation density (*M* = 0.02, *SD* = 0.02): *t*(19) = 4.27, *p* = 0.0004, 95% CI = [0.02, 0.03]. These results demonstrate that the two types of meaning largely account for the same variance in the distributions of fixations over scenes.

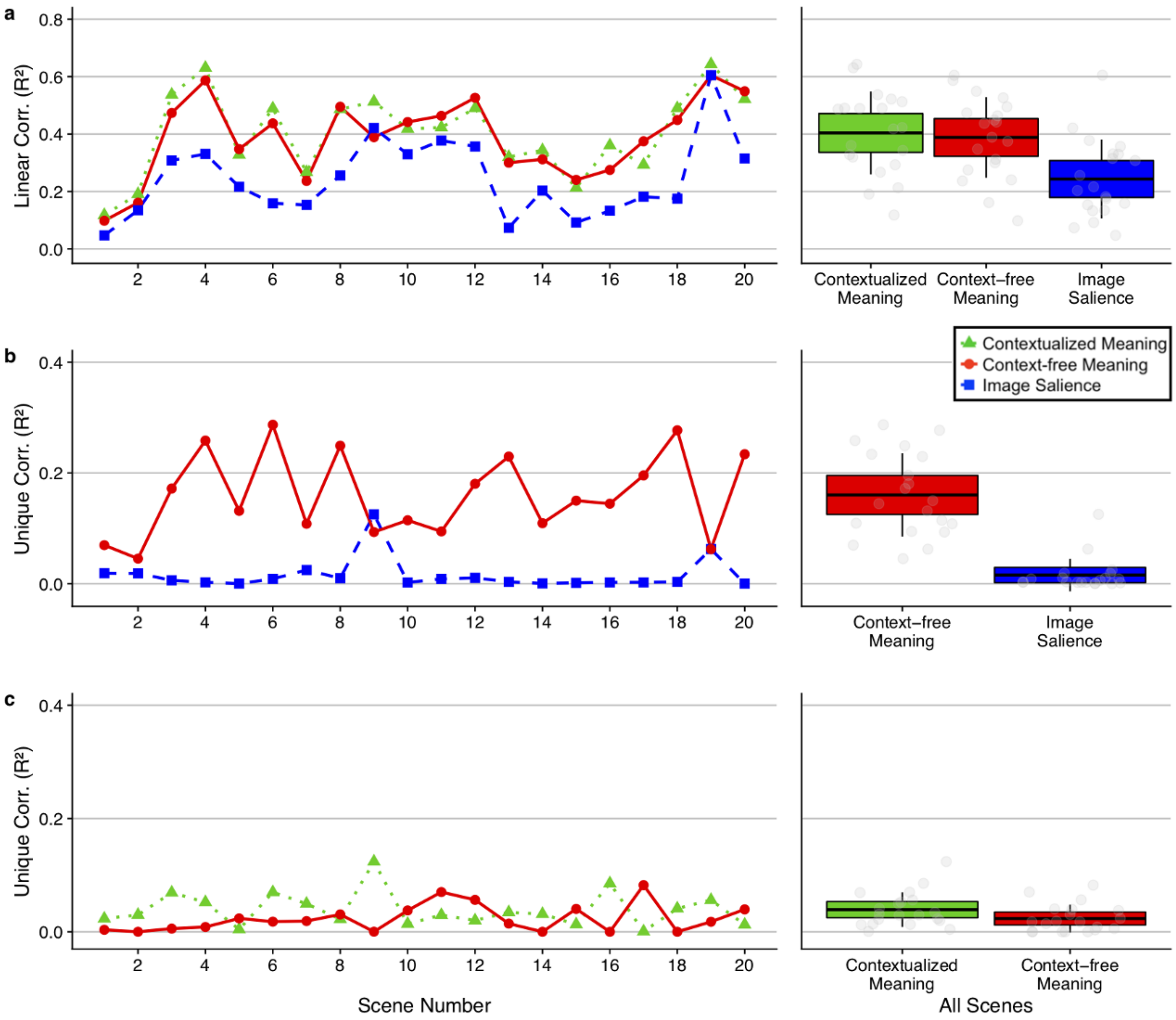


Figure 3.3. Squared Linear and Semi-Partial Correlations by Scene. Line plots show the squared linear (a) and semi-partial correlations (b, c) between the fixation density maps, contextualized meaning (green triangles), context-free meaning (red circles), and image salience (blue squares). The scatter plots show the grand mean (black horizontal line), 95% confidence intervals (colored boxes), and one standard deviation (black vertical line), for contextualized meaning, context-free meaning, and salience across all 20 scenes for each analysis.

Ordinal Fixation Analyses

It has been suggested that when a scene first appears, attention might initially be guided by image salience, with meaning playing a larger role as viewing unfolds (Anderson & Donk, 2017; Anderson et al., 2015; Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999). This hypothesis predicts that the correlation between image salience and fixation density maps should be greater for earlier than later fixations, with salience dominating meaning in the earliest fixations. Alternatively, it could be that meaning guides attention from scene onset due to rapid gist apprehension (Oliva & Torralba, 2006; Potter et al., 2014) and the use of schema to activate memory representations of where likely objects will be located in the scene (Henderson, 2003; Henderson & Hollingworth, 1999; Torralba et al., 2006) for attentional prioritization. This hypothesis predicts that meaning should account for attentional guidance at the earliest moments of scene viewing. To test these competing hypotheses in the free viewing task, we conducted an ordinal fixation analysis for the first three fixations, in which density maps were generated for each sequential fixation for each scene. The analyses focused on the first three of these fixations (1st fixation, 2nd fixation, and 3rd fixation) and proceeded as in the main analyses, with p-values corrected for multiple comparisons using the Bonferroni correction.

Context-free meaning vs. image salience. For the squared linear correlations, context-free meaning accounted for 38%, 31%, and 20% of the variance in the first three fixations whereas salience accounted for 10%, 15%, and 11% of this variance (Figure 3.4), with all three ordinal fixation meaning versus salience comparisons significant (fixation 1: $t(19) = 7.71$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.20, 0.36]; fixation 2: $t(19) = 5.48$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.10, 0.22]; fixation 3: $t(19) = 3.06$, Bonferroni-corrected $p = 0.02$, 95% CI = [0.03, 0.15]). For the semi-partial correlations, meaning accounted for 30%, 19%, and 12% of the unique variance in the first three fixations (fixation 1: $t(19) = 8.92$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.23, 0.37]; fixation 2: $t(19) = 7.06$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.14, 0.25]; fixation 3: $t(19) = 4.65$, Bonferroni-corrected $p = 0.001$, 95% CI = [0.07, 0.17]) and image salience accounted for 2%, 3%, and 3% of this variance (fixation 1: $t(19) = 3.00$, Bonferroni-corrected $p = 0.04$, 95% CI = [0.005, 0.03]; fixation 2: $t(19)$

= 4.51, Bonferroni-corrected p = 0.001, 95% CI = [0.02, 0.05]; fixation 3: t(19) = 4.06, Bonferroni-corrected p < 0.004, 95% CI = [0.02, 0.05]), (Figure 3.4).[1]

Overall, the ordinal fixation analysis comparing context-free meaning and image salience showed that meaning was a better predictor than salience early in free scene viewing, contrary to the salience first hypothesis. This effect is consistent with and extends our past work on early influences of meaning (Henderson & Hayes, 2017; Henderson et al., 2018; Peacock et al., 2019).

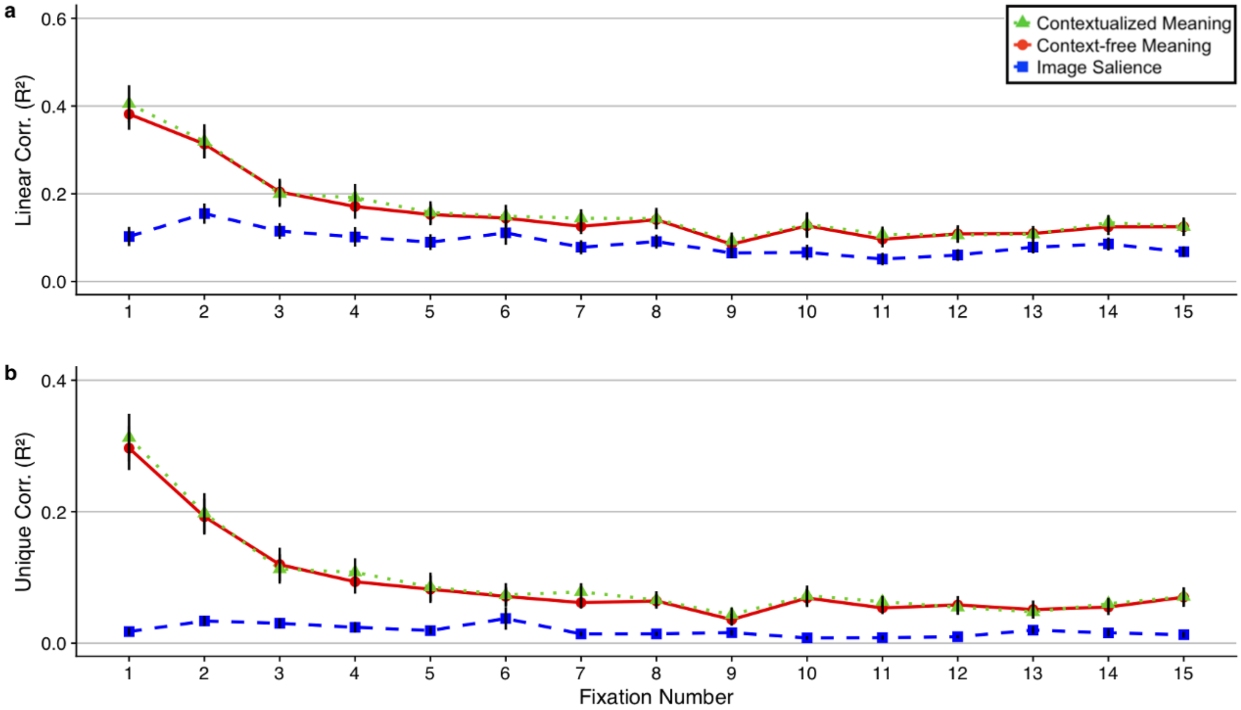


Figure 3.4. Ordinal Fixation Analysis. The line plots show the squared linear correlations (a) and semi-partial correlations (b) between the fixation density maps, contextualized meaning (green triangles), context-free meaning (red circle), and image salience (blue square) as a function of fixation number collapsed across scenes. Analyses focused on the first three fixations and fifteen fixations are displayed for comparison. Error bars represent the standard error of the mean.

General Discussion

The current study was designed to assess several questions related to understanding the roles of meaning and image salience in predicting attentional guidance in real-world scenes. The

[1] As in the main analyses, contextualized and context-free meaning produced identical results across the first three fixations (all Bonferroni corrected ps > 0.05).

main question was whether the previously observed advantage for meaning over image salience would extend to a free viewing task that does not impose any specific top-down task constraints. Earlier comparisons of meaning and image salience have used explicit viewing tasks (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019). However, it is common to use free viewing with no explicit task in the saliency literature, and indeed saliency model benchmarks are based on free viewing (Bylinskii et al., 2015), so it was important to extend the previous results to free viewing. The present results using free viewing were consistent with previous studies that have compared meaning and image salience: meaning accounted for significantly more of the overall variance in the spatial distribution of fixation density than salience. Furthermore, when the variance explained by meaning was statistically controlled, salience explained no more of the unique variance in fixation density, but when the variance explained by salience was controlled, meaning continued to explain substantial unique variance. In addition, contrary to the idea that image salience plays a major role during early scene viewing, these results held for the earliest fixations.[2] This pattern of results replicates the previous findings and extends them to free viewing. We note that when we weighted the fixations by duration to produce duration-weighted fixation density maps, all of the results held, consistent with Henderson and Hayes (2018).

The current results are consistent with the results of recent research suggesting that meaning continues to strongly influence attentional guidance in scenes even when meaning is not directly relevant to the viewer's task. Peacock et al. (2019) used tasks in which participants were

[2] We note that in the ordinal fixation analyses, the correlations decline as a function of fixation. This is an artifact of using center-biased maps to predict fixations. When using prediction maps that do not contain center bias, this artificial bump in meaning is not observed, as has been shown in previous meaning map studies (Peacock et al., 2019).

cued to either count bright patches within a scene or rate scenes for their overall brightness. Despite the fact that salience was task-relevant and meaning was task-irrelevant, meaning continued to guide attention. The convergence of the findings reported by Peacock et al. (2019) and the current study suggests that the influence of meaning over salience on attention is robust and not readily influenced by tasks designed to minimize attention to meaning nor tasks designed to reduce top-down influences of task on attention.

To date, meaning maps have been based on context-free meaning in the sense that the maps have been created from ratings of scene patches that are presented to raters without their scenes (Henderson & Hayes, 2017, 2018; Henderson et al., 2018). The present study expanded the concept of meaning maps to contextualized meaning generated from ratings of scene patches that are presented in the context of their scenes. The question was whether meaning maps that reflected local meaning assessed in the context of the overall scene would better predict fixation density than context-free meaning maps. The results showed that the contextualized meaning maps were significantly correlated with and predicted fixation density similarly to context-free meaning maps. This convergence suggests that our original results were not due to peculiarities of the specific way meaning ratings were obtained. This result also shows that the context-free meaning maps do not seem to be losing much critical semantic information despite the fact that sometimes only parts of large objects and scene regions are shown in the rated patches.

An interesting issue is why the contextualized and context-free meaning maps were similarly related to attention. One possibility is that local scene meaning ratings are not significantly affected by global scene context in the absence of object-scene inconsistencies. Indeed, studies using scenes containing local object manipulations find that objects contain greater meaning when they are inappropriate (versus appropriate) to the global context of the scene which influences attention (Henderson et al., 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009). The literature has also shown that scene gist (Oliva & Torralba, 2006; Potter et al., 2014) and schema representations guide our expectations of what local objects will appear in a scene given its global context (Henderson, 2003; Henderson & Hollingworth, 1999; Torralba et al., 2006). Given the importance of global scene context on attention, future conceptions of contextualized meaning maps will need to be made with scenes containing object-scene inconsistencies to fully predict how global scene context influences the meaning of local elements and attention to those elements.

There are a few other reasons why the contextualized and context-free meaning maps maps similarly predicted attention. The first is that the two maps were highly correlated with each other, and that the shared variance in meaning across the two map types did most of the work in guiding attention. The second is that the current study used a passive viewing task, whereas studies showing an effect of scene context on eye movements have used active tasks such as change blindness (Spotorno et al., 2013; Stirk & Underwood, 2007), memorization, and search (Henderson et al., 1999; Võ & Henderson, 2009). To better understand how global scene context influences attention to local scene elements, future studies may wish to use active viewing tasks in conjunction with scenes containing object-scene inconsistencies.

Conclusion

The current work used a free viewing task in scenes to investigate the relationships between meaning and image salience on attention, as operationalized by fixation density, without introducing additional top-down task biases. We found that meaning was more related to attention than image salience both when assessing the overall spatial distribution of attention and when focusing only on the early guidance of attention. Additionally, the concept of contextualized meaning maps was introduced and compared to previously used context-free meaning maps. Contextualized meaning maps capture the spatial distribution of semantic features based on how informative or recognizable scene patches are in the context of the scenes from which they derive. We found that contextualized and context-free meaning maps predicted attention equally well. In total, these findings show that meaning plays a dominant role in real-world attentional guidance in free viewing, with little influence from image salience.

References

Anderson, N. C., & Donk, M. (2017). Salient object changes influence overt attentional prioritization and object-based targeting in natural scenes. *PlosOne*. https://doi.org/:10.1371/journal.pone.0172132

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 1–22. https://doi.org/10.1167/15.5.9

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62–70. https://doi.org/10.1037/h0036799

Buswell, G. T. (1935). How people look at pictures: a study of the psychology and perception in art.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). Mit saliency benchmark. Retrieved from http://saliency.mit.edu/results_mit300.html

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What do different evaluation metrics tell us about saliency models?. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3), 740-757.

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision*, *14*(13), 3.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69. https://doi.org/10.1109/TIP.2012.2210727

Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: Effects of object–scene consistency. *Visual Cognition*, *16*(2-3), 375-390.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6-6.

Chen, Y., & Zelinsky, G. J. (2019). Is there a shape to the attention spotlight? Computing saliency over proto-objects predicts fixations during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(1), 139.

Clarke, A. D., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41-51.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317-329.

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), 1–19.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception?. *Perception*, *36*(8), 1123-1138.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14–14. https://doi.org/10.1167/14.1.14

Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology*, *24*(13), R622-R628.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*.

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

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in*

Cognitive Sciences, *7*(11), 498–504. https://doi.org/10.1016/j.tics.2003.09.006

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219-222.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences*, *21*(1), 15-23.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York, NY, US: Psychology Press.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*, 743–747. https://doi.org/10.1038/s41562-017-0208-0

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scenes: evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 1–18. https://doi.org/10.1089/jmf.2012.0243

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports*, *8*.

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

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850-856.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*(1), 161–169.

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).

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

Koch, C., & Ullman, S. (1987). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Matters of Intelligence*, *4*(4), 115–141. https://doi.org/10.1007/978-94-009-3833-5_5

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

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*(11), 547–552.

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

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614-621.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. In *Progress in Brain Research* (Vol. 155, pp. 23–36). Elsevier.

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. https://doi.org/10.1016/S0042-

6989(01)00250-4

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, and Psychophysics*. https://doi.org/10.3758/s13414-018-1607-7

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, and Psychophysics*, *76*(2), 270–279.

Rahman, S., & Bruce, N. (2015). Visual saliency prediction and evaluation across different perceptual tasks. *PlosOne*. https://doi.org/10.1371/journal.pone.0138053

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 16-16.

Spotorno, S., Tatler, B. W., & Faure, S. (2013). Semantic consistency versus perceptual salience in visual scenes: Findings from change detection. *Acta Psychologica*, *142*(2), 168-176.

Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision*, *7*(10), 3-3.

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), 5-5.

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.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, *43*(3), 333-346.

Võ, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision2*, *9*(3), 1–15. https://doi.org/10.1167/9.3.24

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 54.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171-211). Springer, Boston, MA.

Chapter 4: Meaning and Expected Surfaces Combine to Guide Attention During Visual Search in Scenes

The following chapter consists of a manuscript that is under review at

Journal of Vision.

Abstract

How do spatial constraints and meaningful scene regions interact to control overt attention during visual search for objects in real-world scenes? To answer this question, we combined novel surface maps of the likely locations of target objects with maps of the spatial distribution of scene semantic content. The surface maps captured likely target surfaces as continuous probabilities. Meaning was represented by meaning maps highlighting the distribution of semantic content in local scene regions and objects. Attention was indexed by eye movements during search for target objects that varied in the likelihood they would appear on specific surfaces. The interaction between surface maps and meaning maps was analyzed to test whether fixations were directed to meaningful scene regions on target-related surfaces. Overall, meaningful scene regions were more likely to be fixated if they appeared on target-related surfaces than if they appeared on target-unrelated surfaces. These findings suggest that the visual system prioritizes meaningful scene regions on target-related surfaces during visual search in scenes.

Due to processing limitations, the visual system must select and prioritize only the most relevant visual information from moment to moment during real world visual search. This selection process is accomplished via eye movements. However, it is unclear why some aspects of the world are prioritized over others for analysis. Previous work has found influences of target features (Malcolm & Henderson, 2009; Navalpakkam & Itti, 2005; Vickery et al., 2005; Wolfe & Horowitz, 2017; Zelinsky, 2008), scene context/spatial constraint (Castelhano & Witherspoon, 2016; Neider & Zelinsky, 2006; Pereira & Castelhano, 2014, 2019), memory (Draschkow et al., 2014; Võ & Wolfe, 2013), and interactions among these sources (Bahle et al., 2018; Bahle & Hollingworth, 2019; Castelhano & Heaven, 2010; Ehinger et al., 2009; Malcolm & Henderson, 2010; Torralba et al., 2006; Wolfe & Horowitz, 2017; Zelinsky et al., 2006; Zelinsky et al., 2020). Although recent work has independently demonstrated that the visual system may also prioritize scene regions high in meaning for fixation during search (Hayes & Henderson, 2019; Peacock et al., under review), it is unknown how scene meaning interacts with other known sources of search guidance. The present study therefore aimed to understand how meaning interacts with one of these known sources of guidance, spatial constraint (i.e., scene regions likely to contain the search target; Brockmole & Henderson, 2006; Brockmole & Võ, 2010; Ehinger et al., 2009; Neider & Zelinsky, 2006; Pereira & Castelhano, 2019; Torralba et al., 2006). To investigate this question, we developed continuously graded surface maps representing the likely locations of a search target, and paired these with meaning maps representing semantic densities in scenes (Henderson & Hayes, 2017).

Surfaces as constraints on search in scenes

The semantic representation of an object in the context of a given scene guides attention during visual search (Biederman, Mezzanotte, & Rabinowitz, 1982; Henderson et al., 2007;

Henderson, Malcolm, & Schandl, 2009; Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978). Viewers searching for an object, such as a pillow, will first fixate semantically appropriate locations (e.g., bed) over inappropriate locations (e.g., table), suggesting that these expected spatial constraints efficiently direct attention to task- and semantically-relevant information (Brockmole & Henderson, 2006; Brockmole & Võ, 2010; Ehinger et al., 2009; Henderson et al., 1999; Loftus & Mackworth, 1978; Neider & Zelinsky, 2006; Pereira & Castelhano, 2019; Torralba et al., 2006).

Spatial constraint has been modeled in different ways. Torralba et al. (2006) successfully predicted the likely locations participants would search for an object in a scene using horizontal bands that represented where a given target object was most likely to be located given the global physical structure of that scene. These bands were learned from a large number of scene exemplars. An issue with this approach, however, is that the predicted spatial constraints were coarse and were not tied to surfaces or objects in a particular scene. Indeed, when participants in Torralba and colleagues' study searched for coffee mugs, they sometimes looked at specific surfaces associated with coffee mugs outside of the region predicted by the horizontal band.

This was remedied by Pereira & Castelhano (2019) who operationalized spatial constraint as the upper (e.g., ceilings, walls), middle (e.g., countertops, tables), and lower (e.g., floors) horizontal surface regions associated with target objects within a scene. A limitation of the Pereira & Castelhano (2019) approach, however, was that their method generated binary spatial constraints: only surfaces within a given horizontal surface region were taken to be predictive of target object location whereas other scene regions were not predictive. Furthermore, all of the surfaces within a given horizontal band were equally predictive of target object location. However, it seems likely that there is a continuous distribution of surface constraints for many target objects (e.g., garbage bins might be more likely to appear on the sidewalk than in the road even though both sidewalks and roads appear in lower scene regions). In the present study we offer an approach to spatial constraint based on scene surfaces that provides a continuum of constraint.

To generate continuous surface maps, we first parsed scenes into their constituent elements (objects and surfaces) and had a group of participants assign labels to those elements. We then asked a separate group of participants to rank the labels of the elements in each scene based upon the degree to which those elements could serve as the location for each of three search targets (garbage bins, drinking glasses, and paintings). For example, for a drinking glass, "table" would likely be ranked higher than "ceiling". Scene elements were ranked in a generic scene-independent manner: we presented the targets and surface elements using labels without a visual scene (Figure 4.2). The element rankings were then mapped back onto scenes to capture target-surface relationships in a continuous fashion. Because surfaces in the foreground occlude background surfaces, we used image-computable three-dimensional depth information (Laina et al., 2016) to account for occlusion. Finally, we accounted for the tendency of objects to extend above the tops of surfaces by generating a target object height constant for each object and its highly ranked surface elements. The height constant reflected how tall a given target object would appear at a given depth. The resulting surface maps continuously represented the likely locations of search target objects in scenes while taking into account depth from the viewer.

Meaning as a constraint on search in scenes

Meaning maps represent the continuous spatial distribution of local semantic densities in scenes (Henderson & Hayes, 2017), allowing direct study of how semantics influence attention during visual search. Recent studies show that meaning predicts eye movements during letter search (Hayes & Henderson, 2019) and during common object search (Peacock et al., under review). Meaning maps provide a framework to test how the spatial distribution of semantic densities interact with other known sources of guidance (e.g., spatial constraint) during visual search. Despite the utility of meaning maps, visual search models have not yet incorporated meaning maps as a source of guidance.

Combining Meaning and Surfaces

Spatial constraint interacts with image salience to guide attention during visual search (Ehinger et al., 2009; Torralba et al., 2006). Given the correlation between image salience and meaning in real-world scenes (Elazary & Itti, 2008; Henderson, 2003; Henderson et al., 2007; Henderson & Hayes, 2017, 2018; Rehrig et al., 2020; Tatler et al., 2011) and the finding that meaning accounts for most if not all of the unique variance in predicting eye fixations when the intercorrelation between meaning and saliency is controlled (Hayes & Henderson, 2019; Henderson & Hayes, 2017, 2018; Peacock et al., under review, 2019b, 2019a, 2020; Rehrig et al., 2020), spatial constraint might also interact with meaning to guide eye movements.

In previous research, spatial constraint has been represented using image-based bands that are not tied to a specific scene surfaces (Torralba et al., 2006), or to surfaces in a binary fashion (Pereira & Castelhano, 2019). Here we represented spatial constraint related to surfaces as a continuum associated with a given target object. Given that meaning predicts attention during visual search (Hayes & Henderson, 2019; Peacock et al., under review) and that eye movements are restricted to meaningful information on surfaces associated with target objects (Castelhano & Heaven, 2011; Castelhano & Henderson, 2003; Castelhano & Witherspoon, 2016; Pereira & Castelhano, 2019), we examined the combined role of target-related surfaces and meaningful scene regions on eye movements during visual search for objects in real-world scenes (Figure 4.1).

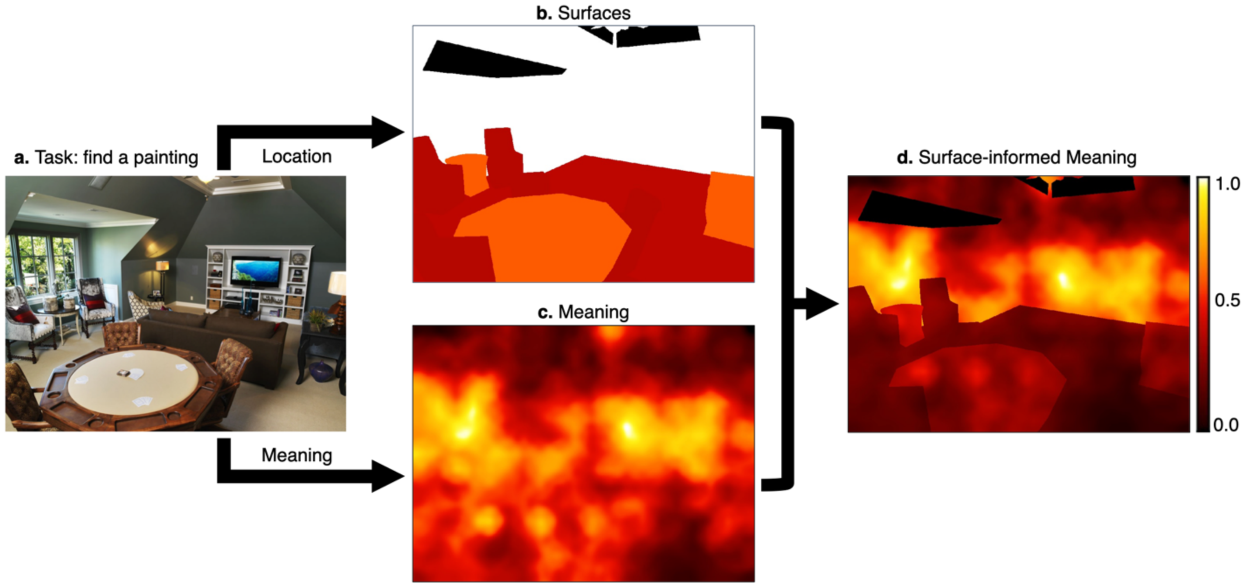


Figure 4.1. Schematic of surface map model. If the goal is to find a painting in a media room (a), the probability that a painting will appear on one surface (walls) over another surface (floors) will drive attention to the more probable region (b). Analogously, meaningful (informative) scene regions are more likely to guide attention than those that are less meaningful (c). Surfaces may inform meaning in that meaningful features on highly predictive surfaces are more likely to be prioritized for attention (white) than those on non-predictive surfaces (black) (d).

Methods

Eyetracking

Participants. The sample size was set with an *a priori* stopping rule of 30 acceptable participants based on prior experiments using these methods (Peacock et al., 2019b, 2019a, 2020). To reach 30 acceptable participants, 37 University of California, Davis, undergraduate students with normal to corrected-to-normal vision initially participated in the experiment (28 females, average age = 20.51). All participants were naïve to the purpose of the study and provided consent. Eye movement data from each participant were inspected for excessive artifacts due to blinks or loss of calibration. Following Henderson and Hayes (2017), we averaged the percent signal ([number of good samples / total number of samples] x 100) for each trial using custom MATLAB code. The percent signal across trials was averaged for each participant and compared to an *a priori* 75% criterion for signal. Overall, 0 participants were excluded based on this criterion of poor eyetracking quality. Individual trials that had less than 75% eyetracking signal were also excluded. Only 10 total trials (0.44% of the total data) were excluded based upon this criterion.

Participants were also excluded if they did not correctly do the task. The percentage of target absent trials in which each participant erroneously indicated there were targets (even though the scene was target absent) was calculated. If this occurred on over 25% of trials, that participant was excluded, resulting in removal of 7 participants. These criteria resulted in analyses based on a total of 30 acceptable participants as per the stopping rule.

Apparatus. Eye movements were recorded using an EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01° rms) sampling at 1000 Hz (SR Research, 2010b). Participants sat 85 cm away from a 21" monitor, so that the scenes subtended approximately 26.5° x 20° of visual angle at 1024x768 pixels. Head movements were minimized using a chin and forehead rest. Viewing of the scenes was binocular, but eye movements were recorded from the right eye. The experiment was controlled using SR Research Experiment Builder software (SR Research 2010a). Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and $9500°/s^2$; SR Research, 2010b). Resulting segmented eye movement data were imported offline into Matlab using the EDFConverter tool. The first fixation, always located at the center of the display as a result of the pretrial fixation marker, was eliminated from analysis. Given that we were interested in search activity and not target decision processes, we only analyzed data from target absent trials.

Fixations that landed off the screen, and any fixations that were less than 50ms or greater than 1500ms were eliminated as outliers. Occasionally, saccade amplitudes are not segmented correctly by EyeLink's standard algorithm, resulting in large values. Given this, saccade amplitudes > 25° were also excluded. Fixations corresponding to these saccades were included as long as they met the other exclusion criteria. This outlier removal process resulted in loss of 2.22% of the data.

Stimuli. 105 digitized photographs (1024 x 768 pixels) of indoor and outdoor real-world scenes were selected for this study, with 35 scenes dedicated to each target object (i.e., 35 scenes for garbage bins, 35 scenes for drinking glasses, 35 scenes for paintings). Ten scenes from each target set were target present and 25 scenes from each set were target absent. Target present scenes had one or more target objects in the scene and served as fillers to ensure that participants explored each scene fully. Data analysis focused on target absent scenes so that influences of the target itself on eye movements would be excluded. All instruction, calibration, and response screens were luminance matched to the average luminance (*M* = 0.43 L) of the scenes.

Procedure. Each run of the experiment consisted of six practice trials and 105 randomized experimental trials split into three counterbalanced target object blocks (35 trials in each block). In each trial, a central fixation was shown on the screen for 400ms to orient participants to the center of the screen where a word cue would appear. Then, a word cue was presented for 800ms indicating the search target for that scene. Following the word cue, the central fixation cross re-appeared for 400ms prior to the search phase of the experiment. The search scene was then presented for 10s (Torralba et al., 2006). While the search scene was present on the screen, participants were instructed to count the number of target objects in the scene and to press "Enter" on a keyboard when all of the objects were found. Possible answers were either “zero targets” or “one or more targets”. Participants were instructed that there could be multiple targets present in the scene to encourage them to fully explore the scene. At the end of each trial, participants used the button box to indicate how many targets were present in the scene. Two practice trials (one target present and one target absent) were administered before the experiment for each target object (a total of six practice trials), providing participants an opportunity to ask any questions they had before beginning the experimental trials.

After the practice trials, a 9-point calibration procedure was performed to map the participants’ eye positions to screen locations. Successful calibration required an average error of less than 0.49° and a maximum error of 0.99°. In order to maintain calibration throughout the experiment, a calibration check screen preceded each trial. If the calibration error exceeded 1.00°, the eye tracker was recalibrated.

Surface Maps

Participants. Ninety-six University of California, Davis, undergraduate students who did not participate in the eye-tracking study participated across three survey studies (garbage bin N = 34, drinking glass N = 32, painting N = 30). All participants were naïve to the purpose of the study and provided informed consent. The sample size was set with an *a priori* stopping rule of 30 acceptable participants for each rating study (90 participants total after the *a priori* participant exclusion criterion was applied). Participants were removed if they were guessing: if a participant did not include either of the top two rankings from the rest of the participants in their study in more than 25% of trials, they were excluded from analysis. This resulted in minimal participant loss (4 participants from the garbage bin task, 2 participants from the drinking glass task, and 0 participants from the painting task).

Scene labeling and segmentation. All scene elements that were present in any of the 105 scenes were first identified to form a set of all possible scene element labels. Elements were defined as objects (e.g., pencil), groups of densely overlapping objects (e.g., pencils), and surfaces (e.g., desk, wall) within a scene. Then, from this global set of labels, each label was mapped to an individual element or elements within each scene using the Computer Vision Annotation Tool (CVAT, https://github.com/opencv/cvat) (Figure 4.2a).

Labels corresponding to the segmented elements were used to generate surface rankings for each target in each scene. Only unique and singular labels from the segmented scenes were used for the ranking task for each scene. Any repeated or plural labels were subsequently re-added during analysis and given the same weight as the unique and singular labels, respectively. Labels that were synonyms of the unique singular label were also excluded from the ranking task. For the target "painting", the following labels were excluded: drawing, drawings, picture, pictures, painting, paintings, poster, posters. For the target "drinking glass", the following labels were excluded: glass, glasses, cup, cups, mug, mugs. For the target, "garbage bin", the following labels were excluded: trashcan, dumpster, trash bin, bin.

Procedure. Separate on-line surveys were administered for each target object via Qualtrics. For example, for "drinking glass", participants were instructed to indicate the degree to which each element label named a surface that a drinking glass could be placed upon. Participants were asked to drag and drop the labels into a provided box on the computer screen, and to rank order them based upon how likely a drinking glass would be to appear on that given surface (Figure 4.2b). Participants were instructed not to rank (i.e., not to drag into the box) labels that were not surfaces a drinking glass would appear upon. Before beginning the survey, participants were given an example ranking question (Figure 4.2b). For drinking glass, the example labels were counter, plant, and chair. Participants were instructed that drinking glasses could be found on a counter and a chair. However, because drinking glasses are more likely to appear on a counter than a chair, counter should be ranked higher than chair. In this example, participants were told that plant should be left out of the box because drinking glasses do not appear on plants. The instructions for garbage bins and paintings were the same except the most likely surface in each example ranking question was modified. For garbage bins, “counter” was replaced with “floor” and for paintings, “counter” was replaced with “wall”.

For each target object, there were 35 ranking trials corresponding to the 35 scenes for that target object category, presented in a random order for each participant. The labels corresponding to a given scene were provided in a randomized order to the left of the ranking column (Figure 4.2b).

Generating surface weights. We first generated weights corresponding to each label’s ranking for each participant in each scene. To calculate each label’s weight, first the total number of labels that each participant ranked was summed for each scene. Then, a proportion was computed to serve as the ranking. If a label was placed first out of 21 ranked labels for a given scene, it would receive a participant-level weighting of 21/21 (Figure 4.2c). If a label was placed second out of 21 ranked labels, it would be given a participant-level weighting of 20/21. If a label was unranked, it would be given a participant-level weight of zero. If a given participant’s rankings for a given scene did not include one of the top two ranked labels from the rest of the participants for that scene, then that participant’s data for that scene was excluded. This resulted in loss of 4.29% of the data from the garbage bins, 1.91% of the data from the drinking glasses, and 4.29% of the data from the paintings. To compute the final weight for each label, we averaged each label's weight across participants. This process resulted in a single weight for each label corresponding to each element in each scene.

Eliminating small and non-predictive elements. Because our primary question asked whether target-related surfaces guide attention to meaningful scene regions on those surfaces, it was necessary to exclude smaller elements that were not predictive of target object location, but that were located on larger elements. For example, a spoon is a small element that might be found on a table, but because a drinking glass is not likely to appear on a spoon, the spoon rating creates a "hole" in the table map. It was therefore necessary to exclude small elements that were also non-predictive.

To eliminate small elements, we compared the size of each element to a size threshold for each target object category [size = area of element in pixels / area of scene in pixels]. The size threshold was the mean size of the most predictive elements (i.e., elements with surface weights greater than or equal to 0.4) for each target object category: garbage threshold = 0.14, painting threshold = 0.19, glass threshold = 0.09 (Figure 4.2d). If a given element's size was less than the size threshold then it was tagged for possible deletion.

To eliminate non-predictive element ratings from predictive elements in a principled manner, we first ranked each element in descending order by scene based upon its surface weighting on the X axis and plotted the weighting (Figure 4.2d) on the Y axis, respectively. We then fit an exponential function [$y = e^{(-x)}$] to the weighting data (Figure 4.2d). Elements that were under the weight asymptote for a given scene were tagged for possible deletion.

If a given element was under both the weight and size thresholds for a given scene, it was excluded from the resulting surface map. However, if it was under one or the other but not both, it was included in the resulting surface map. This method allowed us to eliminate elements that were small but also non-predictive.

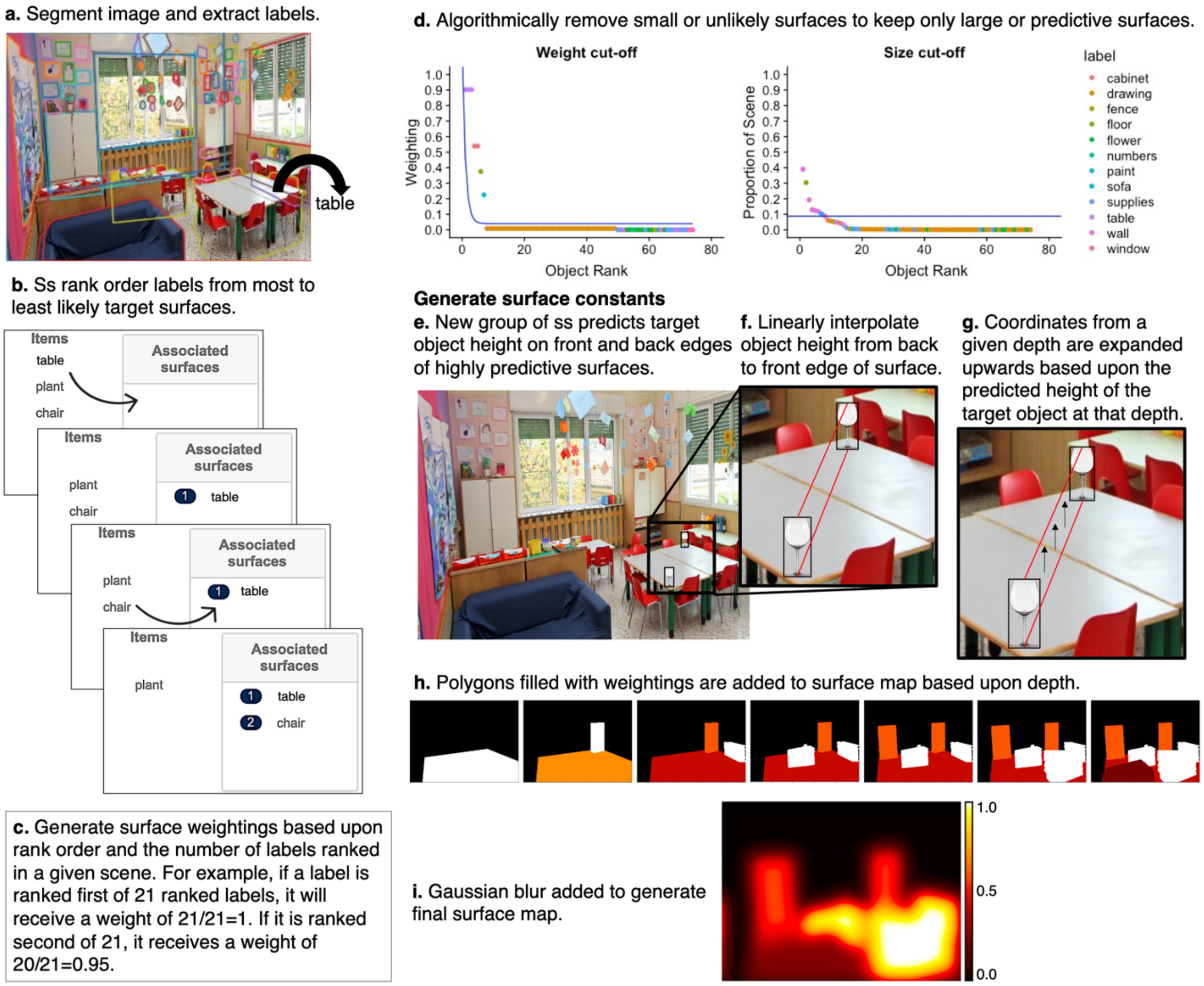
Above-surface constant. Because objects tend to extend in space above the surfaces/elements they sit on, we added a height constant to the most predictive horizontal support surfaces to account for the regions that target objects occupy above these surfaces.

To generate the value of the above-surface constant, seven undergraduate research assistants that were naïve to the purpose of the study indicated how tall an average sized target would appear on either the front or back edge of highly predictive surface elements (corresponding to labels weighted 0.5 or greater) in each scene (Figure 4.2e). We then estimated how tall a given target object would be from the back to the front of the surface elements using linear interpolation (Figure 4.2f). We separated the segmentation for a given surface element into 10 slices based upon the y dimension and expanded the coordinates based upon how tall the target object was estimated to be at that slice (Figure 4.2g). Both the expanded coordinates and the original coordinates were added to the resulting surface map as participants were predicted to look on and above predictive surfaces (Figure 4.2h).

Depth maps. Because surfaces in the foreground occlude background surfaces, we used image-computable depth maps (Laina et al., 2016) to account for occlusion of surface elements in the surface maps described below. Depth maps provide a measure of the predicted depth of each pixel within an image and therefore allowed us to estimate how deep a given surface element was within a scene. With this information, we were able to add deeper (and likely occluded) surfaces into a scene first and later add in closer (and likely non-occluded) elements.

Surface map generation. After finalizing the weights and constants for each surface element, we generated empty surface maps by first creating a 768 x 1024 array of zeros. We then replaced the existing values on the surface map with each element's weighting based upon that element's spatial location and depth relative to the other elements in the scene to account for foreground elements occluding background elements (Figure 4.2h). Here, elements were added from the back (deepest) to the front (shallowest) based upon each element's median depth generated from the depth maps described above (Laina et al., 2016). Constant values for elements corresponding to highly predictive surfaces were added at the same depth as the respective element. A Gaussian low-pass filter with a circular boundary and a cutoff frequency of −6dB (a window size of approximately 2° of visual angle) was applied to each map. The Gaussian low-pass function is from the MIT Saliency Benchmark code[3].

[3] https://github.com/cvzoya/saliency/blob/master/code_forMetrics/antonioGaussian.m



a. Segment image and extract labels.

b. Ss rank order labels from most to least likely target surfaces.

c. Generate surface weightings based upon rank order and the number of labels ranked in a given scene. For example, if a label is ranked first of 21 ranked labels, it will receive a weight of 21/21=1. If it is ranked second of 21, it receives a weight of 20/21=0.95.

d. Algorithmically remove small or unlikely surfaces to keep only large or predictive surfaces.

Generate surface constants

e. New group of ss predicts target object height on front and back edges of highly predictive surfaces.

f. Linearly interpolate object height from back to front edge of surface.

g. Coordinates from a given depth are expanded upwards based upon the predicted height of the target object at that depth.

h. Polygons filled with weightings are added to surface map based upon depth.

i. Gaussian blur added to generate final surface map.

Figure 4.2. Surface map generation. After images were segmented and labeled (a), participants ranked labels independent of scenes by how likely a given target object would be to appear on that surface (b). Surface weightings were then generated (c) and small / unlikely surfaces were removed. Surface constants were generated by linearly interpolating participant generated size predictions from the back to front edges of elements. Maps were made by adding polygons filled with weightings from the back/deepest scene region to the front of the scene (h). A gaussian blur was added to generate the final surface map (i).

Meaning Maps

We used the meaning map technique developed by Henderson and Hayes (2017) (see https://osf.io/654uh/ for code and instructions). To create meaning maps, scene-patch ratings were performed by 434 participants on Amazon Mechanical Turk. Participants were recruited from the United States, had a hit approval rate of 99% and 500 hits approved, and were allowed to participate in the study only once. Participants were paid $0.50 per assignment, and all participants provided informed consent. Rating stimuli were the same 105 digitized (1,024 × 768 pixels) photographs of real-world scenes used for the visual search task. Each scene was decomposed into a series of partially overlapping (tiled) circular patches at two spatial scales. The full patch stimulus set consisted of 31,500 unique fine patches (87-pixel diameter) and 11,340 unique coarse patches (205-pixel diameter), for a total of 42,840 scene patches. The optimal meaning-map grid density for each patch size was previously determined by simulating the recovery of known image properties as reported in Henderson & Hayes (2018).

Each participant rated 300 random patches extracted from 105 scenes. Participants were instructed to assess the meaningfulness of each patch based on how informative or recognizable it was. They were first given examples of two low-meaning and two high-meaning scene patches, to make sure they understood the rating task, and then they rated the meaningfulness of scene patches on a 6-point Likert scale (very low, low, somewhat low, somewhat high, high, very high). Patches were presented in random order and without scene context, so ratings were based on context-free judgments. Each unique patch was rated three times by independent raters for a total of 128,520 ratings. However, due to the large degree of overlap across patches, each patch contained rating information from 27 independent raters for each fine patch and 63 independent raters for each coarse patch. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The average fine and course rating maps were then combined into a single map using the simple average and a light Gaussian filter was applied using the MATLAB function ‘imgaussfilt.m’ set at 10.

Center Proximity Map

A center proximity map served as a global representation of how close each location in the scene image was from the scene center (Figure 4.4d). Specifically, it measured the inverted Euclidean distance from the center pixel of the scene to all other pixels in the scene image. The center proximity measure was used in the mixed-effects models described below to account for and control the role of center bias, the tendency to fixate centrally (Bindemann, 2010; Hayes & Henderson, 2021; Tatler, 2007; Tseng et al., 2009) (Figure 4.4d).

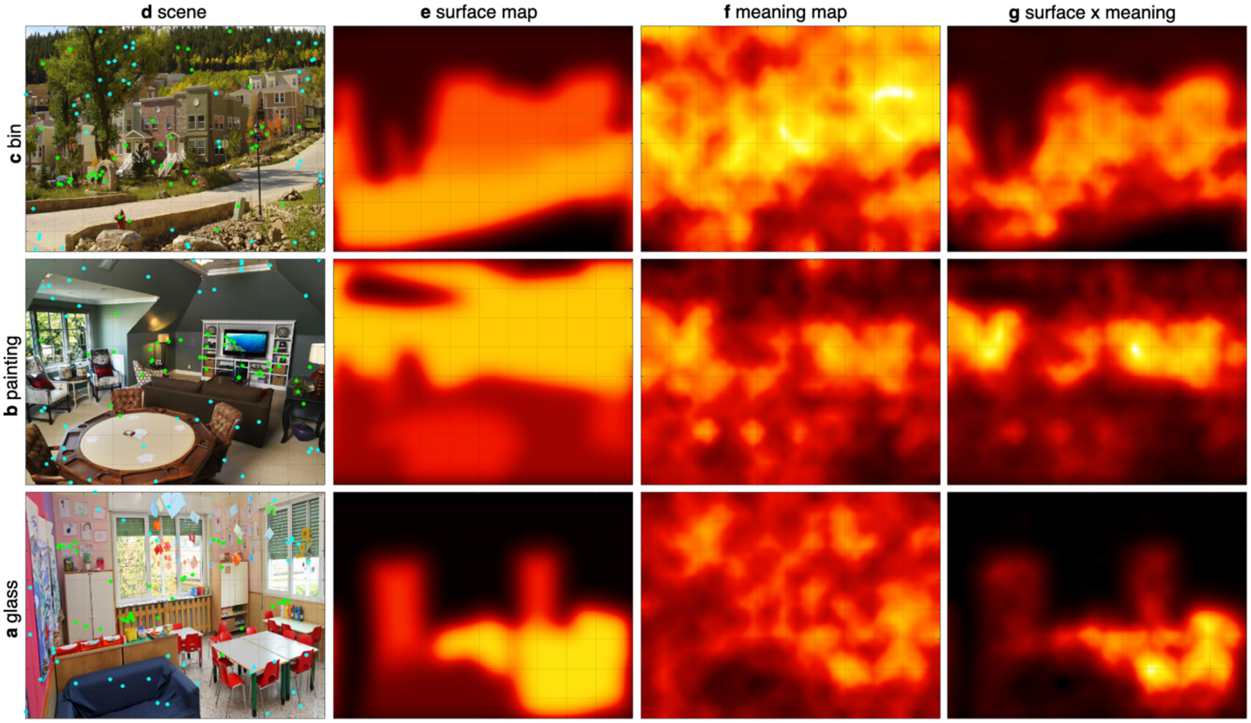


Figure 4.3. Map examples. The figure shows an example of each map type for drinking glasses (a), paintings (b), and garbage bins (c). Each column represents an example scene with fixated (green) vs. non-fixated (cyan) regions for a single participant (d), with each respective surface map (e) meaning map (f), and hypothesized visualization of the surface by meaning interaction.

Eyetracking Search Analysis

To test whether surfaces and meaning interact to predict fixated and non-fixated regions while also taking center proximity and scene-by-scene variation into account, we used a general linear mixed effects (GLME) model with the link logit ('binomial') distribution (Hayes &

Henderson, 2021; Nuthmann et al., 2017). We focused analyses on the eye movement data corresponding to target absent scenes since we were interested in search behavior with regard to expected target locations as opposed to actual target features. Before submitting the data to the GLME, we z-normalized surface maps and meaning maps within each target object category to a common scale. Analyses were conducted separately for each target object because each of the targets is found in different scene regions, and the surfaces they reside upon are different sizes (e.g., floor surfaces are much larger than countertops). The center proximity map was z-normalized as well.

For each fixation, we computed the mean map values by taking the average over a 3-degree window (113-pixels in diameter) around each fixation in the surface map (Figure 4.4b), meaning map (Figure 4.4c), and center proximity map (Figure 4.4d). To represent scene features that were not associated with overt attention for each participant, we randomly sampled an equal number of scene locations where each particular participant did not look in each scene they viewed. The only constraint for the random sampling of the non-fixated scene regions was that the non-fixated 3-degree windows could not overlap with any of the 3-degree windows of the fixated locations.

The dependent variable was whether a region was fixated or not. The fixed effects were the meaning values, the surface values, and the center proximity value. Although the primary effect of interest was the interaction between surfaces and meaning, we modeled the three-way interaction between surfaces, meaning, and center proximity to ensure that any effects were not due to center bias. Additionally, we included a random intercept of scene. Including a random intercept of participant did not account for any variance so this was excluded from each model. We hypothesized that both meaning and surfaces would influence probability of fixation, with highly meaningful scene regions appearing on highly predictive surfaces with the highest probability (Figure 4.4e).

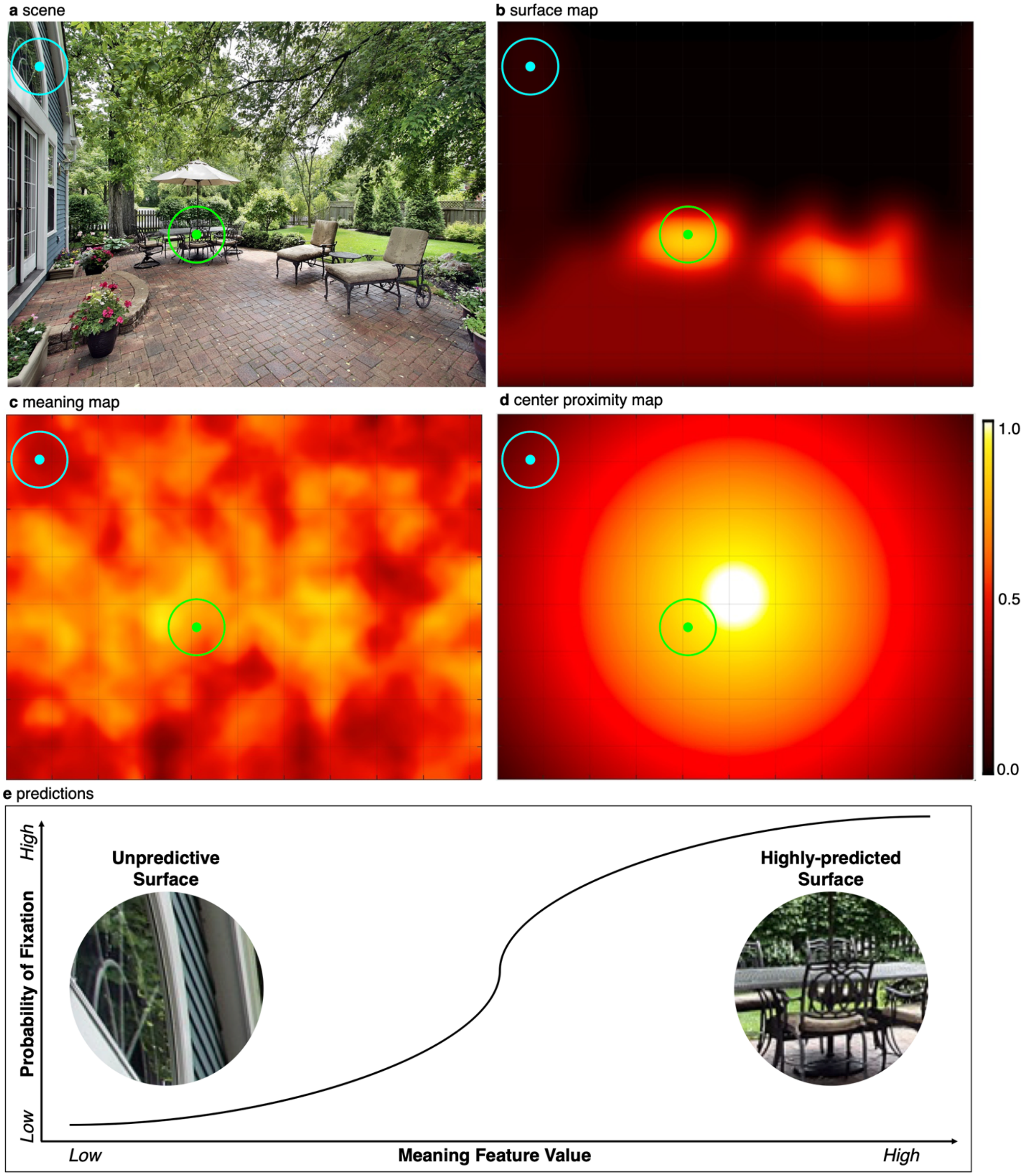


Figure 4.4. Analysis and predictions. The figure shows an example scene (a), surface map (b), meaning map (c), and center proximity map (d) with hypothetical fixated (green) versus non-fixated (cyan) windows. Predicted results (e) shows that meaningful scene regions have a higher probability of fixation if these regions overlap with highly predictive surfaces. If meaningful

scene regions do not overlap with highly predictive surfaces, these regions are less likely to be fixated.

Results

Our primary question asked whether fixations are directed to meaningful scene regions that occur on target-related surfaces during search in scenes. Figure 4.5 summarizes the primary data. The plots show that all three variables were related to fixations during search for all three targets, with fixations more likely to be directed to the scene centers, relevant surfaces, and meaningful regions. To analyze these data, we used the GLME model described above with fixed effects of meaning, surfaces, and center proximity predicting whether a region was fixated or not. The primary effect of interest was the surfaces by meaning interaction. We also modeled the three-way interaction between surfaces, meaning, and center proximity to control for the effect of center bias.

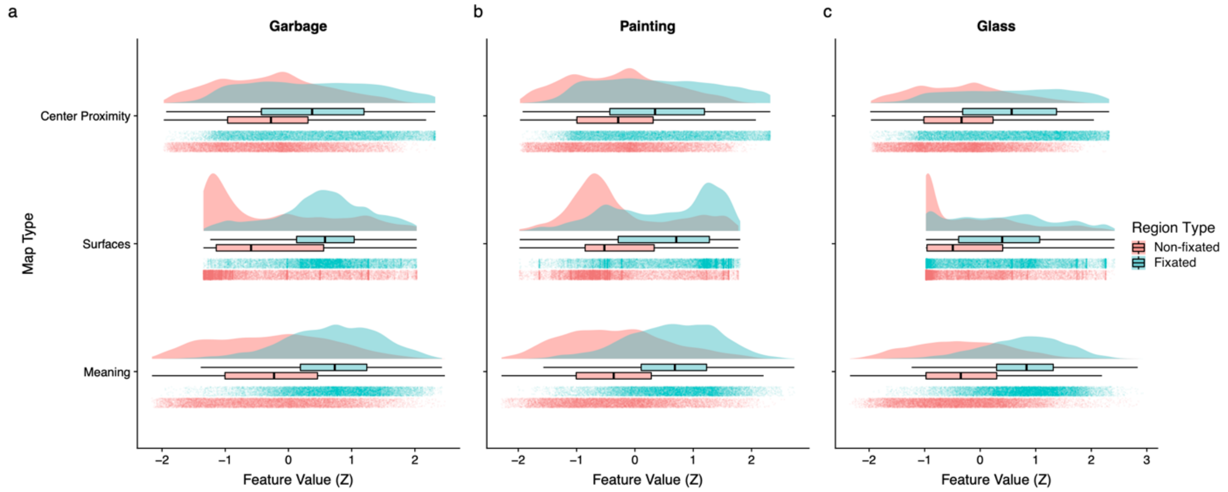


Figure 4.5. Summary plots of the raw eye movement data. Raincloud plots show the center proximity, surface, and meaning z-normalized feature values on fixated (blue) and non-fixated (pink) scene regions for garbage bins (a), paintings (b), and drinking glasses (c). For each box plot, the whiskers refer to the minimum (25% quartile – 1.5*interquartile range) and maximum (75% quartile + 1.5*interquartile range) feature values, the box refers to the 25% and 75% quantiles, and the central, vertical line refers to the median. Each dot corresponds to the average feature value for a given fixated or non-fixated window.

The GLME model results for meaning are visualized in Figure 4.6 and Table 4.1. For drinking glasses, there was a significant three-way interaction between meaning, surfaces, and center proximity; for garbage bins there was a marginal three-way interaction; for paintings there was no significant three-way interaction. For all three target objects there was a significant two-way interaction between meaning and surfaces, which was the primary interaction of interest.

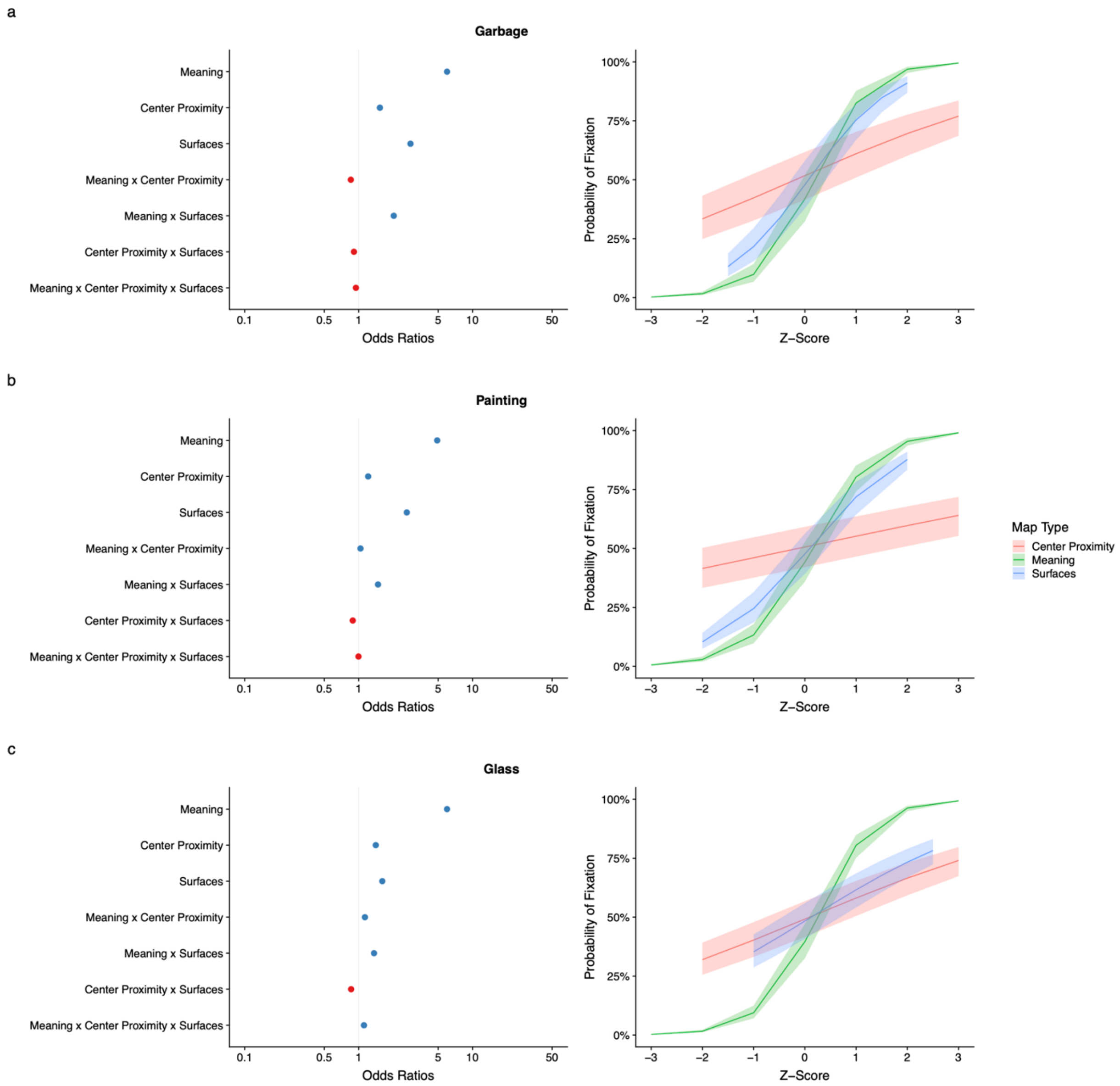


Figure 4.6. Model fits. The odds ratios (left column) and marginal effects (right column) for the garbage (a), painting (b), and drinking glass (c) models are shown. An odds ratio of 1 indicates that neither positive or negative values of a predictor are likely to occur with fixated regions. An

odds ratio of greater than 1 (blue) indicates that positive values of a predictor are more associated with fixated regions whereas an odds ratio of less than 1 (red) indicates that negative values of a predictor are associated with fixated regions. Marginal effects plots (right column) show the probability of fixation for each fixed effect as a function of z-score. Error bands reflect 95% confidence intervals.

Before interpreting the two-way interaction for drinking glasses and garbage bins, we examined the three-way interactions to ensure center proximity was not modulating the meaning by surface effects (Figure 4.7). If the meaning by surface interaction was driven by center proximity, we would expect high meaning and surface values to be fixated at scene centers due to scene-independent viewing biases with no surface by meaning interaction in scene peripheries. For all target objects, meaning values were more likely to be fixated if surface values were greater at scene centers (Figure 4.7a, 4.7d, 4.7g). However, this effect did not change as a function of center proximity: for fixations further from center (Figure 4.7b 4.7e, 4.7h) and in scene peripheries (Figure 4.7c, 4.7f, 4.7i), higher meaning regions were more likely to be fixated if the corresponding surface values were higher. The three-way interaction for garbage cans appears to be the result of the lack of an asymptote in the low probability surfaces (red curves in Figure 4.6) at high meaning values compared to the medium and high probability surfaces (blue and green curves respectively), which may have been due to fewer high-meaning regions on surfaces likely to contain garbage cans (e.g., floors). This result is consistent with the notion that target-related surfaces constrain eye movements to meaningful scene regions irrespective of scene independent viewing biases.

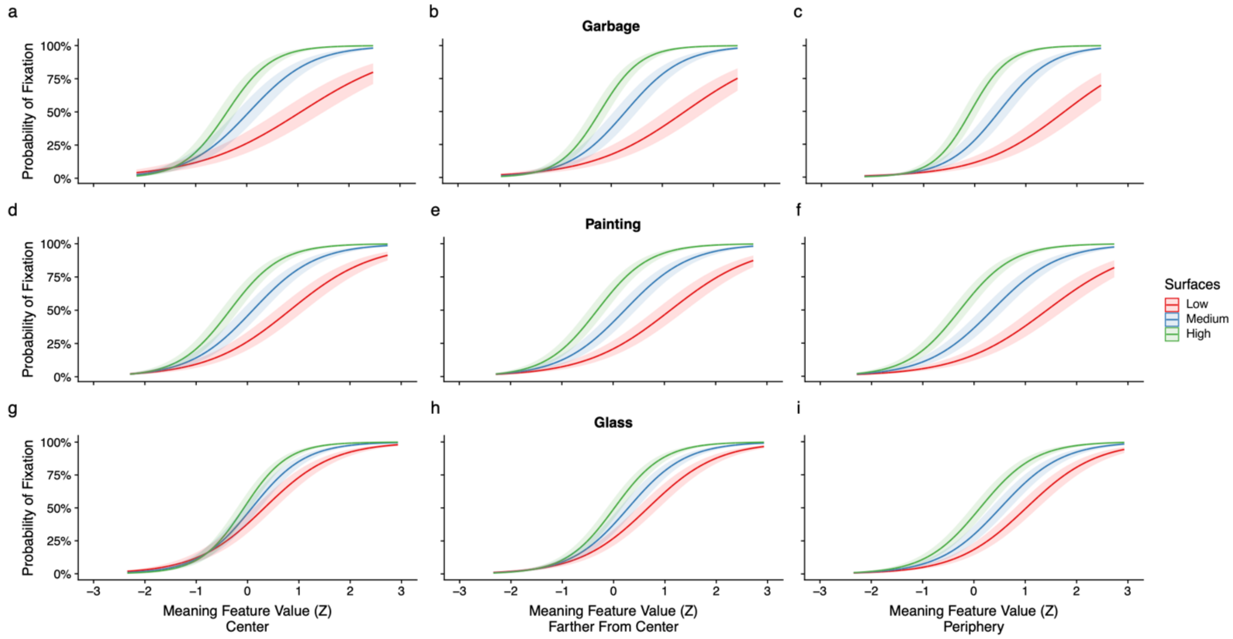


Figure 4.7. Three-way meaning x surfaces x center proximity interaction. This figure shows the probability that meaningful scene regions were fixated on surfaces that were not predictive of target object locations (red), moderately predictive (blue), and highly predictive of target location (green) at scene centers (a, d, g), farther from center (b, e, h), and in scene peripheries (c, f, i) for garbage bins (a, b, c), paintings (d, e, f), and drinking glasses (g, h, i). Error bands reflect 95% confidence intervals.

Table 4.1. Meaning x surface x center proximity GLME results for each target object. Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), z-values, and p-values (p) for each fixed effect, and standard deviations (SD) for the scene random effect.

Garbage

	Fixed Effects					Random Effects, SD
Predictors	β	95% CI	SE	Z-value	P	By-scene
Intercept	-0.51	[-0.93, -0.08]	0.21	-2.42	0.02	1.04
Meaning	1.79	[1.73, 1.85]	0.03	60.58	<0.001	
Center proximity	0.43	[0.40, 0.47]	0.02	22.98	<0.001	
Surfaces	1.05	[1.01, 1.09]	0.02	48.70	<0.001	
Meaning x center proximity	-0.16	[-0.20, -0.11]	0.02	-7.08	<0.001	
Meaning x surfaces	0.71	[0.66, 0.76]	0.03	27.61	<0.001	
Center proximity x surfaces	-0.10	[-0.14, -0.05]	0.02	-4.46	<0.001	
Meaning x center proximity x surfaces	-0.06	[-0.11, 0.001]	0.03	-2.01	0.05	

Painting

	Fixed Effects					Random Effects, SD
Predictors	β	95% CI	SE	Z-value	P	By-scene
Intercept	-0.36	[-0.72, -0.004]	0.18	-2.06	0.04	0.88
Meaning	1.59	[1.54, 1.64]	0.03	60.52	<0.001	
Center proximity	0.19	[0.15, 0.23]	0.02	9.83	<0.001	
Surfaces	0.97	[0.94, 1.01]	0.02	48.60	<0.001	
Meaning x center proximity	0.04	[-0.004, 0.08]	0.02	1.78	0.08	
Meaning x surfaces	0.39	[0.35, 0.44]	0.02	17.22	<0.001	
Center proximity x surfaces	-0.12	[-0.16, -0.08]	0.02	-5.92	<0.001	
Meaning x center proximity x surfaces	-0.003	[-0.05, 0.04]	0.02	-0.13	0.90	

Glass

	Fixed Effects					Random Effects, SD
Predictors	β	95% CI	SE	Z-value	P	By-scene
Intercept	-0.51	[-0.83, -0.19]	0.16	-3.27	0.001	0.78
Meaning	1.79	[1.74, 1.84]	0.03	69.35	<0.001	
Center proximity	0.35	[0.31, 0.38]	0.02	18.21	<0.001	
Surfaces	0.48	[0.44, 0.52]	0.02	25.54	<0.001	
Meaning x center proximity	0.13	[0.08, 0.17]	0.02	5.70	<0.001	
Meaning x surfaces	0.31	[0.26, 0.36]	0.03	12.31	<0.001	
Center proximity x surfaces	-0.15	[-0.19, -0.11]	0.02	-7.44	<0.001	
Meaning x center proximity x surfaces	0.11	[0.06, 0.16]	0.03	4.22	<0.001	

Discussion

The present study tested how spatial constraints related to the expected surfaces associated with a target object interact with meaningful scene regions to control eye movements during visual search in real-world scenes. To this end, we generated surface maps that represented the likely locations of three target objects (garbage bins, drinking glasses, and paintings). The surface maps took three-dimensional depth information into account and represented the likely locations of target objects probabilistically. Surface maps were combined with meaning maps representing the distribution of semantic content across each scene (Henderson & Hayes, 2017). We then examined whether surfaces and meaning interacted to account for fixations in a visual search task in which participants searched for the target objects. The results showed that both likely target surfaces and meaningful regions were more likely to be fixated, with meaningful regions within likely target surfaces most likely to be fixated. This effect persisted regardless of how close to center a given fixation was, suggesting that the effect was not due to scene-independent viewing biases. Our findings provide the first evidence that the visual system constrains search for real-world objects in scenes to meaningful scene regions that are most likely to contain those objects.

Objects that we use and search for daily are constrained by surfaces in different ways, and our surface maps successfully accounted for these differences. Garbage bins and paintings are found on large structural surfaces (floors and walls) that are invariant across scene categories, whereas drinking glasses are found on surfaces that change with scene category (tables/counters in kitchens, desks in offices). Paintings are typically found on vertical surfaces while drinking glasses and garbage bins are typically found on horizontal support surfaces. Finally, target object size and affordances limit where a target object is likely to appear (Castelhano & Witherspoon, 2016). For target objects conforming to these constraints, surface maps bolstered predictions made by meaning maps, thereby suggesting that the surface map method of identifying spatial constraint is sufficiently robust to account for target objects with different properties.

Prior work testing the influences of spatial constraint and image salience on eye movements during visual search shows that combining the two sources of information accounts for fixations significantly better than image salience alone (Ehinger et al., 2009; Torralba et al., 2006). Given that meaning and image salience are correlated yet meaning predicts attention better than image salience during visual search in scenes when this correlation is controlled (Hayes & Henderson, 2019; Peacock et al., under review), a major goal of the current study was to understand whether spatial constraint interacts with meaning to control eye movements. In the same way that the visual system constrains eye movements to physically salient scene regions within a target-defined region of space (Ehinger et al., 2006; Torralba et al., 2006), we found that the visual system also constrains eye movements to meaningful scene regions on target-related surfaces.

Another contribution of the current work is the concept of continuous surface maps. Previous studies have modeled spatial constraint using a single horizontal band (Torralba et al., 2006) or a single horizontal surface representing where a particular object is most likely to be located (Pereira & Castelhano, 2019). The current study introduced graded probabilistic surface maps to account for objects like drinking glasses that may be found on many different surfaces. These surface maps were then combined with meaning maps to predict search eye movements. Combining surfaces and meaning predicted search eye movements significantly better than either source of information alone. This novel combination of surfaces and meaning provides a powerful framework to understand what controls attention during visual search.

Scenes are three-dimensional yet the way we study them is with two-dimensional photographs. Although studies have found ways to deal with nuisances of using two-dimensional photos in the past (e.g., by using non-occluding objects (Nuthmann et al., 2020; Nuthmann & Henderson, 2010), summing representations of occluding objects (Hayes & Henderson, 2021), or by using chimera scenes (Castelhano et al., 2018; Man & Castelhano, 2018), the ability to model scene elements at varying depths is an important variable that should be taken into account in models of scene perception. To account for depth in the present study, we used image-computable depth maps to iteratively layer surface predictions based upon depth into our maps. This method allowed us to continuously model the probabilities of surfaces, even if they were occluded by other surfaces in the scene. We also accounted for the extent to which target objects extend above surfaces at different depths by generating a target object height constant for each object and its highly ranked surface elements. The resulting surface maps were able to continuously represent the likely locations of search target objects in scenes while taking into account each surface's depth from the viewer and the depth-dependent height of the target object, in a way that has not been previously done before.

Our findings are consistent with Pereira and Castelhano (2019) who used an attentional capture paradigm to test whether letter or object distractors that rapidly appeared on target-relevant or irrelevant surfaces were more likely to capture attention. They found that distractors were more likely to be fixated if they appeared on target-relevant surfaces and that this effect was stronger for object distractors. Similarly, we found that meaningful scene regions were more likely to be fixated when they were located on target-related surfaces even when those meaningful regions did not contain the target. Together, this suggests that the visual system may specifically use target-relevant surfaces to constrain search.

Previous work has shown that the gist of the scene is rapidly acquired within ~50ms of scene onset (Castelhano & Henderson, 2008; Greene & Fei-Fei, 2014; Oliva & Torralba, 2001, 2006; Potter, 1975; Potter et al., 2014) and that scene gist can be used to determine which scene regions are most relevant to search (Castelhano & Henderson, 2003). Indeed, past research has found that spatial constraint allows us to make predictions about what scene regions will be most task- or semantically-relevant for attentional prioritization (Brady et al., 2017; Brockmole & Henderson, 2006; Brockmole & Võ, 2010; Ehinger et al., 2009; Neider & Zelinsky, 2006; Torralba et al., 2006). The current results suggest that we may similarly use scene gist to pull out target-relevant surface information.

Conclusions

The present work made two major advances to the visual search literature. The first is the introduction of continuous surface maps, which capture constraints related to the likely locations of target objects in real-world scenes while taking depth information into account. The second major advancement is the novel combination of spatial constraint and meaning. The results show that during visual search, the visual system prioritizes meaningful scene regions on highly predictive surfaces over meaningful scene regions on target-unrelated surfaces.

References

Bahle, B., & Hollingworth, A. (2019). Contrasting episodic and template-based guidance during search through natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*. http://www2.psychology.uiowa.edu/faculty/hollingworth/documents/Bahle_Holl_inpress HPP.pdf

Bahle, B., Matsukura, M., & Hollingworth, A. (2018). Contrasting gist-based and template-based guidance during real-world visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), 367–386. https://doi.org/10.1037/xhp0000468

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*, 2577–2587. https://doi.org/10.1016/j.visres.2010.08.016

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160–1176. http://dx.doi.org/10.1037/xhp0000399

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, *13*(1), 99–108. https://doi.org/10.1080/13506280500165188

Brockmole, J. R., & Võ, M. L.-H. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception & Psychophysics*, *72*(7), 1803–1813. https://doi.org/10.3758/APP.72.7.1803

Castelhano, M. S., Fernandes, S., & Theriault, J. (2018). Examining the Hierarchical Nature of Scene Representations in Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*. https://doi.org/10.1037/xlm0000660

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception, and Psychophysics*, *72*(5), 1283–1297. https://doi.org/10.3758/APP

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, *18*(5), 890–896. https://doi.org/10.3758/s13423-011-0107-8

Castelhano, M. S., & Henderson, J. M. (2003). Flashing scenes and moving windows: An effect of initial scene gist on eye movements. *Journal of Vision*, *3*(9), 67–67. https://doi.org/10.1167/3.9.67

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 660–675. https://doi.org/10.1037/0096-1523.34.3.660

Castelhano, M. S., & Witherspoon, R. L. (2016). How You Use It Matters. *Psychological Science*, *27*(5), 606–621.

Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, *14*(8), 10–10. https://doi.org/10.1167/14.8.10

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6–7), 945–978.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 3–3. https://doi.org/10.1167/8.3.3

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14–14. https://doi.org/10.1167/14.1.14

Hayes, T. R., & Henderson, J. M. (2019). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin and Review*.

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: What a vector space model of semantics can tell us about attention in real-world scenes. *Psychological Science*.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504. https://doi.org/10.1016/j.tics.2003.09.006

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during visual search in real world scenes. In R. P. G. V. Gompel, M. H. Fischer, S. Murray, & Wayne (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 537–562). Elsevier Ltd. https://doi.org/10.1167/9.3.6

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*, 743–747. https://doi.org/10.1038/s41562-017-0208-0

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scenes: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 1–18. https://doi.org/10.1089/jmf.2012.0243

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper Depth Prediction with Fully Convolutional Residual Networks. *ArXiv:1606.00373 [Cs]*. http://arxiv.org/abs/1606.00373

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

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), 8–8.

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), 4–4. https://doi.org/10.1167/10.2.4

Man, L., & Castelhano, M. (2018). Across the planes: Differing impacts of foreground and background information on visual search in scenes. *Journal of Vision*, *18*(10), 384–384. https://doi.org/10.1167/18.10.384

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

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621. https://doi.org/10.1016/j.visres.2005.08.025

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How Well Can Saliency Models Predict Fixation Selection in Scenes Beyond Central Bias? A New Approach to Model

Evaluation Using Generalized Linear Mixed Models. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00491

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 20–20. https://doi.org/10.1167/10.8.20

Nuthmann, A., Schütz, I., & Einhäuser, W. (2020). Salience-based object prioritization during active viewing of naturalistic scenes in young and older adults. *Scientific Reports*, *10*(1), 22057. https://doi.org/10.1038/s41598-020-78203-7

Oliva, A., & Torralba, A. (2001). *Modeling the shape of the scene: A holistic representation of the spatial envelope*. 31.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. In *Progress in Brain Research* (Vol. 155, pp. 23–36). Elsevier.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019a). Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, and Psychophysics*, *81*(1), 20–34. https://doi.org/10.3758/s13414-018-1607-7

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica*.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2020). Center Bias Does Not Account for the Advantage of Meaning Over Salience in Attentional Guidance During Scene Viewing. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01877

Peacock, C. E., Singh, P., Hayes, T. R., & Henderson, J. M. (n.d.). *Meaning guides attention during visual search in real-world scenes*.

Pereira, E. J., & Castelhano, M. S. (2014). Peripheral Guidance in Scenes: The Interaction of Scene Context and Object Content. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 2056–2072.

Pereira, E. J., & Castelhano, M. S. (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-019-01610-z

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966. https://doi.org/10.1126/science.1145183

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, and Psychophysics*, *76*(2), 270–279.

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/xlm0000837

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), 4–4.

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), 5–5.

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), 766–786.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(4). https://doi.org/10.1167/9.7.4

Vickery, T. J., King, L. W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision*, *5*(1), 8–8.

Võ, M. L. H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, *126*(2), 198–212. https://doi.org/10.1016/j.cognition.2012.09.017

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behavior*, *1*(3), 0058–0058.

Zelinsky, G J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–787.

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The Role of Top-down and Bottom-up Processes in Guiding Eye Movements during Visual Search. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1569–1576). MIT Press. http://papers.nips.cc/paper/2805-the-role-of-top-down-and-bottom-up-processes-in-guiding-eye-movements-during-visual-search.pdf

Zelinsky, Gregory J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., Samaras, D., & Hoai, M. (2020). Predicting Goal-directed Attention Control Using Inverse-Reinforcement Learning. *ArXiv:2001.11921 [Cs]*. http://arxiv.org/abs/2001.11921

Chapter 5: Conclusion

The goal of this dissertation was to explore the influences of different scene properties on the guidance of visual attention in real-world scenes. Unlike previous work that used tasks which may have encouraged participants to attend to meaningful scene regions (Henderson & Hayes, 2017, 2018), Chapters 2 and 3 aimed to answer whether meaningful or physically salient scene regions were prioritized for attention during tasks in which either image salience was task-relevant and meaning was task-irrelevant (Chapter 2) or during a free viewing task that did not require attention to either salience or meaning (Chapter 3). Across the experiments described in Chapters 2 and 3, it was found that even in tasks that do not require attention to meaning, the overall and unique variance in attention was significantly more related to meaning than to image salience.

While the overall and unique variance in attention was significantly more related to meaning than to image salience, there is shared variance between meaning and image salience that guides attention. This means that the visual system might occasionally select a salient region over a meaningful region, even if on average meaningful regions were more likely to be selected. Although this is a possibility, the present data do not suggest that this is true. When the shared variance between meaning and image salience was removed, meaning explained substantial unique variance whereas image salience did not. So, while it is possible that salience drives attention to regions that are both meaningful and salient, it seems more likely (given meaning's better predictive power) that meaning is driving attention all of the time. It could also be the case that meaning and saliency work together to guide attention to certain regions. Because meaning is powerful enough to guide attention to regions not predicted by image salience yet the same is not true for image salience, it appears that meaning is the key predictor of attention, but saliency is not.

Previous studies have demonstrated that the visual system prioritizes physically salient information in target-relevant regions during visual search (Ehinger et al., 2009; Torralba et al., 2006). However, because the unique variance in attention was attributed to meaning but not salience in Chapters 2 and 3, Chapter 4 tested whether the visual system selects meaningful regions on target-relevant surfaces for attention. This was found to be true: attention prioritized meaningful scene regions on target-relevant surfaces but not those on target-irrelevant surfaces. This dissertation collectively demonstrates that the human visual system selects scene regions that contain meaningful content based upon our knowledge of the world for attention.

Image salience has been emphasized as a dominant factor in attentional guidance (Borji et al., 2013, 2014; Harel et al., 2006; Itti et al., 1998; Itti & Koch, 2001; Koch & Ullman, 1987). This stands at odds with cognitive relevance theory which proposes that the cognitive system will direct attention to information that is anticipated to be semantically relevant to its current goals and the context of the scene rather than be passively pulled to semantically uninterpreted image features (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003, 2017; Henderson et al., 1999, 2009; Tatler et al., 2011; Yarbus, 1967). The findings of this dissertation support cognitive relevance theory. Across the three studies presented here, we found that attention was directed to scene regions that the cognitive system predicted to be informative based upon world knowledge of semantic content rather than passively pulled by uninterpreted image features (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003, 2017; Henderson et al., 1999, 2009; Tatler et al., 2011; Yarbus, 1967). Furthermore, Chapter 4 demonstrated that top-down forms of knowledge interact, as the cognitive system selects meaningful regions in locations that are task-relevant for attention.

The finding that attention prioritizes semantic content regardless of the setting suggests that models of meaning could be used to solve a variety of real-world attentional guidance problems. For instance, in virtual reality, a computer could use the meaning map model to fully render semantically dense regions that users will likely attend whereas regions that are not semantically rich (e.g., sky) could be represented at a lower resolution. These findings could also be imported into applications, such as driving. If computers can infer where a driver should attend (the road, mirrors), then they could highlight meaningful regions in task-relevant locations to keep drivers focused. This, in turn, would reduce the likelihood of errors and accidents.

Semantic-Based Guidance of Attention

Together, the findings here support a model of attention in which the visual system makes use of world knowledge to orient attention. Overall, this dissertation demonstrates that regardless of the situation, attention is directed to semantically rich information in our environments with respect to our task and goals, supporting cognitive relevance theory (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003, 2017; Henderson et al., 1999, 2009; Tatler et al., 2011; Yarbus, 1967). The findings provide a better understanding of why the visual system selects certain regions of the world for analysis which could be used in real-world settings to highlight only the most relevant information for attention.

References

Anderson, N. C., & Donk, M. (2017). Salient object changes influence overt attentional prioritization and object-based targeting in natural scenes. *PlosOne*. https://doi.org/:10.1371/journal.pone.0172132

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Psychonomic Bulletin and Review, 23*, 1794-1801. doi:10.3758/s13423-016-1035-4

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 1–22. https://doi.org/10.1167/15.5.9

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103*(1), 62-70. doi:10.1037/h0036799

Bahle, B., & Hollingworth, A. (2019). Contrasting episodic and template-based guidance during search through natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*. http://www2.psychology.uiowa.edu/faculty/hollingworth/documents/Bahle_Holl_inpress HPP.pdf

Bahle, B., Matsukura, M., & Hollingworth, A. (2018). Contrasting gist-based and template-based guidance during real-world visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), 367–386. https://doi.org/10.1037/xhp0000468

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Royal Statistical Society B, 57*(1), 289-300.

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*, 2577–2587. https://doi.org/10.1016/j.visres.2010.08.016

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision*, *14*(13), 3.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*(1), 55-69. doi:10.1109/TIP.2012.2210727

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160–1176. http://dx.doi.org/10.1037/xhp0000399

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, *13*(1), 99–108. https://doi.org/10.1080/13506280500165188

Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: Effects of object–scene consistency. *Visual Cognition*, *16*(2-3), 375-390.

Brockmole, J. R., & Võ, M. L.-H. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception & Psychophysics*, *72*(7), 1803–1813. https://doi.org/10.3758/APP.72.7.1803

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology and perception in art*. Oxford, England.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). Mit saliency benchmark. Retrieved from http://saliency.mit.edu/results_mit300.html

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What do different evaluation metrics tell us about saliency models?. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3), 740-757.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*(9), 898-914. doi:10.1.1.79.1508

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception, and Psychophysics*, *72*(5), 1283–1297. https://doi.org/10.3758/APP

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, *18*(5), 890–896. https://doi.org/10.3758/s13423-011-0107-8

Castelhano, M. S., & Henderson, J. M. (2003). Flashing scenes and moving windows: An effect of initial scene gist on eye movements. *Journal of Vision*, *3*(9), 67–67. https://doi.org/10.1167/3.9.67

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(4), 753–763. https://doi.org/10.1037/0096-1523.33.4.753

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 660–675. https://doi.org/10.1037/0096-1523.34.3.660

Castelhano, M. S., & Witherspoon, R. L. (2016). How You Use It Matters. *Psychological Science*, *27*(5), 606–621.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6-6.

Castelhano, M., Fernandes, S., & Theriault, J. (2018). Examining the Hierarchical Nature of Scene Representations in Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*. https://doi.org/10.1037/xlm0000660

Chen, Y., & Zelinsky, G. J. (2019). Is there a shape to the attention spotlight? Computing saliency over proto-objects predicts fixations during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(1), 139.

Clarke, A. D., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41-51.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317–329. https://doi.org/10.1007/BF00868064

Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, *14*(8), 10–10. https://doi.org/10.1167/14.8.10

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6–7), 945–978.

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), 1–19.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 3–3. https://doi.org/10.1167/8.3.3

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception?. *Perception*, *36*(8), 1123-1138.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14-14. doi:10.1167/14.1.14

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*.

Hayes, T. R., & Henderson, J. M. (2019). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin and Review*.

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: What a vector space model of semantics can tell us about attention in real-world scenes. *Psychological Science*.

Hayhoe, M. M., & Ballard, D. H. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(6), 49-63. doi:10.1167/3.1.6

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188-194. doi:10.1016/j.tics.2005.02.009

Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology*, *24*(13), R622-R628.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504. https://doi.org/10.1016/j.tics.2003.09.006

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16*(4), 219-222. doi:https://doi.org/10.1111/j.1467-8721.2007.00507.x

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences*, *21*(1), 15–23. https://doi.org/10.1016/j.tics.2016.11.003

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York, NY, US: Psychology Press.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*, 743–747. https://doi.org/10.1038/s41562-017-0208-0

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scenes: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 1–18. https://doi.org/10.1089/jmf.2012.0243

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

Henderson, J. M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, *10*(5), 438-443.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. v. Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 537-562): Elsevier Ltd.

Henderson, J. M., Hayes, T. R. Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports, 8(13504).* doi: 10.1038/s41598-018-31894-5

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach. *Vision*, *3*(2), 19. https://doi.org/10.3390/vision3020019

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports*, *8*.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850-856.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews, 2*(3), 1-11.

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*(1), 161–169.

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).

Koch, C., & Ullman, S. (1987). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Matters of Intelligence*, *4*(4), 115–141. https://doi.org/10.1007/978-94-009-3833-5_5

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

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper Depth Prediction with Fully Convolutional Residual Networks. *ArXiv:1606.00373 [Cs]*. http://arxiv.org/abs/1606.00373

Land, M. F., & Hayhoe, M. (2001). In what ways to eye movements contribute to everyday activities? *Vision Research, 41*(25-26), 3559-3565. doi:https://doi.org/10.1016/S0042-6989(01)00102-X

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

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*(11), 547–552.

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), 8–8.

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), 4–4. https://doi.org/10.1167/10.2.4

Man, L., & Castelhano, M. (2018). Across the planes: Differing impacts of foreground and background information on visual search in scenes. *Journal of Vision*, *18*(10), 384–384. https://doi.org/10.1167/18.10.384

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205-231. doi:https://doi.org/10.1016/j.visres.2004.07.042

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*(4), 605-617. doi: https://doi.org/10.1016/j.neuron.2007.01.018

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614-621.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 20–20. https://doi.org/10.1167/10.8.20

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How Well Can Saliency Models Predict Fixation Selection in Scenes Beyond Central Bias? A New Approach to Model Evaluation Using Generalized Linear Mixed Models. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00491

Nuthmann, A., Schütz, I., & Einhäuser, W. (2020). Salience-based object prioritization during active viewing of naturalistic scenes in young and older adults. *Scientific Reports*, *10*(1), 22057. https://doi.org/10.1038/s41598-020-78203-7

Oliva, A., & Torralba, A. (2001). *Modeling the shape of the scene: A holistic representation of the spatial envelope*. 31.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. In *Progress in Brain Research* (Vol. 155, pp. 23–36). Elsevier.

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. https://doi.org/10.1016/S0042-6989(01)00250-4

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019a). Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, and Psychophysics*, *81*(1), 20–34. https://doi.org/10.3758/s13414-018-1607-7

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica*.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2020). Center Bias Does Not Account for the Advantage of Meaning Over Salience in Attentional Guidance During Scene Viewing. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01877

Peacock, C. E., Singh, P., Hayes, T. R., & Henderson, J. M. (n.d.). *Meaning guides attention during visual search in real-world scenes*.

Pereira, E. J., & Castelhano, M. S. (2014). Peripheral Guidance in Scenes: The Interaction of Scene Context and Object Content. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 2056–2072.

Pereira, E. J., & Castelhano, M. S. (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-019-01610-z

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966. https://doi.org/10.1126/science.1145183

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, and Psychophysics*, *76*(2), 270–279.

Rahman, S., & Bruce, N. (2015). Visual saliency prediction and evaluation across different perceptual tasks. *PlosOne*. doi:10.1371/journal.pone.0138053

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/xlm0000837

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 16-16.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2016). Task and context determine where you look. *Journal of Vision, 7*(16), 1-20. doi:10.1167/7.14.16

Spotorno, S., Tatler, B. W., & Faure, S. (2013). Semantic consistency versus perceptual salience in visual scenes: Findings from change detection. *Acta Psychologica*, *142*(2), 168-176.

Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision*, *7*(10), 3-3.

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), 4–4.

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), 5-5.

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), 766-786. doi:10.1037/0033-295X.113.4.766

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

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(4). https://doi.org/10.1167/9.7.4

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, *43*(3), 333-346.

Vickery, T. J., King, L. W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision*, *5*(1), 8–8.

Võ, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision2*, *9*(3), 1–15. https://doi.org/10.1167/9.3.24

Võ, M. L.-H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, *126*(2), 198–212. https://doi.org/10.1016/j.cognition.2012.09.017

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behavior*, *1*(3), 0058–0058.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419–433.

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 54.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye Movements and Vision* (pp. 171–211). Springer. https://doi.org/10.1007/978-1-4899-5379-7_8

Zelinsky, G J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–787.

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The Role of Top-down and Bottom-up Processes in Guiding Eye Movements during Visual Search. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1569–1576). MIT Press. http://papers.nips.cc/paper/2805-the-role-of-top-down-and-bottom-up-processes-in-guiding-eye-movements-during-visual-search.pdf

Zelinsky, Gregory J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., Samaras, D., & Hoai, M. (2020). Predicting Goal-directed Attention Control Using Inverse-Reinforcement Learning. *ArXiv:2001.11921 [Cs]*. http://arxiv.org/abs/2001.11921