Meaning and expected surfaces combine to guide attention during visual search in scenes

Candace E. Peacock
Center for Mind and Brain, University of California, Davis, Davis, CA, USA
Department of Psychology, University of California, Davis, Davis, CA, USA

Deborah A. Cronin
Center for Mind and Brain, University of California, Davis, Davis, CA, USA

Taylor R. Hayes
Center for Mind and Brain, University of California, Davis, Davis, CA, USA

John M. Henderson
Center for Mind and Brain, University of California, Davis, Davis, CA, USA
Department of Psychology, University of California, Davis, Davis, CA, USA

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. Attention was indexed by eye movements during the 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.

Introduction

Owing 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). Recent work (Hayes & Henderson, 2019) suggests that the visual system may also prioritize local scene regions that are high in meaning during visual search, but it is unknown how local, context-free meaning (i.e., meaning maps; for a review, see Henderson et al., 2019) 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).

Citation: Peacock, C. E., Cronin, D. A., Hayes, T. R., & Henderson, J. M. (2021). Meaning and expected surfaces combine to guide attention during visual search in scenes. *Journal of Vision*, *21*(11):1, 1–18, https://doi.org/10.1167/jov.21.11.1.

https://doi.org/10.1167/jov.21.11.1 Received April 5, 2021; published October 5, 2021 ISSN 1534-7362 Copyright 2021 The Authors

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 the study by Torralba and colleagues 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 and 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 approach taken by Pereira and Castelhano (2019), 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 on 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 (see Figure 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 considering 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). 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 usefulness 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, p. 201; Rehrig et al., 2020; Tatler et al., 2011) and the finding that meaning accounts for most if not all of the variance in eye fixations when the intercorrelation

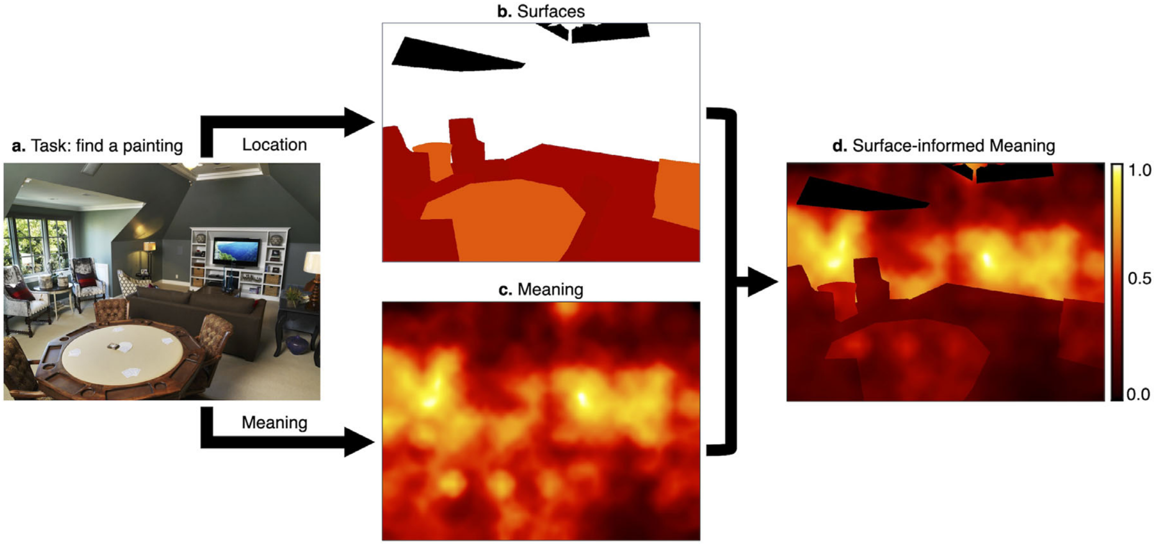


Figure 1. Schematic of surface map model. If the goal is to find a painting in a media room (a), the probability that a painting will appear on one surface (walls) over another surface (floors) will drive attention to the more probable region (b). Analogously, meaningful (informative) scene regions are 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 nonpredictive surfaces (black) (d).

between meaning and saliency is controlled (Hayes & Henderson, 2019; Henderson & Hayes, 2017, 2018; Peacock et al., 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 surface (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) and that eye movements are restricted to scene regions 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 1).

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 owing to blinks or loss of calibration. Following Henderson and Hayes (2017), we averaged the percent signal ([number of good samples/total number of samples] × 100) for each trial using custom MATLAB code. The percent signal across trials was averaged for each participant and compared with an a priori 75% criterion for signal. Overall, no 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 on this criterion.

Participants were also excluded if they did not do the task correctly. 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 more than 25% of trials, that participant was excluded, resulting in the removal of seven 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° × 20.0°

of visual angle at 1024 × 768 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 the 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 50 ms or greater than 1500 ms were eliminated as outliers. Occasionally, saccade amplitudes are not segmented correctly by EyeLink's standard algorithm, resulting in large values. Given this, saccade amplitudes of more than 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

We selected 105 digitized photographs (1024 × 768 pixels) of indoor and outdoor real-world scenes 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.

Paintings, drinking glasses, and garbage bins were selected as the target objects because these objects reside in the upper, middle, and lower horizontal regions of scenes, respectively. This approach allowed us sample target locations across the full areas of the scenes, consistent with Torralba et al. (2006) and Pereira and Castelhano (2019).

To select suitable target absent scenes, we first identified scenes that did not contain the target object from a large "in-house" database of annotated scenes. Care was taken to ensure that there would have been enough space for each of the target objects in these scenes. From here, only indoor scenes were used for paintings, because paintings typically reside on indoor walls. Both outdoor urban scenes and indoor scenes were used for garbage bins, because garbage bins typically appear on the floor in manmade settings. Finally, indoor (e.g., kitchens, offices, bars) and outdoor scenes (e.g., back patios) that contained manmade horizontal support surfaces were selected for drinking glasses.

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 400 ms to orient participants to the center of the screen where a word cue would appear. Then, a word cue was presented for 800 ms indicating the search target for that scene. After the word cue, the central fixation cross reappeared for 400 ms before 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 explore the scene fully. 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 nine-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°. 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

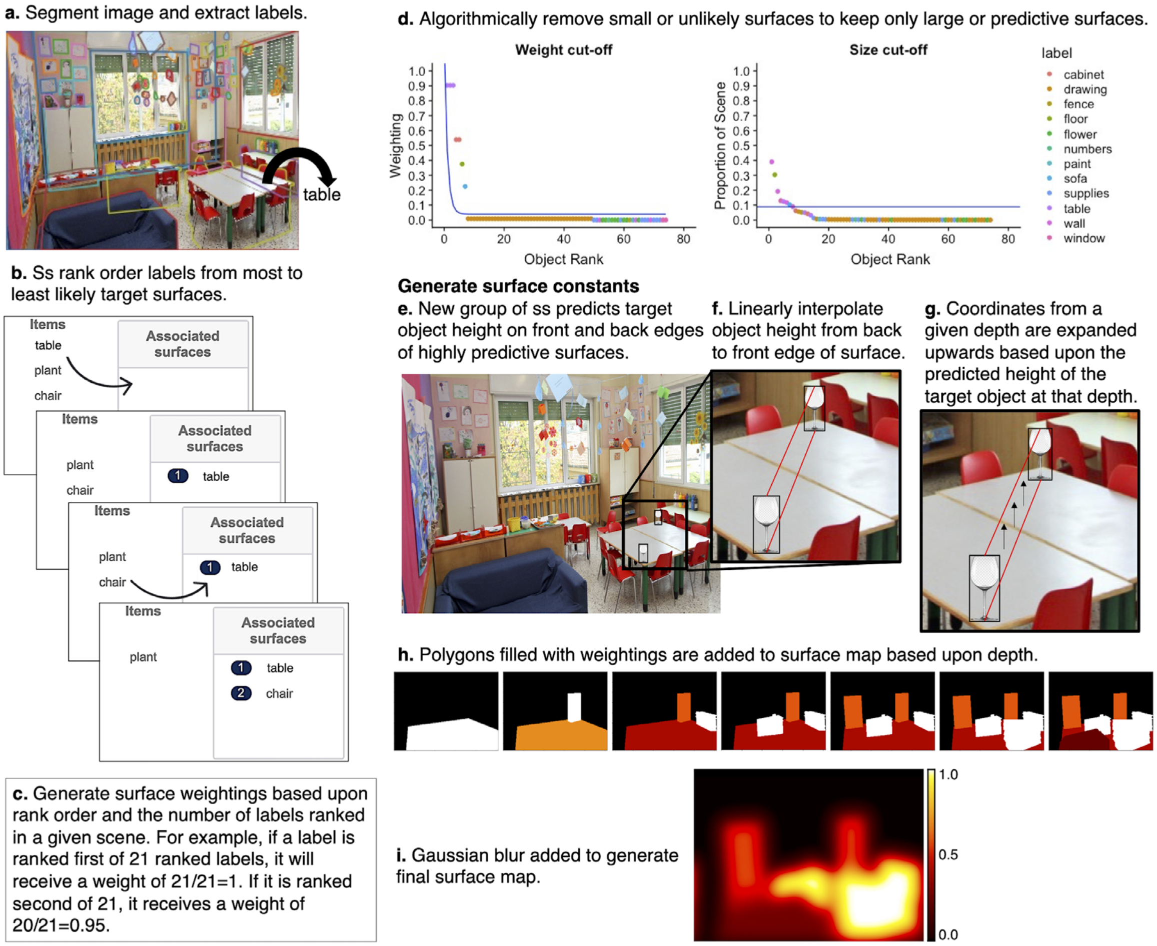
Participants

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 (four participants from the garbage bin task, two participants from the drinking glass task, and no participants from the painting task).

Scene labeling and segmentation

All scene elements that were present in any of the 105 target present and absent 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 2a).

Labels corresponding with 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. Although we did not analyze the target present scenes in the present study, we still acquired their surface rankings. To decrease confusion to participants for these target present scenes, we excluded labels that were synonyms of the target object in these scenes. 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



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 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).

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 on. Participants were asked to drag and drop the labels into a provided box on the computer screen, and to rank order them based on how likely a drinking glass would be to appear on that given surface (Figure 2b). Participants were instructed not to rank (i.e., not to drag into the box) labels that were not surfaces upon which a drinking glass would appear. Before beginning the survey, participants were given an example ranking question (Figure 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 with the 35 scenes for that target object category, presented in a random order for each participant. The labels corresponding with a given scene were provided in a randomized order to the left of the ranking column (Figure 2b).

Generating surface weights

We first generated weights corresponding with 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 of 21 (Figure 2c). If a label was placed second out of 21 ranked labels, it would be given a participant-level weighting of 20 of 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 were excluded. This resulted in the 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 with each element in each scene.

Eliminating small and nonpredictive 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.

To eliminate small elements, we compared the size of each element with 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 2d). If a given element's size was less than the size threshold then it was tagged for possible deletion.

To eliminate nonpredictive element ratings from predictive elements in a principled manner, we first ranked each element in descending order by scene based on its surface weighting on the x-axis and plotted the weighting (Figure 2d) on the y-axis, respectively. We then fit an exponential function [$y = e^{(-x)}$] to the weighting data (Figure 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 keep elements that were small but also predictive.

Above-surface constant

Because objects tend to extend in space above the surfaces or 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 who were naïve to the purpose of the study indicated how tall an average sized target would seem to be on either the front or back edge of highly predictive surface elements (corresponding with labels weighted

0.5 or greater) in each scene (Figure 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 2f). We separated the segmentation for a given surface element into 10 slices based on the y dimension and expanded the coordinates based on how tall the target object was estimated to be at that slice (Figure 2g). Both the expanded coordinates and the original coordinates were added to the resulting surface map, because participants were predicted to look on and above predictive surfaces (Figure 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. The New York University Depth Dataset is a database of scenes with ground truth depth values obtained using Microsoft Kinect (Silberman et al., 2012). This database has been used to establish benchmarks for various depth map algorithms. The depth map algorithm we used in the present article is the state of the art in terms of predicting these ground truth values (Laina et al., 2016). 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 nonoccluded) elements.

Surface map generation

After finalizing the weights and constants for each surface element, we generated empty surface maps by first creating a 768 × 1024 array of zeros. We then replaced the existing values on the surface map with each element's weighting based on that element's spatial location and depth relative to the other elements in the scene to account for foreground elements occluding background elements (Figure 2h). Here, elements were added from the back (deepest) to the front (shallowest) based on each element's median depth generated from the depth maps (Laina et al., 2016). Constant values for elements corresponding with 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 −6 dB (a window size of approximately 2° of visual angle) was applied to each map. The Gaussian low-pass function is from the Massachusetts Institute of Technology's Saliency Benchmark code.[1] Although we did not explicitly conduct tests of subjective depth judgments, the rank ordering generally agreed with our own subjective assessments. The only instances where they did not agree were for surfaces that extended from the front of a space to the back of a space, such as a floor. In these instances, these surfaces were set to the deepest depth so that foreground objects could be placed on top of them.

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 and 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, owing 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 (Figure 3). 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.

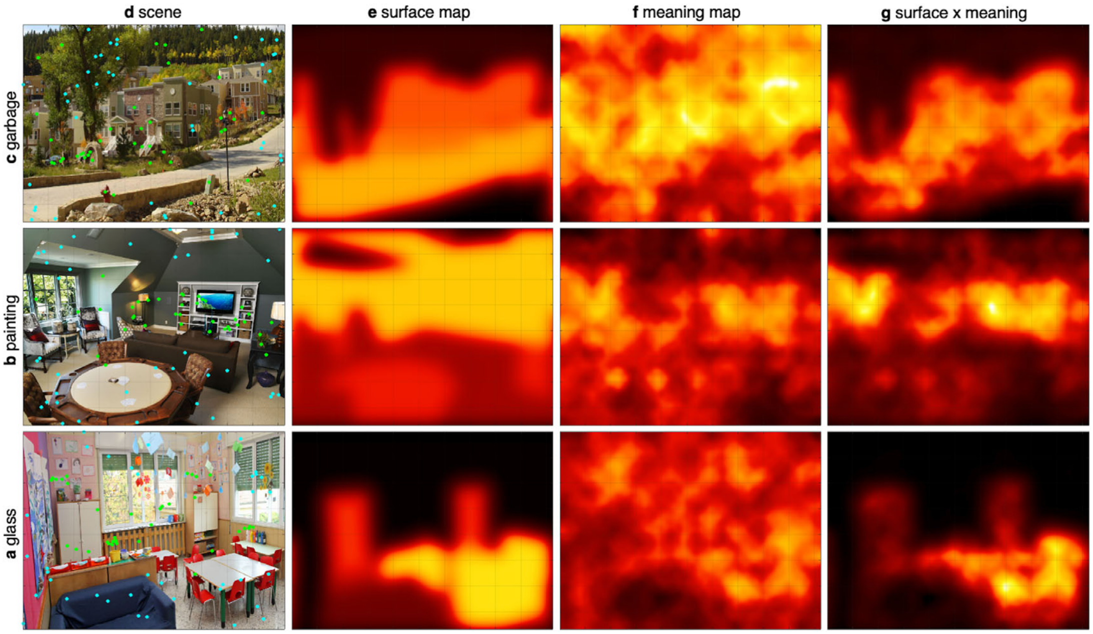


Figure 3. Map examples. The figure shows an example of each map type for drinking glasses (a), paintings (b), and garbage bins (c). Each column represents an example scene with fixated (green) versus nonfixated (cyan) regions for a single participant (d), with each respective surface map (e) meaning map (f), and hypothesized interaction of surfaces and meaning.

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 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 in the Eyetracking search analysis to account for and control the role of center bias, the tendency to fixate centrally, and photographer bias which is the tendency for photographers to place information of interest to humans in the center of a photograph (Bindemann, 2010; Hayes & Henderson, 2021; Tatler, 2007; Tseng et al., 2009) (Figure 4d).

Eyetracking search analysis

To test whether surfaces and meaning interact to predict fixated and nonfixated 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 with target-absent scenes because 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° window (113 pixels in diameter) around each fixation in the surface map (Figure 4b), meaning map (Figure 4c), and center proximity map (Figure 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 nonfixated scene regions was that the nonfixated 3° windows could not overlap with any of the 3° 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

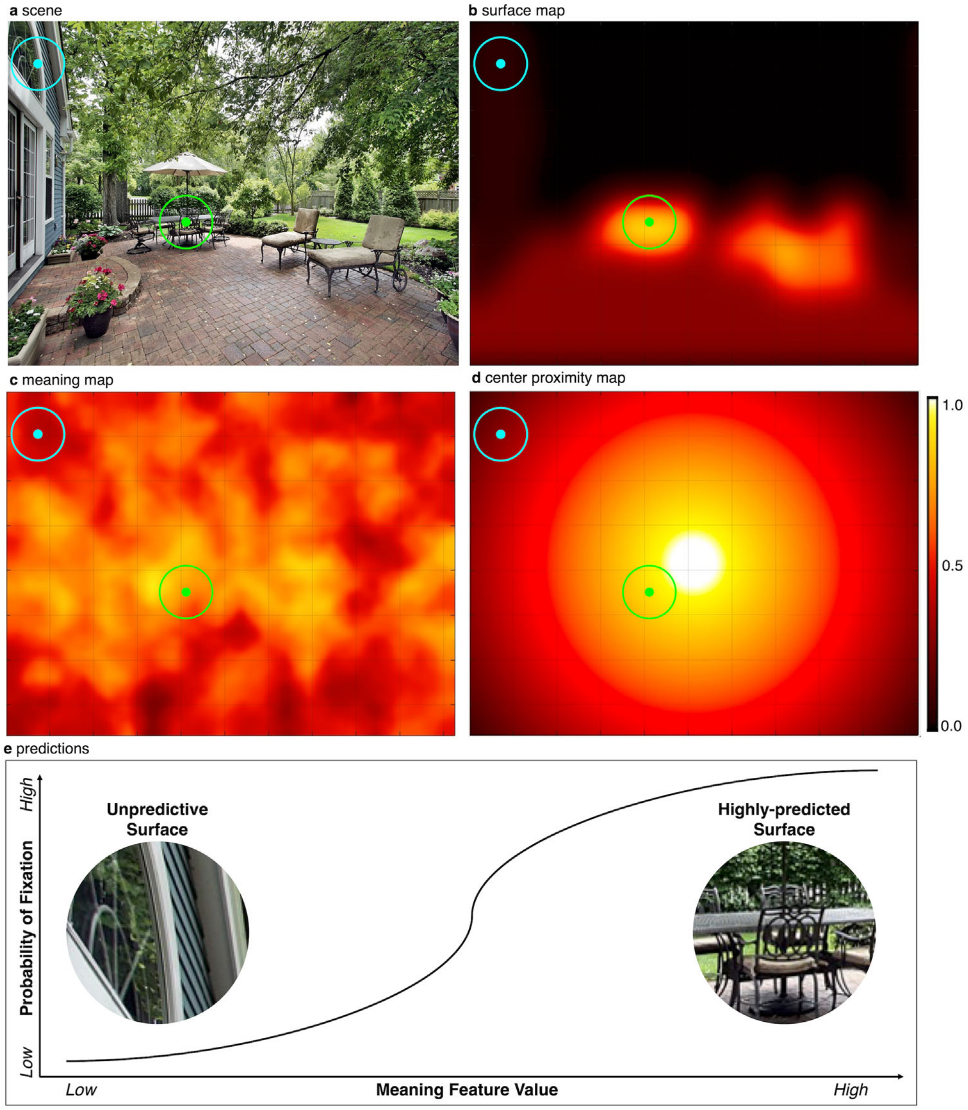


Figure 4. Analysis and predictions. The figure shows an example scene (a), surface map (b), meaning map (c), and center proximity map (d) with hypothetical fixated (green) versus nonfixated (cyan) windows. Predicted results (e) shows that meaningful scene regions have a higher probability of fixation 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.

due to center bias. Additionally, we included a random intercept of scene. Including a random intercept of participant did not account for significant 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 4e).

Examining common structure in surface maps

Given our hypothesis that scene-specific surfaces will predict fixations better than scene-independent bands, we also tested whether the surface maps for one scene predict fixations for the same scene better than fixations for another scene. This process allowed us to examine whether there is common structure in the surface maps or whether the surface maps capture scene-dependent variance in where fixations are directed.

To test whether surface maps for a given scene (scene A) predict fixated locations for the same scene better than fixated locations for another scene (scene B), we computed the mean surface feature values of one scene (e.g., scene A) using 3° windows corresponding with all fixations for a given participant from another scene (e.g., scene B). This process was repeated for each scene and participant. From here, we computed the average surface map value at fixation across participants for a given scene. This resulted in a 25 × 25 matrix for each target object in which the diagonals corresponded with the fixations from the same scene (i.e., scene A surface map values from scene A fixations) and off-diagonals which corresponded with fixations from another scene (e.g.. scene A surface map values from scene B fixations).

Theoretically, if each surface map is capturing scene-dependent variance in where people search for objects, then the diagonal of the matrix should have a larger value than the off-diagonal value. Conversely, if the models are only capturing some common structure in where people search for these objects, then the matrices should be uniform. To test this, difference calculations were computed, which produced 25 × 25 difference matrices for each target object. Difference scores were computed by taking the difference between the surface feature map values for fixations from the same scene (i.e., the diagonals) and the surface map values using fixations from other scenes (off-diagonals). If a given surface map was more strongly tied with the fixations from the same scene than another scene, then the difference score was positive. If a given surface map was more strongly tied with fixations from another scene than the same scene, then the difference score was negative. Difference scores along the diagonal were 0. The average difference score from each scene was then computed and submitted to a one-sample *t* test comparing the difference scores for each target object.

Results

Eye-tracking search analysis

Our primary question asked whether fixations are directed to meaningful scene regions that occur on target-related surfaces during search in scenes. Figure 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 aforementioned GLME model described in the methods 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.

The GLME model results for meaning are visualized in Figure 6 and Table 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; and 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.

We examined the three-way interactions to ensure center proximity was not modulating the meaning by surface effects (Figure 7). If the meaning by surface interaction was driven by center proximity, we would expect high meaning and surface values to be fixated

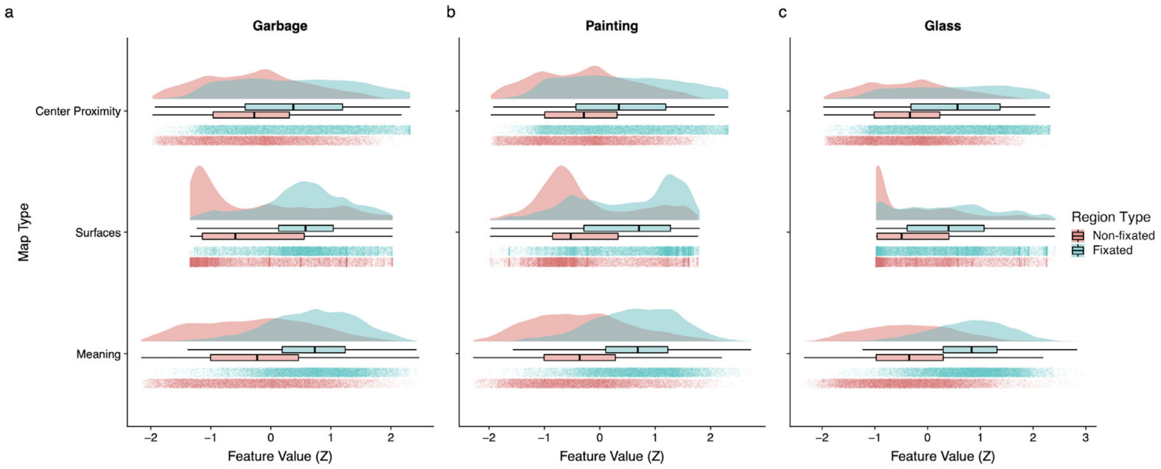


Figure 5. Summary plots of the raw eye movement data. Raincloud plots show the center proximity, surface, and meaning z-normalized feature values on fixated (blue) and nonfixated (pink) scene regions for garbage bins (a), paintings (b), and drinking glasses (c). For each box plot, the whiskers refer 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 nonfixated window.

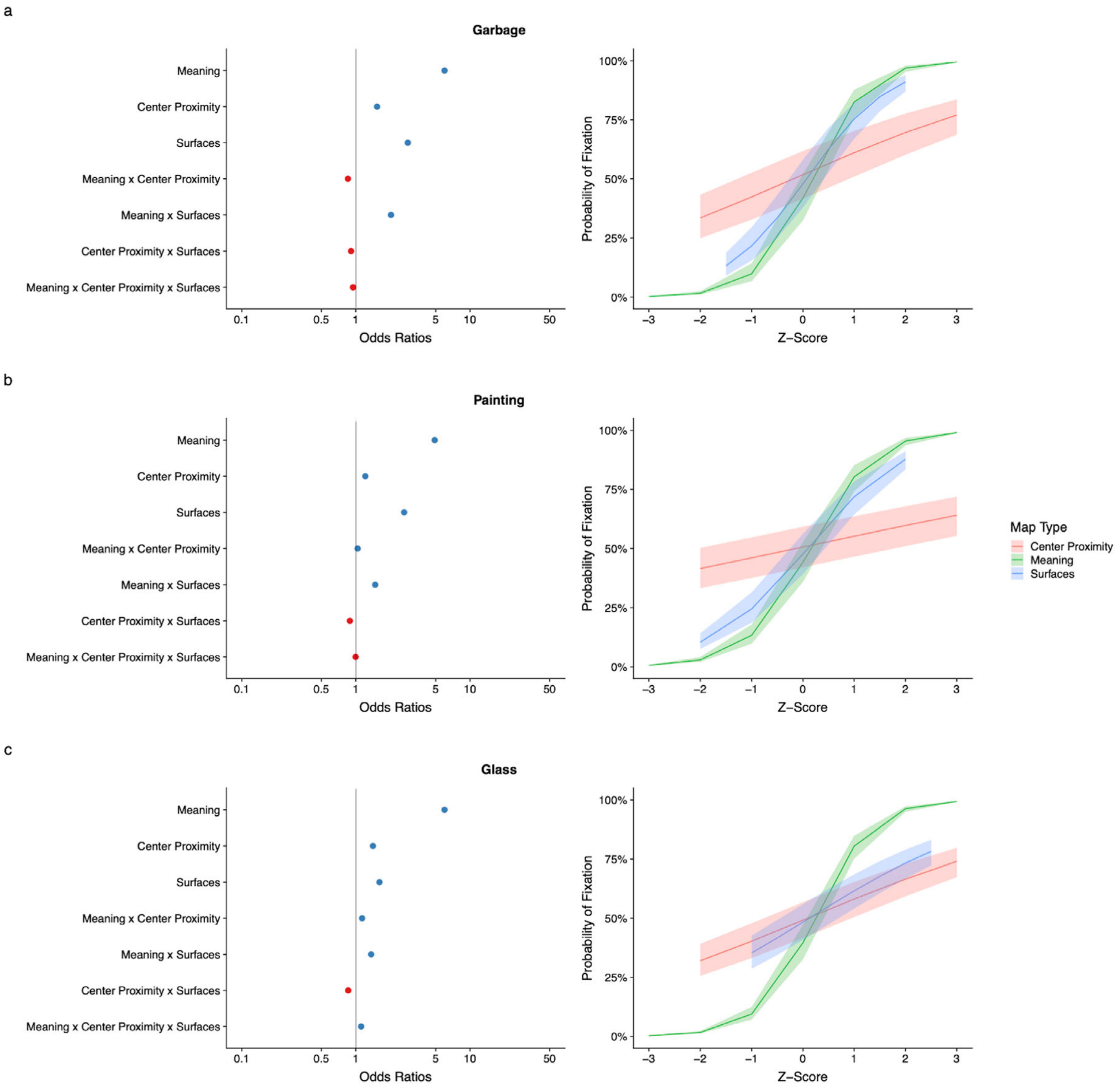


Figure 6. Model fits. The log odds (left column) and the marginal effects (right column) for the garbage (a), painting (b), and drinking glass (c) models are visualized. A log odds of 1 indicate that neither positive nor negative values of a predictor are likely to occur with fixated regions. Al og odds of greater than 1 (blue) indicate that positive values of a predictor are more associated with fixated regions, whereas a log odds 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.

at scene centers owing 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 (Figures 7a, 7d, 7g). However, this effect did not change as a function of center proximity: for fixations further from center (Figures 7b 7e, 7h) and in scene peripheries (Figures 7c, 7f, 7i), higher meaning regions were more likely to be fixated if the corresponding surface values were higher. The three-way interaction for garbage cans seems to be the result of the lack of an asymptote in the low probability surfaces (red curves in Figure 6) at high meaning values compared with 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.

Examining common structure in surface maps

Given our hypothesis that scene-specific surfaces will predict fixations better than scene-independent

	Fixed effects					Random effects, SD
Predictors	β	95% CI	SE	Z-statistic	*p* value	By-scene
Garbage						
Intercept	−0.51	[−0.93 to −0.08]	0.21	−2.42	0.02	1.04
Meaning	1.79	[1.73 to 1.85]	0.03	60.58	<0.001	
Center proximity	0.43	[0.40 to 0.47]	0.02	22.98	<0.001	
Surfaces	1.05	[1.01 to 1.09]	0.02	48.70	<0.001	
Meaning × Center proximity	−0.16	[−0.20 to −0.11]	0.02	−7.08	<0.001	
Meaning × Surfaces	0.71	[0.66 to 0.76]	0.03	27.61	<0.001	
Center proximity × Surfaces	−0.10	[−0.14 to −0.05]	0.02	−4.46	<0.001	
Meaning × Center proximity × Surfaces	−0.06	[−0.11 to 0.001]	0.03	−2.01	0.05	
Painting						
Intercept	−0.36	[−0.72 to −0.004]	0.18	−2.06	0.04	0.88
Meaning	1.59	[1.54 to 1.64]	0.03	60.52	<0.001	
Center proximity	0.19	[0.15 to 0.23]	0.02	9.83	<0.001	
Surfaces	0.97	[0.94 to 1.01]	0.02	48.60	<0.001	
Meaning × Center proximity	0.04	[−0.004 to 0.08]	0.02	1.78	0.08	
Meaning × Surfaces	0.39	[0.35 to 0.44]	0.02	17.22	<0.001	
Center proximity × Surfaces	−0.12	[−0.16 to −0.08]	0.02	−5.92	<0.001	
Meaning × Center proximity × Surfaces	−0.003	[−0.05 to 0.04]	0.02	−0.13	0.90	
Glass						
Intercept	−0.51	[−0.83 to −0.19]	0.16	−3.27	0.001	0.78
Meaning	1.79	[1.74 to 1.84]	0.03	69.35	<0.001	
Center proximity	0.35	[0.31 to 0.38]	0.02	18.21	<0.001	
Surfaces	0.48	[0.44 to 0.52]	0.02	25.54	<0.001	
Meaning × Center proximity	0.13	[0.08 to 0.17]	0.02	5.70	<0.001	
Meaning × Surfaces	0.31	[0.26 to 0.36]	0.03	12.31	<0.001	
Center proximity × Surfaces	−0.15	[−0.19 to −0.11]	0.02	−7.44	<0.001	
Meaning × Center proximity × Surfaces	0.11	[0.06 to 0.16]	0.03	4.22	<0.001	

Table 1. Meaning × Surface × center proximity GLME results for Each target object. *Notes:* beta estimates (β), 95% confidence intervals (CI), standard errors (SE), z-statistic, and *p* values (*p*) for each fixed effect and standard deviations (SD) for the scene random effect. CI = confidence interval; SD = standard deviation; SE = standard error.

bands, we examined whether the surface maps for one scene predict fixations for the same scene better than fixations for another scene. This process allowed us to test whether there is a common structure in the surface maps or whether the surface maps capture scene-dependent variance in where fixations are directed. To test this, we used the difference matrices described elsewhere in this article. If a given surface map was more strongly tied with the fixations from the same scene than another scene, then the difference score was positive. If a given surface map was more strongly tied with fixations from another scene than the same scene, then the difference score was negative (Figure 8). Difference scores along the diagonal were 0.

Overall, the surface maps for a given scene were significantly more related to the fixations from the same scene than another scene for each of the target objects, garbage: $M = 0.25$, $SD = 0.32$, $t(24) = 3.93$, $p < 0.001$, 95% confidence interval = 0.12–0.38; painting: $M = 0.12$, $SD = 0.16$, $t(24) = 3.71$, $p = 0.001$, 95% confidence interval = 0.05–0.19; and glass: $M = 0.19$, $SD = 0.21$, $t(24) = 4.44$, $p < 0.001$, 95% confidence interval = 0.10–0.27. This finding suggests that the surface maps are indeed capturing scene-dependent variance in where people search for objects rather than scene-independent biases in where people search for objects.

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

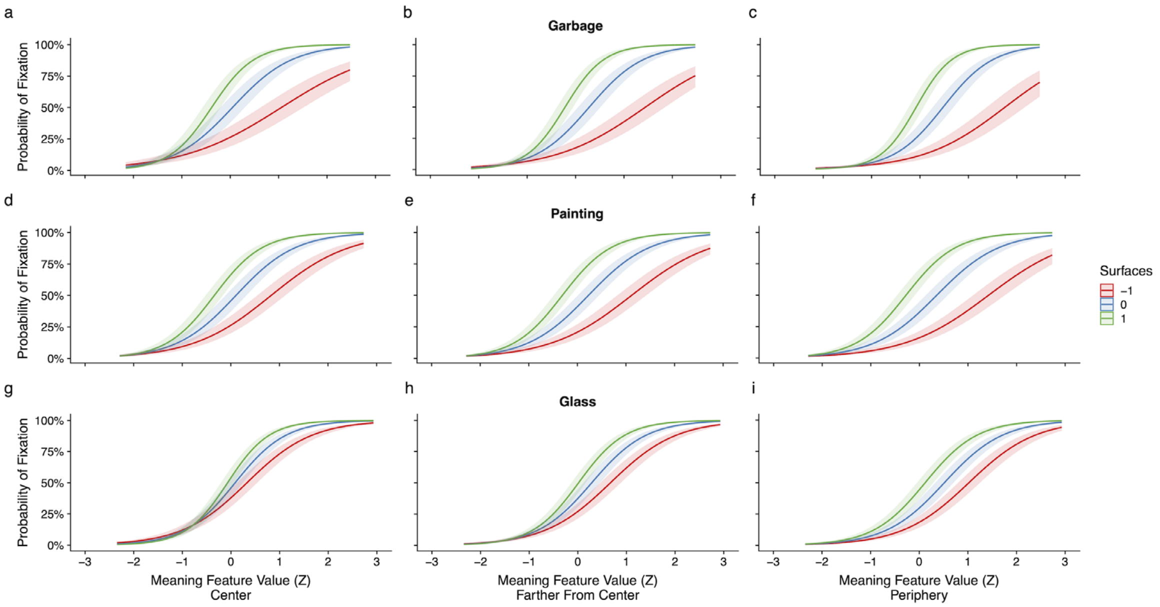


Figure 7. Three-way Meaning × Surfaces × Center proximity interaction. This figure shows the probability that meaningful scene regions were fixated on surfaces that were not predictive of target object locations (red), moderately predictive (blue), and highly predictive of target location (green) at 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.

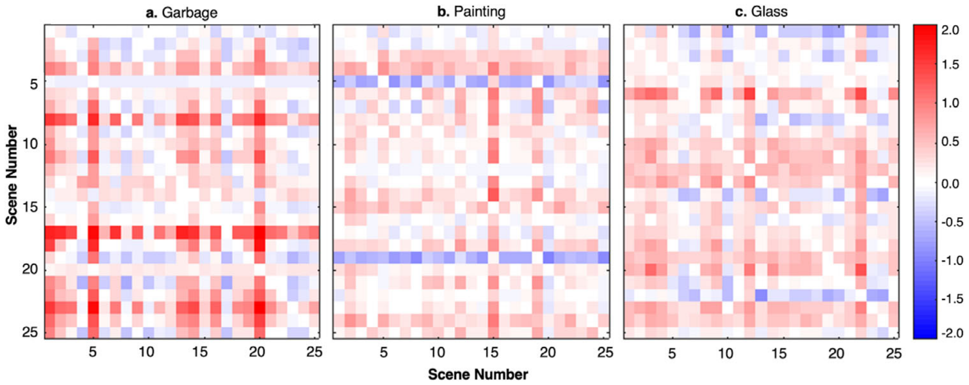


Figure 8. Examining common structures in surface maps via difference matrices. Differences matrices are visualized for garbage bins (a), paintings (b), and drinking glasses (c). The diagonals correspond with the difference between the surface map values and fixations for the same scene and itself (which equals 0). Off-diagonals correspond with the difference between surface map values and fixations for the same scene minus surface map values and fixations for another scene. Red refers to positive difference scores (i.e., the surface map for a given scene is more strongly related to fixations from the same scene than another scene). Blue corresponds to negative difference scores (i.e., fixations from another scene are more strongly related to a given surface map than fixations from the same scene).

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 locally meaningful (recognizable and informative) 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), 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., 2009; 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. Here, we found that fixations from a given scene were more related to surface maps from the same scene than fixations from another scene. This finding, in total, suggests that surface maps capture scene-dependent variance in where people search for objects rather than scene-independent biases in where people attend. 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 the control of 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 nonoccluding 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 on 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 considering 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 those from Pereira and Castelhano (2019), who used an attentional capture paradigm to test whether letter or object distractors that briefly 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 finding suggests that the visual system may specifically use target-relevant surfaces to constrain the search.

Cognitive guidance theory (Henderson, 2003, 2017; Henderson et al., 1999, 2009) proposes that people will orient their attention to information that is relevant to the cognitive system. This can include semantically informative information in scenes and task-relevant scene regions. Although surfaces (e.g., walls) are not necessarily semantically informative (because walls are blank), they were task relevant in the present study. This factor suggests that the cognitive system integrates knowledge of the task and knowledge of the environment to highlight regions of the scene that are both task-relevant and semantically informative.

Previous work has shown that the gist of the scene (basic-level category) is rapidly acquired within approximately 50 ms 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.

Keywords: scene perception, eye movements, meaning, spatial constraint, visual search

Acknowledgments

The authors thank Michelle Galat, Erika Agbabian, Aimee Nguyen, Mehvish Javed, Nat Finnegan, and Viviana Padilla for assisting in data collection. We would also like to thank Elizabeth Hall and Anthony Lagunda for annotating the scenes used in the present study.

Supported by the National Eye Institute of the National Institutes of Health (grant number R01EY027792). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This project also received support from the UC Davis Vision Science Training Program (T32 EY015387).

Commercial relationships: none.

Corresponding author: Candace E. Peacock.
Email: cepeacock@ucdavis.edu.
Address: Center for Mind and Brain, 267 Cousteau Place, University of California, Davis, CA 95618, USA.

Footnote

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

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, 45*(4), 523–536, https://doi.org/10.1037/xhp0000624.

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.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*, 143–177.

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*(9), 1619–1633, 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, https://doi.org/10.1177/0956797616629130.

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, https://doi.org/10.1080/13506280902834720.

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, 26* (5), 1683–1689, https://doi.org/10.3758/s13423-019-01642-5.

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, 32* (8), 1262–1270, https://doi.org/10.1177/0956797621994768.

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. (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., 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). New York: 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., 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., Malcolm, G. L., & Schandl, Charles. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin and Review, 16*(5), 850–856, https://doi.org/10.3758/PBR.16.5.850.

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, https://doi.org/10.1167/9.11.8.

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. S. (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, https://doi.org/10.1016/j.visres.2004.07.042.

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*, 491, 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. *International Journal of Computer Vision, 42*(3), 145–175.

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). New York: 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, 198*, 102889, https://doi.org/10.1016/j.actpsy.2019.102889.

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*, 1877, https://doi.org/10.3389/fpsyg.2020.01877.

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, http://dx.doi.org/10.1037/a0037524.

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, 26* (4), 1273–1281, 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, https://doi.org/10.3758/s13414-013-0605-z.

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, 46* (9), 1659–1681, http://dx.doi.org/10.1037/xlm0000837.

Research, S. R. (2010a). Experiment builder user's manual. Mississauga, ON: SR Research Ltd.

Research, S. R. (2010b). EyeLink 1000 user's manual, version 1.5.2. Mississauga, ON: SR Research Ltd.

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision – ECCV 2012* (Vol. *7576*, pp. 746–760). Berlin, Heidelberg: Springer, https://doi.org/10.1007/978-3-642-33715-4_54.

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, https://doi.org/10.1167/7.14.4.

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, https://doi.org/10.1167/11.5.5.

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, https://doi.org/10.1037/0033-295X.113.4.766.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7), 4, 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, https://doi.org/10.1167/5.1.8.

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, https://doi.org/10.1038/s41562-017-0058.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*(4), 787–787, https://doi.org/10.1037/a0013118.

Zelinsky, G. 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.

Zelinsky, G. J., 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). Camridge, MA: 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.