Title

Searching for meaning: Local scene semantics guide attention during natural visual search in scenes.

Permalink

https://escholarship.org/uc/item/7b42z9c4

Journal

Quarterly Journal of Experimental Psychology, 76(3)

Authors

Peacock, Candace
Singh, Praveena
Hayes, Taylor
et al.

Publication Date

2023-03-01

DOI

10.1177/17470218221101334

Peer reviewed

Published in final edited form as:

Q J Exp Psychol (Hove). 2023 March ; 76(3): 632–648. doi:10.1177/17470218221101334.

Searching for meaning: Local scene semantics guide attention during natural visual search in scenes

Candace E Peacock[1,2,*], **Praveena Singh**[3,*], **Taylor R Hayes**[1], **Gwendolyn Rehrig**[2], **John M Henderson**[1,2]

[1]Center for Mind and Brain, University of California, Davis, Davis, CA, USA

[2]Department of Psychology, University of California, Davis, Davis, CA, USA

[3]Center for Neuroscience, University of California, Davis, Davis, CA, USA

Abstract

Models of visual search in scenes include image salience as a source of attentional guidance. However, because scene meaning is correlated with image salience, it could be that the salience predictor in these models is driven by meaning. To test this proposal, we generated meaning maps that represented the spatial distribution of semantic informativeness in scenes, and salience maps which represented the spatial distribution of conspicuous image features and tested their influence on fixation densities from two object search tasks in real-world scenes. The results showed that meaning accounted for significantly greater variance in fixation densities than image salience, both overall and in early attention across both studies. Here, meaning explained 58% and 63% of the theoretical ceiling of variance in attention across both studies, respectively. Furthermore, both studies demonstrated that fast initial saccades were not more likely to be directed to higher salience regions than slower initial saccades, and initial saccades of all latencies were directed to regions containing higher meaning than salience. Together, these results demonstrated that even though meaning was task-neutral, the visual system still selected meaningful over salient scene regions for attention during search.

Keywords

Attention; scene perception; eye movements; meaning; image salience

Article reuse guidelines: sagepub.com/journals-permissions

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

[*]Contributed equally.

Author contributions

C.E.P., P.S., and J.M.H. conceived of the study. P.S. collected the data. C.E.P., J.M.H., T.R.H., and G.R. conceived of the analyses. C.E.P. and P.S. analysed the data. C.E.P. and P.S. wrote the manuscript. C.E.P. prepared the figures. C.E.P., P.S., J.M.H., T.R.H., and G.R. revised/edited the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplementary material

The supplementary material is available at qjep.sagepub.com.

When searching for an object in the real world, observers must select and prioritise information that is relevant for the current task at the expense of irrelevant information. However, the process by which this occurs is unclear. Previous work testing the influence of scene information on attentional prioritisation during visual search has found influences of target features (Malcolm & Henderson, 2009; Navalpakkam & Itti, 2005; Vickery et al., 2005; Wolfe & Horowitz, 2017; Zelinsky, 2008), scene context (Castelhano & Witherspoon, 2016; Neider & Zelinsky, 2006; Pereira & Castelhano, 2014, 2019), image salience (Anderson et al., 2015), and various combinations thereof (Castelhano & Heaven, 2010; Ehinger et al., 2009; Malcolm & Henderson, 2010; Torralba et al., 2006; Wolfe & Horowitz, 2017; Zelinsky et al., 2006, 2020) on eye fixations. Although image salience predicts various behaviours during object search, such as fixation allocation (Henderson et al., 2007) and fast first saccades (Anderson et al., 2015), scene semantics influence search behaviours as well (Cornelissen & Võ, 2017; Hayes & Henderson, 2019; Henderson et al., 2007). Furthermore, because scene semantics and image salience are correlated (Elazary & Itti, 2008; Henderson, 2003; Henderson et al., 2007; Henderson & Hayes, 2017, 2018; Rehrig, Peacock, et al., 2020; Tatler et al., 2011) and effects that appear to be due to image salience often are found to be due to semantics in a variety of tasks (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; 2019a, 2019b, 2020; Rehrig, Hayes, et al., 2020; Rehrig, Peacock, et al., 2020) salience effects in search might also be semantic effects instead.

Image salience and visual search

Image salience, which is defined as the local conspicuity of low-level image features (colours, intensities, and orientations) in a scene, has been implicated as a guidance factor during visual search. Here, image salience has been found to interact with (1) search efficiency, (2) other search guidance factors (e.g., scene context, target features), and (3) the allocation of attention during search. With regard to search efficiency, it has been demonstrated that the salience of targets influences search performance (Biggs et al., 2014; Nothdurft, 2006; Nuthmann et al., 2021; Treisman & Gelade, 1980) such that search is more efficient when targets are more salient in both scenes (Nuthmann et al., 2021) and in target arrays (Biggs et al., 2014; Nothdurft, 2006; Treisman & Gelade, 1980). This work collectively demonstrates that the salience of target objects influences search efficiency.

Image salience has also been found to interact with other sources of guidance in scenes, such as target features and scene context (Ehinger et al., 2009; Torralba et al., 2006). This work demonstrates that target features, scene context, and image salience interact to facilitate visual search. For example, Torralba et al. (2006) found that people tended to look in locally salient regions that were relevant to finding a target object, and a combined scene context and image salience model predicted fixation locations better than image salience alone. However, for scenes in which the target object could be placed anywhere (i.e., coffee mugs), salience better predicted eye movements than the combined model. This suggests that whereas image salience interacts with other guidance factors, it can independently predict search fixations as well, depending on the demands of the task (but see Zelinsky et al., 2006 who combined Gaussian mixtures of computationally-derived target guidance and salience guidance signals, and found the mixture that best predicted fixation behaviour in a search task had a zero salience component).

Prior work has also demonstrated that search fixations are associated with image salience. Indeed, one study demonstrated that short latency, first saccades were more likely to land on a region of the image with high salience than long latency, first (and subsequent) saccades during visual search (Anderson et al., 2015). Another study demonstrated that during visual search in scenes, intensity, contrast, and edge density differed at fixated versus non-fixated scene regions, implicating a role of image salience in fixation allocation during search (Henderson et al., 2007). However, the same regions that differed in image features also differed in ratings of semantic informativeness, suggesting that it might not be the image features, themselves, that predict fixated locations. However, image salience and scene semantics were not directly pitted against one another, so it was unclear whether image salience or semantics were more related to fixation allocation. Although it has been suggested that a semantic analog of salience would likely predict attention better in models of visual search (Ehinger et al., 2009; Henderson et al., 2007), to our knowledge, no studies to date have explicitly tested whether semantics uniquely predict eye movements better than image salience when searching for objects in natural scenes. Given that the semantic informativeness of scene regions may better predict search fixations than image salience, the goal of the present study was to explicitly test whether high-level scene meaning outperforms image salience in predicting fixation allocation during visual search.

Meaning and visual search

The hypothesis that scene meaning would outperform image salience in predicting fixation allocation during search would be consistent with cognitive guidance theory, which proposes that attention is generally guided by cognitive factors, such as semantic knowledge gained from experience, including information about a scene’s likely semantic content and the spatial distribution of that content (Henderson, 2007). Cognitive guidance theory proposes that because scene regions are selected based upon semantic informativeness, attentional priority will be based on the predicted informativeness (meaning) of scene regions, whereas image salience will produce an unranked landscape of targets. Indeed, studies using object-scene consistency manipulations have found that the meaning of objects, rather than the physical salience of image features, guides attention during visual search (Biederman et al., 1982; Henderson et al., 1999, 2007, 2009). The issues with these studies, however, is that they typically modify a single object within a scene, resulting in only a small portion of the scene being useful for analysis.

Meaning maps (Henderson & Hayes, 2017) represent the spatial distribution of local semantic density across a scene, allowing us to study how semantics influence attention during visual search across natural scenes as opposed to a single region. Furthermore, because meaning maps are represented in the same format as image salience (in terms of the local landscape of priority), they allow researchers to directly compare the unique influences of high-level meaning and low-level image salience on attention. Since the introduction of meaning maps, many studies have shown that meaning accounts for significantly more unique variance in attention (operationalised as fixation density) than image salience

(Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019a, 2019b, 2020; Rehrig, Hayes, et al., 2020; Rehrig, Peacock, et al., 2020). This work has been extended to visual search: even when meaning is task-neutral, it continues to guide attention during visual search for embedded letters beyond the role of image salience (Hayes & Henderson, 2019). In total, these studies suggest an obligatory role of task-neutral meaning on attentional selection. However, it is unknown whether task-neutral meaning continues to guide attention during object search in scenes. If this were the case, then this would suggest a mechanism whereby the visual system selects scene regions based upon semantic informativeness and would open the door to study how meaning interacts with other known guidance factors (e.g., scene context, target features) during search.

Present study

In total, the present work tested whether meaning predicts search fixations better than image salience during two object search tasks in which target objects were either present (Experiment 1) or absent (Experiment 2). If attention is more related to task-neutral meaning than to task-neutral image salience during object search, this would provide evidence for a hypothesis that scene regions are, in part, selected based on semantic informativeness during search.

Experiment 1

Method

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., 2019a, 2019b, 2020). To reach 30 acceptable participants, 35 University of California, Davis, undergraduate students with normal to corrected-to-normal vision initially participated in the experiment (26 females, average age = 21.06). All participants were naïve to the purpose of the study and provided verbal consent. Eye movement data from each participant were inspected for excessive artefacts due to blinks or loss of calibration. We then computed the signal percentage ([number of good samples/total number of samples] × 100) for each trial using custom MATLAB code. The signal percentage across trials was averaged for each participant and compared with an a priori 75% criterion for signal. Overall, one participant was excluded based on this criterion due to poor eyetracking quality. Individual trials that had less than 75% signal were also excluded. Overall, only two total trials (0.17% of the total data) were excluded based upon this criterion.

Participants were also excluded if they did not correctly do the task. Here, the percentage of trials in which each participant indicated they had found the target, but the data showed they had not fixated the target, was calculated. If this occurred on over 25% of trials, that entire participant was excluded, resulting in removal of five participants (one of these participants also had poor eyetracking quality as defined above). 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 at 0.01°rms) sampling at 1,000 Hz (SR Research, 2010b).

Participants sat 85 cm away from a 21 in. monitor, so that the scenes subtended approximately 26.5° × 20° of visual angle at 1,024 × 768 pixels. Head movements were mini-mised 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 the 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 9,500°/s^2; SR Research, 2010b). Eye movement data were imported offline into MATLAB 2018a (Mathworks, Inc., Natick, MA) using the EDFConverter tool. The first fixation, always located at the centre of the display as a result of the pretrial fixation marker, was eliminated from analysis.

Because we were interested in search eye movements leading up to target decisions, and not target decision processes themselves, and given that the meaning and salience of the targets was not explicitly controlled, only eye movements leading up to the first fixation on the target were included in the analyses. Furthermore, the target region was also excluded/masked in the fixation density maps, meaning maps, and salience maps. As a result, the correlation analyses did not include the meaning and image salience of the regions containing target objects.

Fixations that landed off the screen and any fixations that were less than 50 ms or greater than 1,500 ms in duration were eliminated as outliers. Saccade amplitudes that landed off the screen (>25°) were also excluded. Fixations corresponding to these saccades were included if they met the other exclusion criteria. This outlier removal process resulted in loss of 0.009% of the data across all subjects. Summary statistics for the eye movement measures can be found in Table S1 in the online Supplementary Material.

Stimuli.—Forty digitised photographs (1,024 × 768 pixels) of indoor and outdoor real-world scenes were selected for this study. Scenes were luminance matched across the set by converting the RGB image of each scene to LAB space (i.e., 0 = *darkest*, 100 = *brightest*), scaling the luminance channel of all scenes from 0 to 1, and then adjusting them to the average luminance of the set. Luminance matching took place prior to meaning and salience map generation. All instruction, calibration, and response screens were also luminance matched to the average luminance ($M = 0.43$ L) of the scenes.

Procedure.—Each experimental session consisted of two practice trials and 40 experimental trials in which the target was present. Participants were instructed to search each scene for a different target object located within that scene. First, a central fixation was shown on the screen for 400 ms to orient participants to the centre of the screen where a word cue would appear (Figure 1). Then, a word cue was presented to them for 800 ms, informing them what the search target would be for that scene (Choe et al., 2017; Rayner et al., 2009). We used a word cue as opposed to a visual cue to encourage observers to use scene-related information rather than target-specific features to find the target (Castelhano & Heaven, 2010; Malcolm & Henderson, 2009; Vickery et al., 2005). Following the word cue, the central fixation cross reappeared for 400 ms prior to the search phase of the experiment (Malcolm & Henderson, 2009; Rayner et al., 2009). The search scene was then presented and participants were given 10,000 ms to locate the target (Figure 1). Participants were instructed to maintain fixation on the target once they found it and to press “Enter” on a keyboard. If the target was not found, the scene automatically timed out after 10,000 ms and the next trial began. Two practice trials were administered before the experiment, 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°. 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.

Map creation

Meaning maps.—For this study, we used the same 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 165 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 US$0.50 per assignment, and all participants provided informed consent. Rating stimuli were the same 40 digitised (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 12,000 unique fine patches (87-pixel diameter) and 4,320 unique coarse patches (205-pixel diameter), for a total of 16,320 scene patches. 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).

Each participant rated 300 random patches extracted from 40 scenes. Participants were instructed to assess the meaningfulness of each patch based on how informative or recognisable 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-type 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 judgements. Each unique patch was rated three times by three independent raters for a total of 48,960 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. Because the location of the search targets was not explicitly controlled for in the current study, no centre bias was added to the meaning maps to better account for eye movements directed to targets in scene peripheries.

Fixation density maps.—Fixation density maps were generated from the eye movement data (Figure 2; Figure S1). 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 cut-off frequency of −6 dB (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 Salience Benchmark code.[1]

Unbiased graph-based visual salience.—Graph-based visual salience (GBVS) is a prominent salience model that combines maps of low-level image features to create salience maps (Harel et al., 2006). Salience maps for each scene were generated using the GBVS toolbox with default settings (Figure 2; Figure S1). See "makeGBVSParams.m" for these default parameters.

Centre bias is a natural feature of GBVS salience maps. However, because the location of the search targets was not explicitly controlled for in the current study, we opted to use meaning and salience maps that did not contain centre bias to account for eye movements directed to targets in the periphery. To generate unbiased GBVS maps, we used the whitening method (Rahman & Bruce, 2015), a two-step normalisation approach in which each salience map is normalised to have 0 mean and unit variance. After this, a second, pixel-wise normalisation is performed so that each pixel location across all the salience maps has 0 mean and unit variance (Figure 2; Figure S1).

Histogram matching.—To equate the power of the meaning and salience maps in terms of total density, image histogram matching was used with the fixation density map for each scene serving as the reference image for the corresponding meaning and salience maps (Henderson & Hayes, 2017). Image histogram matching is desirable because it normalises an input image to a reference image, ensuring that the distribution of power in the two maps is similar, thus making them more equally comparable. Using the ground-truth fixation density maps as the reference for both meaning and salience allowed us to directly compare the meaning and salience maps. The "imhistmatch" function from the MATLAB Image Processing Toolbox was used to accomplish image histogram matching.

Results

Correlation analysis.—Following our past work, we used linear correlations (Pearson's *r*) to test the degree to which meaning and image salience accounted for variance in fixation density maps. We chose linear correlation because it is easy to interpret, it operates at the map-level, it is sensitive to small differences in predictors, it makes few assumptions, it is intuitive, it can be visualised, it generally balances the various positives and negatives of different analysis approaches, and it allows us to tease apart variance due to salience and meaning (Bylinskii et al., 2019). In addition, Pearson's *r* is dimensionless, which makes it easy to compare across studies. Two-tailed, paired *t*-tests were used to statistically test the

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

relative ability of the salience and meaning maps to linearly predict the fixation density maps.

Given that the meaning and salience maps were significantly correlated to each other in the present study (*M* = 0.19, *SD* = 0.16) as evidenced by a one-sample *t*-test ($t(39) = 7.32$, $p < .001$, 95% CI = [0.13, 0.22]) and that our primary research question concerned the ability of meaning and salience to *independently* guide visual search, we also used semi-partial correlations. 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. Both *t*-tests are reported with a 95% confidence interval (CI) to indicate the range of values that are 95% certain to contain the difference between the groups. If the 95% CI overlaps with 0, then there is no statistical difference between the groups.

For the linear correlations, meaning explained significantly more of the variance in fixation density (*M* = 0.15, *SD* = 0.13) than image salience (*M* = 0.05, *SD* = 0.06): $t(39) = 5.13$, $p < .001$, 95% CI = [0.06, 0.14], achieved Cohen's *d* = 0.99.[2] For the semi-partial correlations, meaning captured 13% of the unique variance in fixation density (*M* = 0.13, *SD* = 0.11): $t(39) = 7.46$, $p < .001$, 95% CI = [0.10, 0.17], and salience explained 3% of the unique variance (*M* = 0.03, *SD* = 0.04): $t(39) = 4.77$, $p < .001$, 95% CI = [0.02, 0.05] (Figure 3). Although meaning and image salience capture significant unique variance in fixation density, meaning is a significantly better predictor of fixation density than image salience.[3] For a visualisation of the scenes and maps that correspond to each scene number, see Figure S1.

We note that the true amount of variance in attention that can be accounted for is well below 100%. To demonstrate this, we estimated the expected maximum for meaning and salience to account for attention in each scene by performing a leave one out cross-validation (LOOCV) (Henderson & Hayes, 2018; Torralba et al., 2006). Here, we computed a fixation density map for each scene for 29 subjects and a test map for the 30th subject. The linear correlation of the group and test maps was computed, and this was repeated for all 30 subjects. Mean correlations by scene and across scenes were then generated. The results are shown as grey, dotted lines in the top panel of Figure 3, with the left panel showing the mean linear correlation for each scene and the right panel showing the grand mean across

[2] To determine whether we obtained adequate effect sizes for the primary comparison of interest, we conducted a sensitivity analysis using G*Power 3.1 (Faul et al., 2007, 2009). We computed the effect size index *d* (Cohen, 1977) and the critical *t* statistic for a two-tailed paired *t*-test with 95% power and a sample size of 40 scenes using the mean and *SD* of differences (*M* = 0.10; *SD* = 0.12). The analysis revealed a critical *t* value of 2.02, a minimum *d* of 0.81, and 0.99 achieved power. Overall, the achieved Cohen's *d* (0.99) was larger than the minimum *d* (0.81).

[3] The graph-based visual salience (GBVS) salience map used was centre emergent, which means that it has implicit centre bias. To ensure that the advantage of meaning over salience was not due to the whitening process that removed the centre bias, we replicated the main result using the Itti et al. (1998) salience model that did not integrate centre bias into its computation. To generate the Itti et al. salience map, we used the "ittikochmap" function from the GBVS toolbox with the following settings: channels = "CIO," unCenterBias = 1, useIttiKochInsteadOfGBVS = 1. Overall, the pattern of results held. When considering the linear relationship between meaning and fixation densities (*M* = 0.15, *SD* = 0.13) and image salience and fixation densities (*M* = 0.09, *SD* = 0.10), meaning predicted attention significantly better than image salience: $t(39) = 2.92$, $p < .01$, 95% CI = [0.02, 0.10]. Meaning (*M* = 0.10, *SD* = 0.11; $t(39) = 6.15$, $p < .001$, 95% CI = [0.07, 0.14]) and image salience (*M* = 0.04, *SD* = 0.05; $t(39) = 5.20$, $p < .001$, 95% CI = [0.03, 0.06]) both uniquely predicted eye movements.

scenes. Across all scenes for Experiment 1, the cross-validation R^2 was 0.24 (SD = 0.08). Given these theoretical ceilings, meaning is explaining approximately 63% (i.e., 0.15/0.24 = 0.63) of the total variance and salience is explaining approximately 21% (i.e., 0.05/0.24 = 0.21) of the total variance. We note that meaning is explaining a large partition (over half) of the variance, which is surprising considering all the other constraints on search. Meaning and salience performed significantly worse than the theoretical maximum as demonstrated by paired t-tests, meaning: $t(39) = -4.54$, $p < .001$, 95% CI = [0.11, 0.19]; salience: $t(39) = -19.62$, $p < .001$, 95% CI = [0.03, 0.07], but this is not surprising considering the other factors that are important during search, such as target features.

Ordinal fixation analysis.—It has been hypothesised that early fixations may be more directly controlled by image salience than by meaning (Anderson et al., 2015; Borji et al., 2013). To test this hypothesis, we conducted an additional analysis focused specifically on early fixations to examine whether meaning still accounted for significantly more variance in fixation density compared with image salience. For the linear correlations, we used two-tailed paired t-tests to test whether meaning showed an advantage over salience for the first three fixations on each scene. For the semi-partial correlations, we used two-tailed one-sample t-tests against 0 to test whether meaning and salience significantly explained unique variance. All p-values were adjusted for multiple comparisons with the false discovery rate (FDR) correction. As shown in Figure 4, the linear correlations showed that meaning (Fixation 1: M = 0.07, SD = 0.11; Fixation 2: M = 0.09, SD = 0.12; Fixation 3: M = 0.08, SD = 0.09) significantly explained fixation density better than salience (Fixation 1: M = 0.01, SD = 0.03; Fixation 2: M = 0.03, SD = 0.05; Fixation 3: M = 0.03, SD = 0.04) on the first three fixations (all ps < .05) (Figure 4).

The semi-partial correlation analyses showed that for the first (M = 0.06, SD = 0.10), second (M = 0.08, SD = 0.11) and third (M = 0.07, SD = 0.09) fixations, meaning explained significant unique variance (all FDR corrected ps < .001). The same was true for image salience (Fixation 1: M = 0.01, SD = 0.02; Fixation 2: M = 0.03, SD = 0.04; Fixation 3: M = 0.02, SD = 0.03) (all FDR corrected ps < .05).

These results replicate and extend previous meaning map results (Hayes & Henderson, 2019; Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019a, 2019b, 2020; Rehrig, Hayes, et al., 2020; Rehrig, Peacock, et al., 2020). The present findings show that meaning is a better predictor of the spatial distribution of attention than image salience during visual search for objects (Figures 3 and 4). This result suggests that while image salience has been implicated as a factor in guiding attention during visual search (Ehinger et al., 2009; Torralba et al., 2006; Treisman & Gelade, 1980; Wolfe et al., 1989; Wolfe & Horowitz, 2017), models of visual search may wish to revise their models to incorporate meaning.

Fast first saccades.—It has been shown that fast initial saccades are driven to higher salience regions than slower initial saccades during visual search (Anderson et al., 2015). However, given that the meaning map literature has shown that what at first appear to be early salience effects are due to meaning, we were interested to test whether the same might hold true for fast first saccades. To test whether fast first saccades were driven by meaning or image salience, we divided initial saccade latencies into quartile bins. These bins were determined by subject as fixation durations vary substantially across individuals (Castelhano & Henderson, 2008; Henderson & Luke, 2014; Luke et al., 2018). We then generated fixation density maps for each scene based upon the landing location of the fixation following the initial saccade separated by saccade latency quartile. Specifically, there were four total fixation density maps for each scene corresponding to the landing position following the first, second, third, and fourth quartile of saccade latencies across all subjects. We then computed the linear and unique correlations between meaning and salience for each fixation density map. We then compared the meaning and salience correlations following the initial saccade latencies in the first quartile (i.e., the fastest saccades) with the meaning and salience correlations in quartiles 2 through 4 (i.e., the slower quartiles) via two-tailed paired sample *t*-tests.

Overall, for the linear correlations, fast first saccades (*M* = 0.03, *SD* = 0.10) were not directed to regions with higher meaning than slower first saccades (*M* = 0.05, *SD* = 0.08): *t*(39) = 1.23, *p* = .22, 95% CI = [−0.02, 0.06]. The same pattern held for image salience (fast first saccades: *M* = 0.01, *SD* = 0.03; slower first saccades: *M* = 0.01, *SD* = 0.03): *t*(39) = 0.99, *p* = .33, 95% CI = [−0.01, 0.02], contrary to the predictions of the hypothesis that fast first saccades are more driven by image salience (Figure 4). We also directly compared whether first saccades at each latency quartile bin were significantly directed to more meaningful or more salient scene regions via two-tailed paired *t*-tests with *p*-values adjusted with the FDR correction. Saccades at the first and second saccade latency quartiles were numerically (but not significantly) directed to meaning (quartile 1: *M* = 0.03, *SD* = 0.10; quartile 2: *M* = 0.04, *SD* = 0.09) over image salience (quartile 1: *M* = 0.01, *SD* = 0.03; quartile 2: *M* = 0.01, *SD* = 0.03); all *p*s > .05. Saccades at the third and fourth saccade latency quartiles were significantly more directed to meaning (quartile 3: *M* = 0.07, *SD* = 0.12; quartile 4: *M* = 0.05, *SD* = 0.08) than to salience (quartile 3: *M* = 0.02, *SD* = 0.03; quartile 4: *M* = 0.02, *SD* = 0.03); all *p*s < .05.

For the unique correlations, fast first saccades (*M* = 0.03, *SD* = 0.10) were not directed to regions with higher meaning than slower first saccades (*M* = 0.05, *SD* = 0.08): *t*(39) = 1.16, *p* = .25, 95% CI = [−0.02, 0.06]. The same held true for image salience (fast first saccades: *M* = 0.01, *SD* = 0.03; slower first saccades: *M* = 0.01, *SD* = 0.02): *t*(39) = 0.72, *p* = .48, 95% CI = [−0.01, 0.02] (Figure 4). We also used two-tailed one-sample *t*-tests against 0 to test whether meaning and salience significantly explained unique variance for each saccade latency quartile. Overall, meaning explained significant unique variance for quartiles 2 through 4 (quartile 2: *M* = 0.04, *SD* = 0.09; quartile 3: *M* = 0.06, *SD* = 0.11; quartile 4: *M* = 0.05, *SD* = 0.08; all *p*s < .05) but not for quartile 1 (*M* = 0.03, *SD* = 0.10; *p* = .10). The same was true for image salience. Here, salience explained significant unique variance for quartiles 2 through 4 (quartile 2: *M* = 0.01, *SD* = 0.02; quartile 3: *M* = 0.01, *SD* = 0.02; quartile 4: *M* = 0.01, *SD* = 0.02; all *p*s < .01) but not for quartile 1 (*M* = 0.01, *SD* = 0.03; *p* = .08).

Overall, the Experiment 1 results demonstrated no fast first effect for either meaning or salience when considering the linear and unique correlations. Furthermore, all saccade latencies were either numerically or significantly more directed to meaning than to image salience.

Experiment 2

It could be that because the Experiment 1 scenes contained target objects, the meaning or the salience of the target objects themselves drove the Experiment 1 results (Biggs et al., 2014; Nothdurft, 2006; Nuthmann et al., 2021; Treisman & Gelade, 1980). To ensure that the effect of interest was not driven by the target present task from Experiment 1, we conducted a second visual search study in which targets were present or absent in scenes. Data analysis focused on target absent scenes so that influences of the target itself on eye movements would be eliminated.

Method

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., 2019a, 2019b, 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 artefacts 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] × 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, 0 participants were excluded based on this criterion of poor eye tracking quality. Individual trials that had less than 75% eye tracking 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 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 1,000 Hz (SR Research, 2010b). Participants sat 85 cm away from a 21” monitor, so that the scenes subtended approximately 26.5° × 20° of visual angle at 1,024 × 768 pixels. Head movements were mini-mised 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 9,500°/s^2; SR Research, 2010b). The resulting segmented eye movement data were imported offline into MATLAB using the EDFConverter tool. The first fixation, always located at the centre 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 analysed data from target absent trials.

Fixations that landed off the screen, and any fixations that were less than 50 ms or greater than 1,500 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 >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.—A total of 105 digitised photographs (1,024 × 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) (Figure 5). In all, 10 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.

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. From here, only indoor scenes were used for paintings, as paintings typically reside on indoor walls. Both outdoor urban scenes and indoor scenes were used for garbage bins, as 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. See Figure S2 to visualise the scenes used in the present study.

Procedure.—Each run of the experiment consisted of 6 practice trials and 105 randomised 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 centre 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. Following the word cue, the central fixation cross reappeared for 400 ms prior to the search phase of the experiment. The search scene was then presented for 10 s (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°. 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.

Map creation

Meaning maps.—Meaning maps for the 75 target absent scenes were generated (Figure 5) using the same method as described in Experiment 1.

Fixation density maps.—Fixation density maps were generated using the same method as Experiment 1.

Unbiased GBVS.—The unbiased GBVS maps were produced using the same method as described in Experiment 1.

Histogram matching.—Image histogram matching was conducted in the same manner as Experiment 1.

Results

Correlation analysis.—As with Experiment 1, the correlation between meaning and salience ($M = 0.20$, $SD = 0.15$) was significant as evidenced by a one-sample t-test: $t(74) = 11.35$, $p < .001$, 95% CI = [0.17, 0.24]. Meaning explained significantly greater linear variance in fixation density ($M = 0.22$, $SD = 0.15$) relative to image salience ($M = 0.06$, $SD = 0.08$) as evidenced by a paired t-test: $t(74) = 7.90$, $p < .001$, 95% CI = [0.12, 0.19]. For the semi-partial correlations, meaning explained significant unique variance ($M = 0.19$, $SD = 0.14$) as per a one-sample t-test: $t(74) = 12.13$, $p < .001$, 95% CI = [0.16, 0.22]. The same held true for image salience ($M = 0.04$, $SD = 0.06$): $t(74) = 5.96$, $p < .001$, 95% CI = [0.03, 0.05]. These analyses suggest that, irrespective of whether scenes contain target objects (Experiment 1) or not (Experiment 2), meaning still explains linear and unique variance as people search for objects in scenes.

As with Experiment 1, we estimated the expected maximum for meaning and salience to account for attention in each scene by performing an LOOCV (Henderson & Hayes, 2018; Torralba et al., 2006). The results are shown as grey, dotted lines in the top panel of Figure 6, with the left panel showing the mean linear correlation for each scene and the right panel showing the grand mean across scenes. Across all scenes for Experiment 2, the cross-validation R^2 was 0.38 ($SD = 0.10$). Given this theoretical ceiling, meaning is explaining approximately 58% (i.e., 0.22/0.38 = 0.58) of the total variance and salience is explaining approximately 16% (i.e., 0.06/0.38 = 0.16) of the total variance. As with Experiment 1, meaning explained over half of the available variance, which is large considering the other factors that guide search. As with Experiment 1, meaning and salience performed significantly worse than the theoretical maximum as demonstrated by paired t-tests (meaning: $t(74) = -8.86$, $p < .001$, 95% CI = [−0.20, −0.13]; salience: $t(74) = -20.46$, *p* < .001, 95% CI = [0.35, 0.28]) but again this is not surprising considering the other factors that are important during search.

Ordinal fixation analysis.—Our next analysis tested whether early fixations are more directly controlled by image salience or meaning. To test this hypothesis, we conducted an ordinal fixation analysis to test whether meaning still accounted for significantly more variance than image salience on early fixations. Here, the linear correlations showed that meaning (Fixation 1: *M* = 0.09, *SD* = 0.13; Fixation 2: *M* = 0.10, *SD* = 0.14) was significantly more related to fixation densities than salience (Fixation 1: *M* = 0.02, *SD* = 0.03; Fixation 2: *M* = 0.04, *SD* = 0.07) for the first two fixations as per paired *t*-tests corrected with the FDR correction (all *p*s < .05) (Figure 7). There was no significant difference between meaning (*M* = 0.08, *SD* = 0.10) and salience (*M* = 0.04, *SD* = 0.08) for the third fixation (*p* > .05).

The semi-partial correlation analyses showed that for the first (*M* = 0.09, *SD* = 0.12), second (*M* = 0.10, *SD* = 0.13), and third (*M* = 0.07, *SD* = 0.09) fixations, meaning explained significant unique variance as per one-sample *t*-tests (all FDR corrected *p*s < .001). The same was true for image salience (Fixation 1: *M* = 0.01, *SD* = 0.03; Fixation 2: *M* = 0.03, *SD* = 0.06; Fixation 3: *M* = 0.04, *SD* = 0.08; all FDR corrected *p*s < .05). Overall, the ordinal fixation analyses for Experiment 2 replicated Experiment 1, suggesting that the presence of the target object in Experiment 1 did not influence the relationship between meaning and visual search behaviours.

Fast first saccades.—Prior work suggests that fast initial saccades are driven to higher salience regions than slower initial saccades during visual search (Anderson et al., 2015). In the present work, we tested whether this fast first salience effect was actually a fast first meaning effect. As per Experiment 1, the linear correlations in Experiment 2 demonstrated that fast first saccades (*M* = 0.07, *SD* = 0.11) were not directed to regions with higher meaning than slower first saccades (*M* = 0.05, *SD* = 0.08): *t*(29) = −0.72, *p* = .22, 95% CI = [−0.02, 0.06]. The same pattern held for image salience (fast first saccades: *M* = 0.02, *SD* = 0.04; slower first saccades: *M* = 0.03, *SD* = 0.04): *t*(29) = 0.54, *p* = .59, 95% CI = [−0.01, 0.01], contrary to the predictions of the hypothesis that fast first saccades are more driven by image salience (Figure 4). We also directly compared whether first saccades at each latency quartile bin were significantly directed to more meaningful or more salient scene regions via two-tailed paired *t*-tests with *p*-values adjusted with the FDR correction. Saccades at all saccade latency quartiles were significantly directed more to meaning (quartile 1: *M* = 0.07, *SD* = 0.11; quartile 2: *M* = 0.08, *SD* = 0.13; quartile 3: *M* = 0.05, *SD* = 0.10; quartile 4: *M* = 0.05, *SD* = 0.08) than to salience (quartile 1: *M* = 0.02, *SD* = 0.04; quartile 2: *M* = 0.03, *SD* = 0.06; quartile 3: *M* = 0.02, *SD* = 0.06; quartile 4: *M* = 0.03, *SD* = 0.05); all *p*s < .05.

When considering the unique correlations, fast first saccades (*M* = 0.06, *SD* = 0.11) were not directed to regions with higher meaning than slower first saccades (*M* = 0.05, *SD* = 0.07): *t*(29) = −0.92, *p* = .36, 95% CI = [–0.03, 0.01]. The same pattern held for image salience (fast first saccades: *M* = 0.02, *SD* = 0.04; slower first saccades: *M* = 0.02, *SD* = 0.04): *t*(29) = 0.13, *p* = .90, 95% CI = [–0.01, 0.01]. We also used two-tailed one-sample *t*-tests against 0 to test whether meaning and salience significantly explained unique variance for each saccade latency quartile. Overall, meaning explained significant unique variance for all quartiles (quartile 1: *M* = 0.06, *SD* = 0.11; quartile 2: *M* = 0.07, *SD* = 0.13; quartile 3: *M* = 0.05, *SD* = 0.10; quartile 4: *M* = 0.05, *SD* = 0.08); all *p*s < .001. The same held true for salience (quartile 1: *M* = 0.02, *SD* = 0.04; quartile 2: *M* = 0.02, *SD* = 0.06; quartile 3: *M* = 0.02, *SD* = 0.06; quartile 4: *M* = 0.02, *SD* = 0.04); all *p*s < .01.

Overall, the results for Experiment 2 generally replicated Experiment 1. Although there was no fast first effect for either meaning or salience (as per Experiment 1), initial saccades at all latencies were significantly more directed to higher meaning regions than higher salience regions.

Discussion

Prior work has tested how target features (Malcolm & Henderson, 2009; Navalpakkam & Itti, 2005; Vickery et al., 2005; Wolfe & Horowitz, 2017; Zelinsky, 2008), scene context (Castelhano & Witherspoon, 2016; Henderson et al., 1999; Neider & Zelinsky, 2006; Pereira & Castelhano, 2014, 2019) image salience (Anderson et al., 2015), and various combinations of these sources (Castelhano & Heaven, 2010; Ehinger et al., 2009; Malcolm & Henderson, 2010; Torralba et al., 2006; Wolfe & Horowitz, 2017; Zelinsky et al., 2006, 2020) influence eye movements during object search in scenes. However, because meaning and image salience are correlated (Elazary & Itti, 2008; Henderson, 2003; Henderson et al., 2007; Henderson & Hayes, 2017, 2018; Rehrig, Peacock, et al., 2020; Tatler et al., 2011), and because recent work has shown that attention prioritises task-neutral meaning over image salience during visual search for embedded letters in scenes (Hayes & Henderson, 2019), the current study tested whether this pattern of results would also hold during visual search for objects in scenes. If task-neutral meaning does indeed predict search eye movements better than image salience, this then provides support for cognitive guidance theory that suggests that attention is guided by cognitive factors including semantic knowledge during search. To investigate this question, we used a visual search task in which viewers searched for objects in photographs of scenes. To compare meaning and salience, we generated meaning maps, which capture the spatial distribution of scene semantics, and salience maps, which capture the spatial distribution of image salience, and compared how well each predicted attention (as operationalised by fixation density) during search for the target.

Present findings

The main results across two studies indicated that although meaning and salience explained significant unique variance in attention, meaning explained significantly more of the variance in fixation density than image salience. This trend held for ordinal fixation analyses, which showed that attention prioritised meaning over image salience from the earliest points in time, both overall and when only unique variance was considered. The advantage of meaning over salience in predicting search fixations extends the findings of Hayes and Henderson (2019), who found the same pattern of results during visual search for embedded targets, as well as other meaning mapping studies (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019a, 2019b, 2020; Rehrig, Hayes, et al., 2020; Rehrig, Peacock, et al., 2020), and further suggest a role of meaning in visual search. However, we note that these prior studies found no effect of image salience whereas the current study did. Although the ordinal fixation analysis correlations appeared to be low in the present study, this was driven by the use of unbiased meaning and salience maps. In a prior study using both centre-biased and unbiased maps, it was found that centre bias artifactually inflated the variance explained (Peacock et al., 2019a). The correlations between the first, second, and third fixations in the present study are on par with the correlations demonstrated in Peacock et al. (2019a).

It has also been suggested that fixations following fast initial saccade latencies are driven to higher salience regions than slower initial saccade latencies during visual search (Anderson et al., 2015), presumably because image salience is available earlier than scene meaning (Anderson et al., 2015, 2016). To investigate this issue here, we expanded the time course analyses by testing how meaning and image salience influenced the first attended location as a function of the latency of the saccade from the initial experimenter-defined central start location. Contrary to the hypothesis that fixations following fast initial saccades are more driven by image salience, our results showed that fast first saccades are no more likely to be directed to regions with significantly higher salience than are slower first saccades. The same pattern was observed for meaning: fast first saccades are no more likely to be directed to regions with higher meaning than are slower first saccades. Importantly, we also found that first saccades of all latencies, including fast first saccades, were directed to regions with greater meaning than image salience, although these effects were stronger in Experiment 2. These results suggest that in the context of search for objects in natural scenes, meaning plays an important role in guiding overt attention from the earliest points in time. This result is also consistent with recent evidence that the spatial distribution of meaning in scenes is available almost as quickly as the spatial distribution of salience, and both well within the latency of the first saccade (Kiat et al., 2022).

Importantly, the observed pattern of results replicated across two studies, in which targets were either present or absent. This pattern suggests that irrespective of the meaning and salience of target objects or the target features themselves, meaning plays a role in search.

Mechanism

Cognitive guidance theory states that attention is guided by scene regions that are semantically informative and cognitively relevant (Henderson, 2007; Henderson & Hayes, 2017, 2018; Henderson et al., 2018). The present study provides evidence of cognitive guidance theory suggesting that people use their stored semantic knowledge to guide their search processes. The present work suggests that even though meaning is not relevant to the search task, people cannot ignore (or "turn off") the meaning of what they are looking at. Indeed, prior scene studies have shown that people cannot help but look at task-neutral scene meaning during search for bright patches (Peacock et al., 2019a) or during search for letters overlaid on top of a scene (Cornelison & Võ, 2017; Hayes & Henderson, 2019). Furthermore, recent physiological work shows that the distribution of meaning is computed rapidly (prior to the first fixation) and demonstrates a neural basis for meaning-based guidance of attention that is distinct from image salience (Kiat et al., 2022). Together, the findings of the present work and that of past work demonstrate that irrespective of what a viewer is doing, a scene and its objects carry semantic content (e.g., a kitchen is a kitchen irrespective of whether you're making dinner or looking for a coffee cup) and viewers cannot help but use that semantic content to inform where they move their eyes next. This suggests that people use their semantic knowledge gained from experience to rapidly make predictions about a scene's likely semantic content and the spatial distribution of that content for attentional prioritisation (Henderson, 2017). Given the obligatory role of meaning during search, it stands to reason that models of search may wish to incorporate task-neutral meaning into their models (see the following section).

Other guidance factors

Studies have shown that target features (Malcolm & Henderson, 2009; Navalpakkam & Itti, 2005; Vickery et al., 2005; Wolfe & Horowitz, 2017; Zelinsky, 2008), scene context (Castelhano & Witherspoon, 2016; Neider & Zelinsky, 2006; Pereira & Castelhano, 2014, 2019), memory (Draschkow et al., 2014; Võ & Wolfe, 2013), and other guidance factors influence search fixations. Despite the many constraints on fixation placement during visual search, the present study still finds (surprisingly) that task-neutral meaning continues to be important. In fact, the present study extends a recent finding that the expected locations of target objects interact with the semantic informativeness (i.e., meaning) of scene regions to predict where people look during object search (Peacock et al., 2021). Together, the present findings in the context of prior work suggest that even though meaning maps are generated in a task-neutral fashion, the visual system selects scene regions based on semantic informativeness as represented by those maps, and these scene regions interact with other sources of search guidance. Given these findings, models of search should consider incorporating meaning with other sources of guidance.

Indeed, an interesting future avenue for research will be to test how meaning interacts with other sources of guidance, such as target object features in the search target template. Zelinsky et al. (2006) combined Gaussian mixtures of computationally derived target guidance and salience guidance signals, and found the mixture that best predicted fixation behaviour in a search task had a zero salience component. However, it is unknown whether the same is true for a combination of target features and meaning, or whether target features and meaning would interact to enhance prediction of search guidance. We predict that (1) because our results replicated across target present and absent tasks in which meaning explained a nontrivial 58% and 63% of the theoretical ceiling of variance in attention, respectively, and (2) because it is unlikely that task-neutral meaning is perfectly correlated with target features, both meaning and target features would likely add unique variance to a model of search incorporating the two signals.

Object guidance during search

Prior work has demonstrated that during non-search tasks, objects predict fixation allocation (Castelhano et al., 2009; Chen & Zelinsky, 2019; Cronin et al., 2020; Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Nuthmann et al., 2020; Pajak & Nuthmann, 2013; Stoll et al., 2015; ‘t Hart et al., 2013). Because meaning maps have an object bias and explain search fixations independent of image salience in the present study, it could be the case that objects, in general, are also important for fixation allocation during search. Future work could test whether objects indeed guide attention during search independent of meaning and whether object information interacts with meaning and other guidance factors during search. Knowledge of this would provide converging evidence that general semantic information beyond known search guidance factors are important for search.

Limitations

It could be that the scenes used here were biased in favour of meaning rather than visual salience. When considering the whole scene analysis, meaning had an advantage over image salience in 75% of the scenes (30 of 40 scenes) in Experiment 1 and 77% of the scenes (58 out of 75 scenes) in Experiment 2. Although there are no apparent systematic differences between these scenes and the scenes in which image salience had the advantage in the current study (see Figures S1 and S2), there may be circumstances in which image salience is a better predictor of attention, such as in certain scene categories or tasks. For example, in Experiment 2, salience explained search fixations better than image salience 16% of the time for drinking glasses and paintings and 36% of the time during search for garbage bins, which suggests that salience may have an advantage depending on the task or scenes used. Future meaning mapping studies may wish to use a larger scene set to address whether there are systematic differences in scenes in which meaning or image salience has the advantage.

Salience maps are valuable to computational modelling of search, as they can be obtained for any new image, even images that a model has never seen. For meaning maps to be as easy to apply to computational modelling of search behaviour as image salience, it will be necessary to generate meaning maps for a broader set of scenes than currently exists. One way to automate meaning map generation would involve using a deep learning model to compute meaning maps without depending on crowd-sourcing. It is important to note, however, that our purpose in the present study was not to generate a prediction machine for computer vision, but rather to understand the factors that the human brain uses to guide attention. Therefore, it was not critical for the current question that there be an easy way to take an arbitrary scene and automatically generate a meaning map for it. However, developing models to generate image-computable meaning maps is an important exploration for future modelling endeavours.

Newer (deep) salience models exist (e.g., DeepGaze II: Kümmerer et al., 2016; ICF: Kümmerer et al., 2017). We elected not to use these models in the current study because *deep salience* models are trained and optimised on fixation data and therefore capture additional regularities in looking behaviour beyond image salience (e.g., people are more likely to make short rather than long saccades). Furthermore, a recent study found that deep salience models are correlated with both meaning and image salience (Hayes & Henderson, 2021). Trying to pull apart, then, what deep salience maps are capturing is not a trivial problem given that eye movements are related to many factors. Conversely, *image salience* models simply capture conspicuous low-level image features. Therefore, given our theoretical focus specifically on image salience, it was the more appropriate representation to answer the question of whether high-level scene semantics or low-level image salience guides attention during visual search.

We note that whereas deep salience models are correlated with meaning (Henderson & Hayes, 2021), that does not necessarily mean that these models extract meaning in the same way as human raters, or indeed at all. Hayes and Henderson (2022) found that people rate diffeomorphed patches (patches in which meaning has been removed but visual features retained) to have significantly less meaning than non-diffeomorphed patches, whereas deep net models show a moderate increase. This result suggests that meaning maps indeed reflect local semantic content in scenes, whereas deep salience models reflect something else.

Conclusion

The current work sought to understand how meaning and image salience guide attention during visual search for objects in real-world scenes. We found that meaning accounted for significantly more variance in fixation density than image salience, both overall and early during viewing, and that fast initial saccades were not directed to higher salience regions than slower initial saccades. These findings extend the visual search literature, suggesting that meaning contributes to attentional selection during visual search in real-world scenes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Conner Daniels and Jocelyn Huerta for assisting in data collection.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Eye Institute of the National Institutes of Health (grant no. 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).

References

Anderson NC, Donk M, & Meeter M (2016). The influence of a scene preview on eye movement behavior in natural scenes. Psychonomic Bulletin & Review, 23(6), 1794–1801. [PubMed: 27073087]

Anderson NC, 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. 10.1167/15.5.9

Biederman I, Mezzanotte RJ, & Rabinowitz JC (1982). Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology, 14, 143–177. [PubMed: 7083801]

Biggs AT, Adamo SH, & Mitroff SR (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. Acta Psychologica, 152, 158–165. 10.1016/j.actpsy.2014.08.005 [PubMed: 25226547]

Borji A, Sihite DN, & Itti L (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing, 22(1),55–69. 10.1109/TIP.2012.2210727 [PubMed: 22868572]

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. [PubMed: 29993800]

Castelhano MS, & Heaven C (2010). The relative contribution of scene context and target features to visual search in scenes. Attention, Perception, & Psychophysics, 72(5), 1283–1297.

Castelhano MS, & Henderson JM (2008). The influence of color on the perception of scene gist. Journal of Experimental Psychology: Human Perception and Performance, 34(3),660–675. 10.1037/0096-1523.34.3.660 [PubMed: 18505330]

Castelhano MS, Mack ML, & Henderson JM (2009). Viewing task influences eye movement control during active scene perception. Journal of Vision, 9(3), Article 6. 10.1167/9.3.6

Castelhano MS, & Witherspoon RL (2016). How You Use It Matters. Psychological Science, 27(5), 606–621. [PubMed: 27022016]

Chen Y, & Zelinsky GJ (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),139–154. 10.1037/xhp0000593 [PubMed: 30596438]

Choe KW, Kardan O, Kotabe HK, Henderson JM, & Berman MG (2017). To search or to like: Mapping fixations to differentiate two forms of incidental scene memory. Journal of Vision, 17, Article 8. 10.1167/17.12.8

Cohen J (1977). Statistical power analysis for the behavioral sciences (Rev. ed.). Lawrence Erlbaum Associates.

Cornelissen THW, & Võ ML-H (2017). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. Attention, Perception, & Psychophysics, 79(1), 154–168. 10.3758/s13414-016-1203-7

Cronin DA, Hall EH, Goold JE, Hayes TR, & Henderson JM (2020). Eye Movements in Real-World Scene Photographs: General Characteristics and Effects of Viewing Task. Frontiers in Psychology, 10, Article 2915. 10.3389/fpsyg.2019.02915

Draschkow D, Wolfe JM, & Võ ML-H (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. Journal of Vision, 14(8), Article 10. 10.1167/14.8.10

Ehinger K, Hidalgo-Sotelo B, Torralba A, & Oliva A (2009). Modeling visual search in a thousand scenes: The roles of saliency, target features, and scene context. Journal of Vision, 9(8), Article 1199.

Einhäuser W, Spain M, & Perona P (2008). Objects predict fixations better than early saliency. Journal of Vision, 8(14), Article 18. 10.1167/8.14.18

Elazary L, & Itti L (2008). Interesting objects are visually salient. Journal of Vision, 8(3), Article 3. 10.1167/8.3.3

Faul F, Erdfelder E, Buchner A, & Lang A-G (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149–1160. [PubMed: 19897823]

Faul F, Erdfelder E, Lang A-G, & Buchner A (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175–191. [PubMed: 17695343]

Harel J, Koch C, & Perona P (2006). Graph-based visual saliency. In Neural information processing systems (pp. 1–8).

Harel J, Koch C, & Perona P (2006). Graph-Based Visual Saliency. Advances in Neural Information Processing Systems, 19. https://proceedings.neurips.cc/paper/2006/hash/4db0f8b0fc895da263fd77fc8aecabe4-Abstract.html

Hayes TR, & Henderson JM (2019). Scene semantics involuntarily guide attention during visual search. Psychonomic Bulletin & Review, 26, 1683–1689. [PubMed: 31342407]

Hayes TR, & Henderson JM (2021). Deep saliency models learn low-, mid-, and high-level features to predict scene attention. Scientific Reports, 11(1), Article 18434. 10.1038/s41598-021-97879-z

Hayes TR, & Henderson JM (2022). Meaning maps detect the removal of local semantic scene content but deep saliency models do not. Attention, Perception, & Psychophysics, 84(3), 647–654. 10.3758/s13414-021-02395-x

Henderson JM (2003). Human gaze control during real-world scene perception. TRENDS in Cognitive Sciences, 7(11), 498–504. [PubMed: 14585447]

Henderson JM (2007). Regarding scenes. Current Directions in Psychological Science, 16(4), 219–222.

Henderson JM (2017). Gaze control as prediction. Trends in Cognitive Sciences, 21(1), 15–23. [PubMed: 27931846]

Henderson JM, Brockmole JR, Castelhano MS, & Mack ML (2007). Visual saliency does not account for eye movements during visual search in real world scenes. In Gompel RPGV, Fischer MH, Murray WS & Hill RL (Eds.), Eye movements: A window on mind and brain (pp. 537–562). Elsevier Ltd. 10.1167/9.3.6

Henderson JM, & Hayes TR (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. Nature Human Behaviour, 1, 743–747. 10.1038/s41562-017-0208-0

Henderson JM, Hayes TR, Rehrig G, & Ferreira F (2018). Meaning guides attention during real-world scene description. Scientific Reports, 8, Article 13504.

Henderson JM, & Luke SG (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. Journal of Experimental Psychology: Human Perception and Performance, 40(4), 1390–1400. [PubMed: 24730735]

Henderson JM, Malcolm GL, & Schandl C (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. Psychonomic Bulletin & Review, 16(5), 850–856. [PubMed: 19815788]

Henderson JM, Weeks PA, & 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. 10.1037/0096-1523.25.1.210

Itti L, Koch C, & Niebur E (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259.

Kiat J, Hayes TR, Henderson JM, & Luck SJ (2022). Rapid extraction of the spatial distribution of physical saliency and semantic informativeness from natural scenes in the human brain. Journal of Neuroscience, 42(1), 97–108. [PubMed: 34750229]

Kümmerer M, Wallis TSA, & Bethge M (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. ArXiv:1610.01563 [Cs, q-Bio, Stat]. http://arxiv.org/abs/1610.01563

Kümmerer M, Wallis TS, & Bethge M (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563.

Luke SG, Darowski ES, & Gale SD (2018). Predicting eye-movement characteristics across multiple tasks from working memory and executive control. Memory & Cognition, 46(5), 826–839. [PubMed: 29484579]

Malcolm GL, & Henderson JM (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. Journal of Vision, 9(11), Article 8.

Malcolm GL, & Henderson JM (2010). Combining top-down processes to guide eye movements during real-world scene search. Journal of Vision, 10(2), Article 4. 10.1167/10.2.4

Navalpakkam V, & Itti L (2005). Modeling the influence of task on attention. Vision Research, 45, 205–231. [PubMed: 15581921]

Neider MB, & Zelinsky GJ (2006). Scene context guides eye movements during visual search. Vision Research, 46(5),614–621. 10.1016/j.visres.2005.08.025 [PubMed: 16236336]

Nothdurft H-C (2006). Salience and target selection in visual search. Visual Cognition, 14(4–8),514–542. 10.1080/13506280500194162

Nuthmann A, Clayden AC, & Fisher RB (2021). The effect of target salience and size in visual search within naturalistic scenes under degraded vision. Journal of Vision, 21(4), Article 2. 10.1167/jov.21.4.2

Nuthmann A, & Henderson JM (2010). Object-based attentional selection in scene viewing. Journal of Vision, 10(8), Article 20. 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), Article 22057. 10.1038/s41598-020-78203-7

Pajak M, & Nuthmann A (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. Journal of Vision, 13(5), Article 2. 10.1167/13.5.2

Peacock CE, Cronin DA, Hayes TR, & Henderson JM (2021). Meaning and expected surfaces combine to guide attention during visual search in scenes. Journal of Vision, 21(11), Article 1. 10.1167/jov.21.11.1

Peacock CE, Hayes TR, & Henderson JM (2019a). Meaning guides attention during scene viewing even when it is irrelevant. Attention, Perception, & Psychophysics, 81, 20–34. 10.3758/s13414-018-1607-7

Peacock CE, Hayes TR, & Henderson JM (2019b). The role of meaning in attentional guidance during free viewing of real-world scenes. Acta Psycholigica, 198, Article 102889. 10.1016/j.actpsy.2019.102889

Peacock CE, Hayes TR, & Henderson JM (2020). Center bias does not account for the advantage of meaning over salience in attentional guidance during scene viewing. Frontiers in Psychology, 11, Article 1877. 10.3389/fpsyg.2020.01877

Pereira EJ, & Castelhano MS (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. [PubMed: 25089577]

Pereira EJ, & Castelhano MS (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. Psychonomic Bulletin & Review, 26, 1273–1281. 10.3758/s13423-019-01610-z [PubMed: 31161527]

Rahman S, & Bruce N (2015). Visual saliency prediction and evaluation across different perceptual tasks. PLOS ONE, 10, Article e0138053. 10.1371/journal.pone.0138053

Rayner K, Smith TJ, Malcolm GL, & Henderson JM (2009). Eye movements and visual encoding during scene perception. Psychological Science, 20, 6–10. [PubMed: 19037907]

Rehrig G, Hayes TR, Henderson JM, & Ferreira F (2020). When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention. Memory & Cognition, 48, 1181–1195. 10.3758/s13421-020-01050-4 [PubMed: 32430889]

Rehrig G, Peacock CE, Hayes TR, Henderson JM, & 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, 1659–1681. 10.1037/xlm0000837 [PubMed: 32271065]

SR Research. (2010a). Experiment Builder user’s manual. SR Research Ltd.

SR Research. (2010b). EyeLink 1000 user’s manual, version 1.5.2. SR Research Ltd.

Stoll J, Thrun M, Nuthmann A, & Einhäuser W (2015). Overt attention in natural scenes: Objects dominate features. Vision Research, 107, 36–48. 10.1016/j.visres.2014.11.006 [PubMed: 25478891]

‘t Hart BM, Schmidt HCEF, Roth C, & Einhauser W (2013). Fixations on objects in natural scenes: Dissociating importance from salience. Frontiers in Psychology, 4, Article 455. 10.3389/fpsyg.2013.00455

Tatler BW, Hayhoe MM, Land MF, & Ballard DH (2011). Eye guidance in natural vision: Reinterpreting salience. Journal of Vision, 11(5), Article 5.

Torralba A, Oliva A, Castelhano MS, & Henderson JM (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. [PubMed: 17014302]

Treisman AM, & Gelade G (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97–136. [PubMed: 7351125]

Vickery TJ, King LW, & Jiang Y (2005). Setting up the target template in visual search. Journal of Vision, 5(1), Article 8.

Võ MLH, & Wolfe JM (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. Cognition, 126(2),198–212. 10.1016/j.cognition.2012.09.017 [PubMed: 23177141]

Wolfe JM, Cave KR, & Franzel SL (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. [PubMed: 2527952]

Wolfe JM, & Horowitz TS (2017). Five factors that guide attention in visual search. Nature Human Behavior, 1(3), Article 58.

Zelinsky GJ (2008). A theory of eye movements during target acquisition. Psychological Review, 115(4), 787–787. [PubMed: 18954205]

Zelinsky GJ, 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, 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 Weiss Y, Schölkopf B & Platt JC (Eds.), Advances in neural information processing systems 18 (pp. 1569–1576). MIT Press.

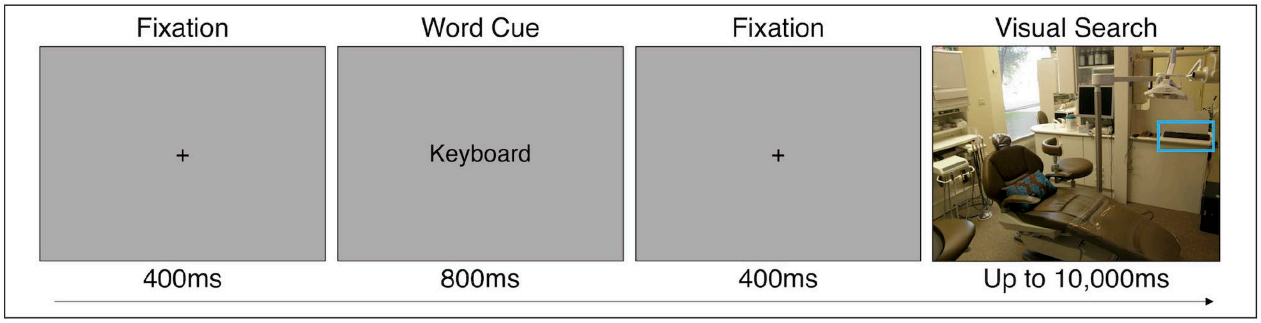

Figure 1.
Trial structure. This figure represents the trial structure for the visual search task. For visualisation purposes, the target location is enclosed in a blue box in the last panel.

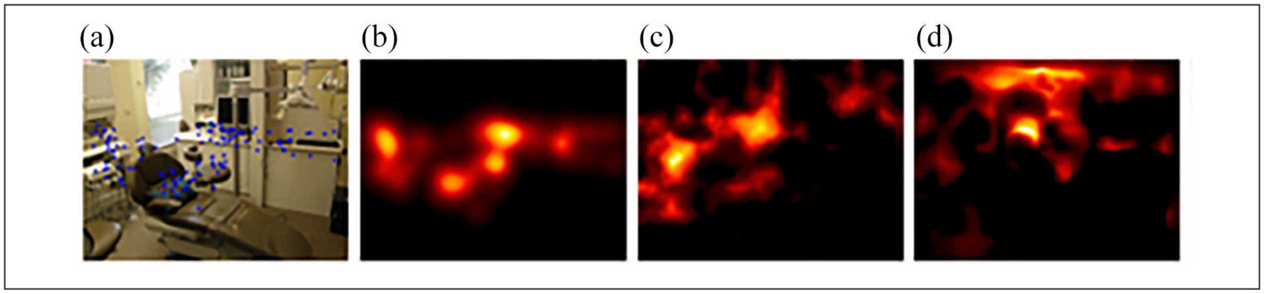

Figure 2.

Example scenes and maps: (a) an example scene used in the study with fixations overlaid; (b) the corresponding fixation density map for the scene; (c) the meaning map; and (d) the salience map for the example scene.

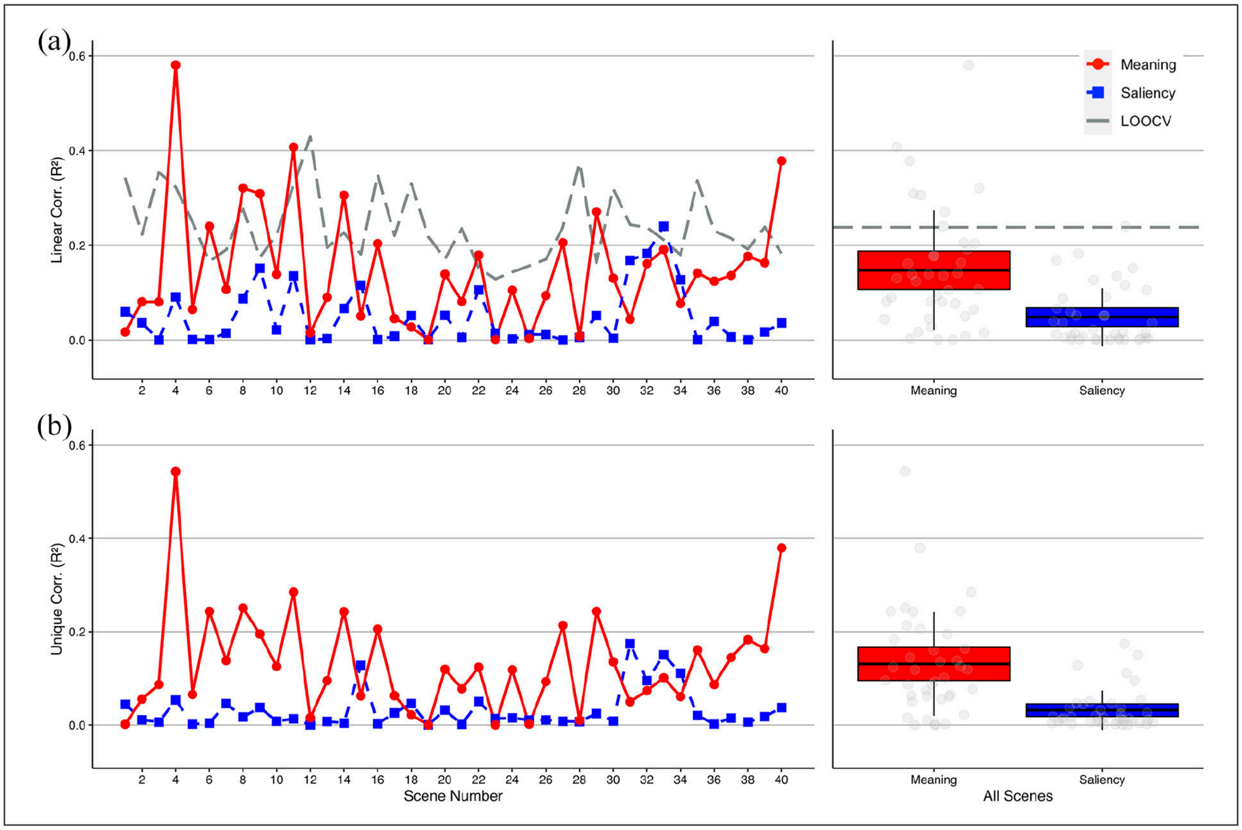

Figure 3.

Experiment 1: meaning versus salience. Line plots show the (a) squared linear and (b) semi-partial correlations between the fixation density maps, meaning (red circles), and image salience (blue squares). The grey, dashed lines in panel (a) correspond to the leave one out cross-validation (LOOCV) that predicted the theoretical ceiling of the variance in attention that could be explained. The scatter plots show the grand mean (black horizontal line), 95% confidence intervals (coloured boxes), and 1 standard deviation (black vertical line), for meaning and image salience across all scenes for each analysis. Each dot represents the correlation for a given scene.

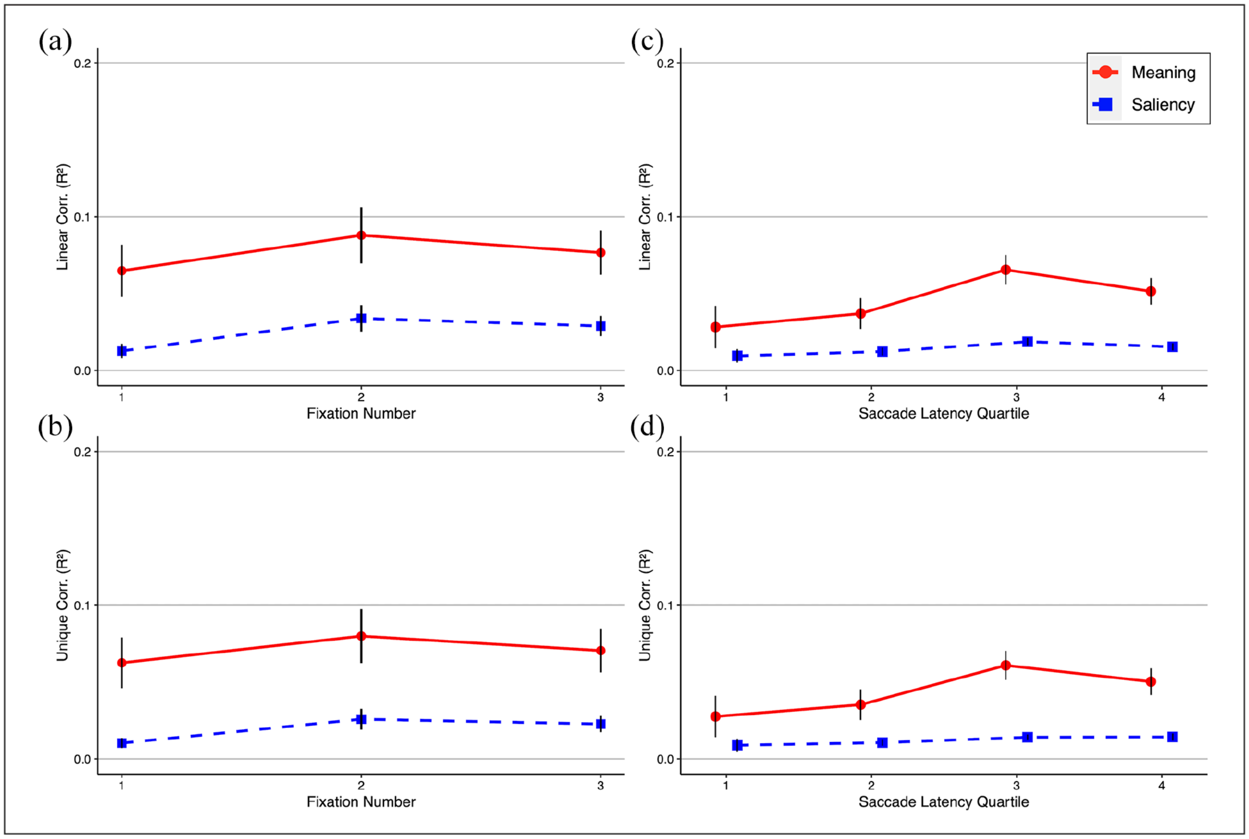

Figure 4.

Experiment 1: correlations of meaning and salience in the first three fixations. Line plots show the (a) squared linear and (b) semi-partial correlations between the fixation density maps, meaning (red circles), and image salience (blue squares) for the first three fixations across all scenes for each analysis. (c) The average meaning (red circles) and salience (blue squares) values for each quartile of initial saccade latency. Error bars refer to the standard error of the mean.

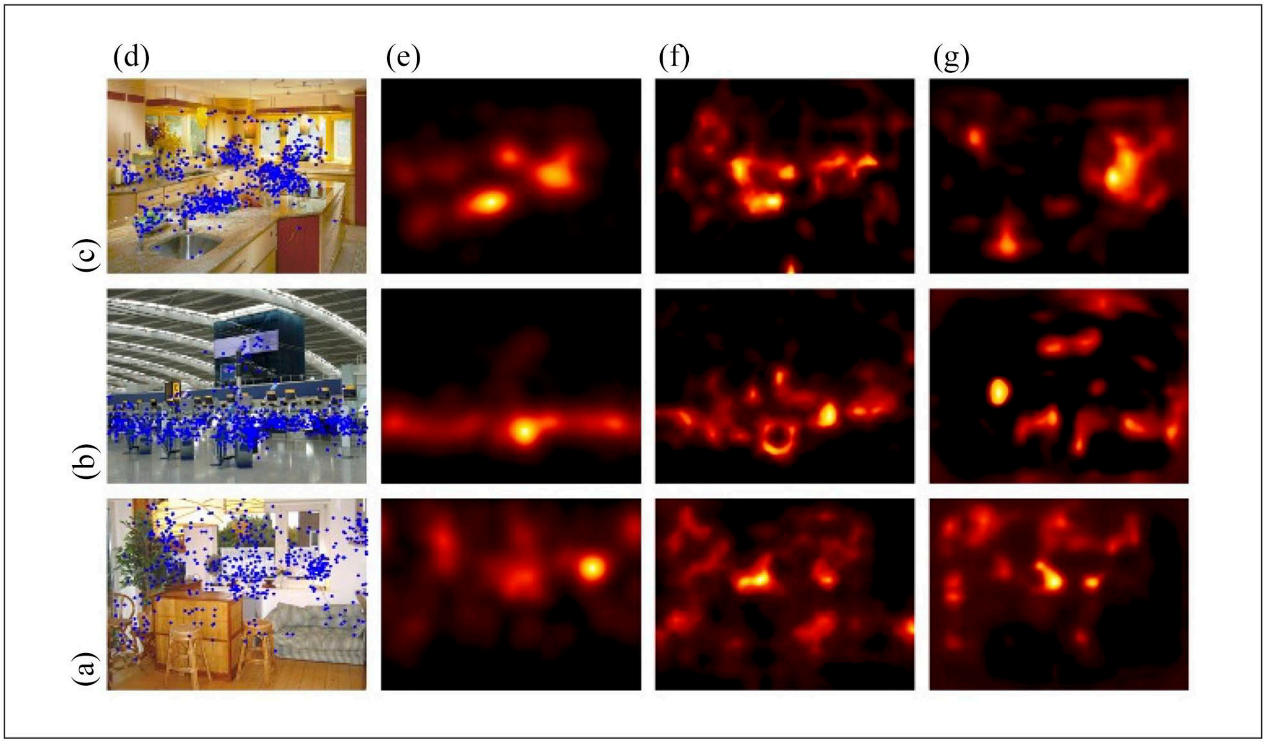

Figure 5.

Example scenes and maps: (a) example target absent scenes used in the study for (a) paintings, (b) garbage bins, and (c), (d) drinking glasses with fixations overlaid. (e) the corresponding fixation density map for the scene. (f) the meaning map and (g) the salience map for the example scene.

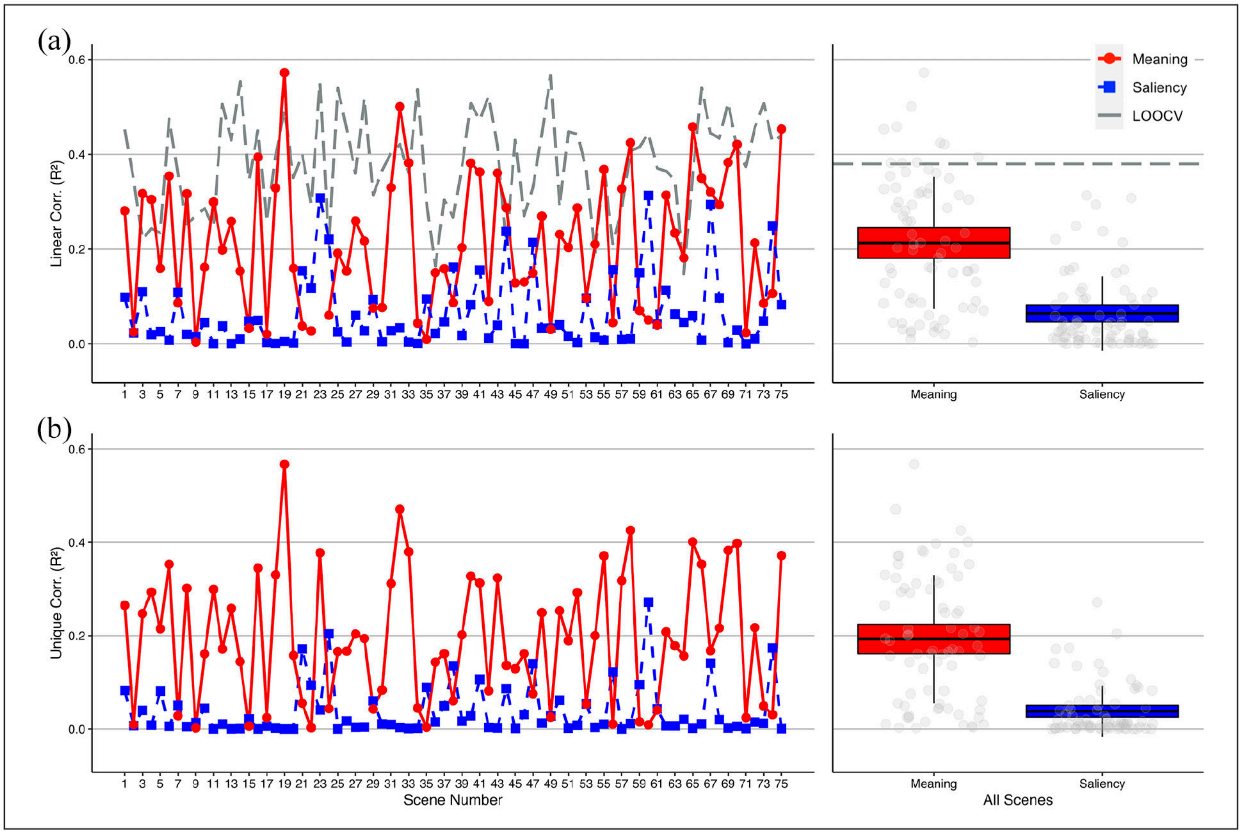

Figure 6.

Experiment 2: meaning versus salience. Line plots show the (a) squared linear and (b) semi-partial correlations between the fixation density maps, meaning (red circles), and image salience (blue squares). The grey, dashed lines in panel (a) correspond to the leave one out cross-validation (LOOCV) that predicted the theoretical ceiling of the variance in attention that could be explained. The scatter plots show the grand mean (black horizontal line), 95% confidence intervals (coloured boxes), and 1 standard deviation (black vertical line), for meaning and image salience across all scenes for each analysis. Each dot represents the correlation for a given scene. People searched for paintings in scenes 1–25. People searched for garbage bins in scenes 26–50, and searched for drinking glasses in scenes 51–75.

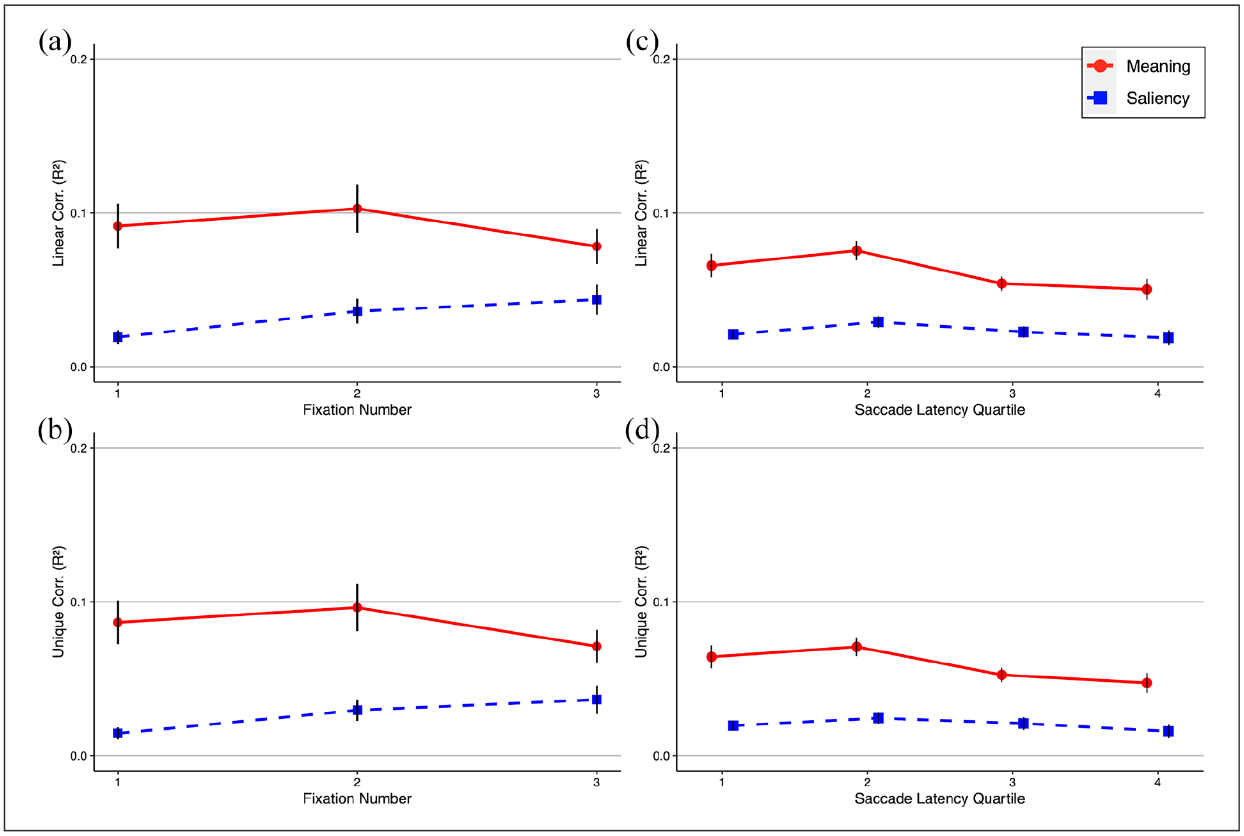

Figure 7.

Experiment 2 correlations of meaning and salience in the first three fixations. Line plots show the (a) squared linear and (b) semi-partial correlations between the fixation density maps, meaning (red circles), and image salience (blue squares) for the first three fixations across all scenes for each analysis, and (c) the average meaning (red circles) and salience (blue squares) values for each quartile of initial saccade latency. Error bars refer to the standard error of the mean.