# Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps

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

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

**We compared the influence of meaning and of salience on attentional guidance in scene images. Meaning was captured by "meaning maps" representing the spatial distribution of semantic information in scenes. Meaning maps were coded in a format that could be directly compared to maps of image salience generated from image features. We investigated the degree to which meaning versus image salience predicted human viewers' spatiotemporal distribution of attention over scenes. Extending previous work, here the distribution of attention was operationalized as duration-weighted fixation density. The results showed that both meaning and image salience predicted the duration-weighted distribution of attention, but that when the correlation between meaning and salience was statistically controlled, meaning accounted for unique variance in attention whereas salience did not. This pattern was observed in early as well as late fixations, fixations including and excluding the centers of the scenes, and fixations following short as well as long saccades. The results strongly suggest that meaning guides attention in real-world scenes. We discuss the results from the perspective of a cognitive-relevance theory of attentional guidance.**

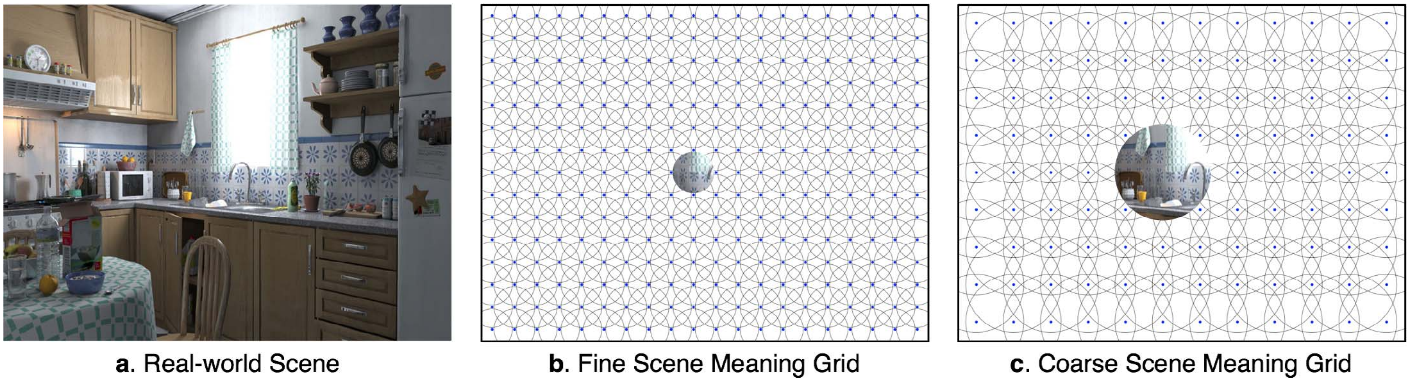
## Introduction

We can attend to only a fraction of the visual stimulation available to us at any given moment. For this reason, visual attention is guided through scenes in real time, with the eyes shifting position about three times each second, on average, to select informative objects and scene regions for scrutiny (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003; Henderson, 2017; Henderson & Hollingworth, 1999; Land & Hayhoe, 2001; Rayner, 2009; Yarbus, 1967). How does the brain determine which scene regions and elements should be attended at any given moment?

Most recent research on attentional guidance in real-world scene images has focused on the idea that attention is primarily driven by low-level image features. Image-guidance theory has its roots in models of attention and visual search that emphasize the attraction of attention by primitive visual features and feature differences (Treisman & Gelade, 1980; Wolfe & Horowitz, 2017). When applied to real-world scenes, the most influential instantiation of this type of theory is based on image salience, which proposes that saliency maps are generated by pooling contrasts in semantically uninterpreted image features such as luminance, color, and edge orientation at multiple spatial scales (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Harel, Koch, & Perona, 2006; Itti & Koch, 2001; Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002). In this theoretical approach, regions that are uniform along these features are considered uninformative, whereas those that differ from neighboring regions across spatial scales are taken to be worthy of attention. That is, differences in salience in the map serve as predictions about the spatial distribution of attention in a scene. In this view, attentional guidance is fundamentally characterized as a reaction to image features in the scene, with attention captured by or pulled to visually salient scene regions (Henderson, 2007). An appeal of image-guidance theory based on image salience is that salience is both neurobiologically inspired and computationally tractable (Henderson, 2017). The saliency-map approach has served an important heuristic function in the study of attention and eye movements in scene perception by providing an explicit model that generates quantitative predictions about attention and eye movements.

Citation: Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6):10, 1–18, https://doi.org/10.1167/18.6.10.

https://doi.org/10.1167/18.6.10 Received November 3, 2017; published June 21, 2018 ISSN 1534-7362 Copyright 2018 The Authors



**a**. Real-world Scene **b**. Fine Scene Meaning Grid **c**. Coarse Scene Meaning Grid

Figure 1. Real-world scene and corresponding tiled patch grids. (a) Example real-world scene. (b–c) Overlapping circular patches used for meaning rating at (b) fine and (c) coarse spatial scales. The blue dots in (b–c) denote the center of each circular patch and the image circles show examples of the content captured by the fine and coarse scales for the example scene.

Despite the substantial influence of the saliency-map approach on research in scene perception, it is well established that the semantic content of a scene and the viewer's task also influence viewing (Buswell, 1936; Yarbus, 1967). Indeed, when directly tested, image salience often does a poor job of accounting for attention in real-world scene viewing (Einhäuser, Rutishauser, & Koch, 2008; Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009; Tatler, Hayhoe, Land, & Ballard, 2011; Underwood, Foulsham, & Humphrey, 2009). To account for these observations, cognitive-guidance models place primary emphasis on cognitive control of attention. In this view, attention is pushed by the cognitive system to scene regions that are semantically informative and cognitively relevant in the current situation (Henderson, 2007). For example, in the cognitive-relevance model (Henderson et al., 2007; Henderson et al., 2009), attention is guided by semantic representations that code the meaning of the scene and its local regions (objects, surfaces, and other interpretable entities) with respect to the viewer's current goals and task (Buswell, 1935; Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Henderson, 2003; Henderson, 2007; Henderson, 2017; Henderson & Hollingworth, 1999; Land & Hayhoe, 2001; Rothkopf, Ballard, & Hayhoe, 2007; Tatler et al., 2011; Turano, Geruschat, & Baker, 2003; Võ & Wolfe, 2013; Yarbus, 1967). The cognitive-relevance model posits that the representations used to assign task relevance and meaning for attentional priority encode knowledge about the world itself (world knowledge) as well as knowledge about the general scene concept (scene schema knowledge) and the current scene instance (episodic scene knowledge; Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999).

Most proponents of image guidance acknowledge that meaning must play some role in attentional guidance. Nevertheless, much of the research on attentional guidance in real-world scene images has been motivated by and focused on image salience as instantiated by saliency maps. One reason for this emphasis is the relative tractability of image salience; it is far easier to quantify image features than it is to quantify meaning (Henderson, 2017). To investigate meaning and compare its influence to image salience, it is necessary to represent both constructs so that comparable quantitative predictions can be generated from them.

To provide a method for directly comparing the influences of meaning and salience on the guidance of attention, we recently developed the concept of *meaning maps* (Henderson & Hayes, 2017). Meaning maps draw inspiration from two classic scene-viewing studies (Antes, 1974; Mackworth & Morandi, 1967). In these studies, images were divided into regions and subjects were asked to rate each region based on how easy it would be to recognize (Antes, 1974) or how informative it was (Mackworth & Morandi, 1967). In both studies, eye movements of a different group of subjects were measured while they viewed the rated images. The key result was that, in general, viewers looked more at the higher-rated regions. We modified and extended these methods to develop meaning maps for images of real-world scenes. We used crowdsourced responses in which we asked subjects who were not aware of our experimental aims to rate the meaningfulness of a large number of scene patches. Specifically, photographs of scenes were divided into a dense array of objectively defined circular overlapping patches at two spatial scales (Figure 1). These patches were then presented to raters independently of the scenes from which they were taken, and raters were asked to indicate how informative or recognizable they judged the patches to be (Figure 2). Finally, we constructed smoothed maps for each scene based on interpolated ratings over a large number of raters (Figure 3). The basic idea of the meaning map is that it

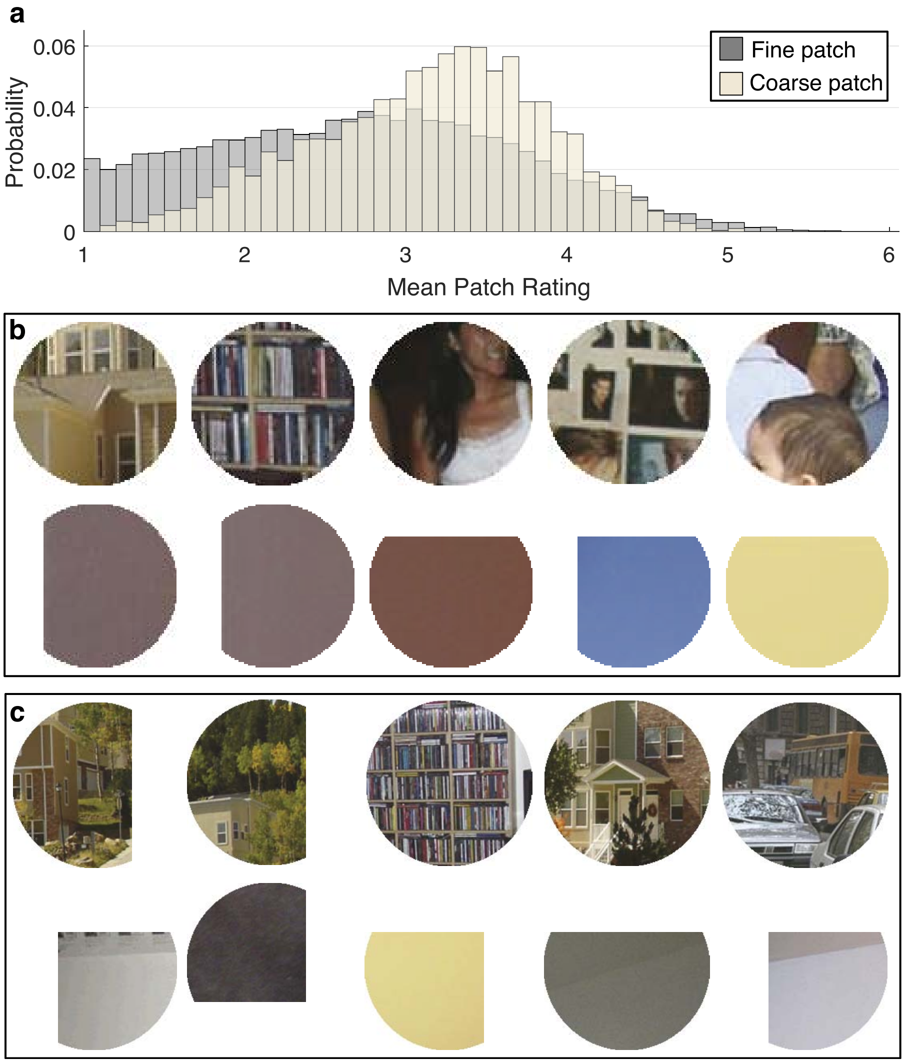


Figure 2. Rating distributions and example high and low patches. (a) Distribution of ratings for fine and coarse patches across all raters and scenes. (b–c) Example highest- and lowest-rated nonoverlapping patches for (b) fine and (c) coarse patches.

captures the spatial distribution of the semantic content of a scene in the same format as a saliency map captures the spatial distribution of image salience. Like image salience, meaning is nonuniformly spatially distributed across scenes, with some scene regions relatively rich in semantic content and others relatively sparse.

A meaning map provides the conceptual analog of a saliency map by capturing the spatial distribution of semantic features (rather than image features) across a

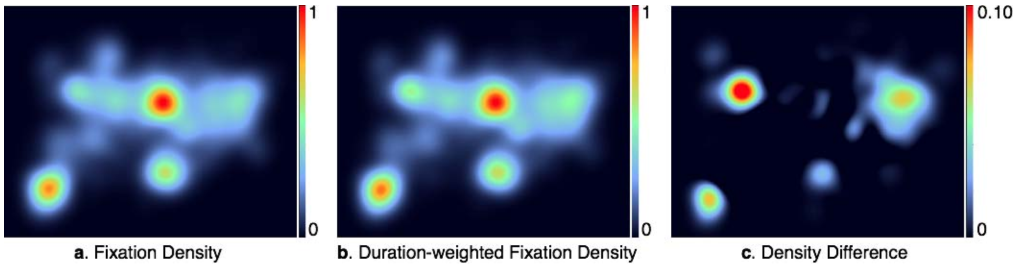


Figure 3. Duration-weighted fixation density. Example (a) fixation density and (b) duration-weighted fixation density, for all fixations on one scene. (c) The density difference depicting the absolute-value difference in the two densities, with hotter regions representing greater difference. Note that the meaning and saliency maps are on the same scale (0–1) and the difference map is on a 10% scale (0–0.10), to highlight the difference in the unweighted and weighted fixation-density maps.

scene. Because meaning maps are represented in the same format as saliency maps, they can be directly compared to saliency maps. A meaning map can be used to generate predictions concerning attentional guidance using the same methods that have been used to test the goodness of fit of predictions from saliency theory (Carmi & Itti, 2006; Itti, Koch, & Niebur, 1998; Parkhurst et al., 2002; Torralba, Oliva, Castelhano, & Henderson, 2006). And the predictions for attentional guidance generated from meaning maps can be compared to those generated from saliency maps. In short, meaning maps and saliency maps provide a foundation for directly contrasting the influences of meaning and salience on attentional guidance.

In an initial study, we investigated the relative ability of meaning maps and saliency maps to predict attentional guidance during scene viewing (Henderson & Hayes, 2017). In that study, and in keeping with the literature on scene perception, attention maps were based on the locations of eye fixations. We found that both meaning and salience could predict the distribution of attention over scenes, with meaning accounting for more variance in attention than salience. However, we also found that meaning and salience were themselves highly correlated. Furthermore, when the variance due to salience was controlled, meaning accounted for a significant amount of the remaining variance in attention; but when meaning was controlled, no further variance in attention was accounted for by salience. These data held for both early and later fixations during viewing, including the very earliest fixations on the scenes. The data strongly suggested that attention is guided by meaning rather than salience.

The present study was designed to extend the original meaning-map results. A potential concern with the original report is that the attention maps were based on fixation locations that did not take into account fixation durations (Henderson & Hayes, 2017). The fixation-location analysis was an important first step because most of the research assessing saliency maps to date has similarly focused on fixation location (Borji et al., 2013; Borji et al., 2014; Harel et al., 2006; Itti & Koch, 2001; Parkhurst et al., 2002). However, fixation durations vary, and this variability reflects a variety of factors including attention related to perceptual and cognitive processing. When more attention is needed on an object or other scene entity, fixations are directed to that entity for more time (Einhäuser & Nuthmann, 2016; Henderson, Nuthmann, & Luke, 2013; Henderson & Pierce, 2008; Henderson & Smith, 2009; Henderson, Weeks, & Hollingworth, 1999; Laubrock, Cajar, & Engbert, 2013; Nuthmann, 2017; Nuthmann, Smith, Engbert, & Henderson, 2010). The distribution of attention over a scene therefore depends on both the location and duration of attentional selection (Henderson, 2003). For this reason, we report here a new set of analyses designed to determine how well meaning and salience predict attentional guidance in scenes accounting for how long attention is focused on each location. We include a new center-knockout procedure to ensure that the results hold when center bias is completely removed from the analysis. Finally, we include a new saccade-amplitude analysis showing that the advantage for meaning over salience holds across all saccade amplitudes.

In summary, the goal of this study was to test current theoretical approaches to attentional guidance in real-world scenes. We applied our recently developed meaning-map method to capture the spatial distribution of semantic content across scenes. We then tested cognitive- and image-guidance theories by comparing the ability of meaning maps and saliency maps to predict attentional guidance during real-world scene viewing, with attention operationalized as the duration-weighted fixations of subjects viewing the scenes.

# Method

## Meaning maps

For this study we used the meaning maps developed by Henderson and Hayes (2017), as described in this section.

### *Subjects*

Scene-patch ratings were performed by 165 subjects on Amazon Mechanical Turk. Subjects 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. Subjects were paid $0.50 per assignment, and all subjects provided informed consent.

### *Stimuli*

Stimuli were 40 digitized (1,024 × 768 pixels) photographs of real-world scenes depicting a variety of indoor and outdoor environments. The full set of scene images can be found in the supplementary materials of Henderson and Hayes (2017). Each scene was decomposed into a series of partially overlapping (tiled) circular patches at two spatial scales (Figure 1). Simulated recovery of known scene properties (e.g., luminance) indicated that the underlying property could be recovered well (98% variance explained) using these two patch sizes (see Appendix), suggesting that this method is sufficiently sensitive to underlying scene structure. 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.

### Procedure

Each subject rated 300 random patches extracted from 40 scenes. Subjects were instructed to assess the meaningfulness of each patch based on how informative or recognizable it was. They were first given examples of two low-meaning and two high-meaning scene patches, to make sure they understood the rating task, and then they rated the meaningfulness of scene patches on a 6-point Likert scale (very low, low, somewhat low, somewhat high, high, very high). Patches were presented in random order and without scene context, so ratings were based on context-free judgments. Each unique patch was rated three times by three independent raters for a total of 48,960 ratings. However, due to the high degree of overlap across patches, each patch contained rating information from 27 independent raters for each fine patch and 63 independent raters for each coarse patch. Figure 2 shows the distribution of ratings and the highest- and lowest-rated nonoverlapping patches across all scenes at the two patch sizes. The lowest-rated patches tended to come from the edges of the pictures, which accounts for their truncated shapes.

Meaning maps were generated from the ratings by averaging, smoothing, and then combining fine and coarse maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The average rating maps were then smoothed using thin-plate spline interpolation (fit using the thinplateinterp method in MATLAB; MathWorks, Natick, MA). Finally, the smoothed maps were combined using a simple average. This procedure was used to create a meaning map for each scene. The final map was blurred by a multiplicative center-bias operation which down-weighted the scores in the periphery to account for the central fixation bias, the commonly observed phenomenon in which subjects concentrate their fixations more centrally and rarely fixate the outside border of a scene (Bindemann, 2010; Borji et al., 2013; Henderson et al., 2007; Tatler, 2007). This center bias operation is commonly applied to saliency maps and emerges in the ones used here.

To investigate the relationship between the generated meaning maps and image-based saliency maps, saliency maps for each scene were computed using the Graph-based Visual Saliency (GBVS) toolbox with default settings (Harel et al., 2006). GBVS is a prominent saliency model that combines maps of neurobiologically inspired low-level image features. The GBVS model also includes a center bias as described for the meaning maps that down-weights the periphery of its maps.

### Histogram matching

Meaning and saliency maps were normalized to a common scale using image-histogram matching, with the duration-weighted fixation map for each scene serving as the reference image for the corresponding meaning and saliency maps. Histogram matching of the meaning and saliency maps was accomplished using the MATLAB function imhistmatch in the Image Processing Toolbox.

## Eye-tracking experiment and attention maps

### Subjects

Seventy-nine University of South Carolina undergraduate students with normal or corrected-to-normal vision participated in the experiment. All were unaware of the purposes of the experiment and provided informed consent. The eye-movement data from each subject were inspected for excessive artifacts caused by blinks or loss of calibration due to incidental movement by examining the mean percentage of signal across all trials using MATLAB. Data from 14 subjects with less than 75% signal were removed, leaving 65 subjects for analysis who tracked very well (mean signal percentage = 91.74%). We have previously used this corpus to investigate individual differences in scan patterns in scene perception (Hayes & Henderson, 2017), as well as for an initial study of meaning maps (Henderson & Hayes, 2017).

### Apparatus

Eye movements were recorded with an EyeLink 1000+ tower-mount eye tracker (spatial resolution 0.01) sampling at 1,000 Hz (SR Research, 2010b). Subjects sat 85 cm away from a 21-in. monitor, so that scenes subtended approximately 27° × 20.4° of visual angle at 1,024 × 768 pixels. Head movements were minimized using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software (SR Research, 2010a).

### Stimuli

Stimuli consisted of the 40 digitized photographs of real-world scenes that were used to create the meaning and saliency maps.

### *Procedure*

Subjects were instructed to view each scene in preparation for a later memory test, which was not administered. Each trial began with fixation on a cross at the center of the display for 300 ms. Following central fixation, each scene was presented for 12 s while eye movements were recorded. Scenes were presented in the same order for all subjects.

A 13-point calibration procedure was performed at the start of each session to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error of less than 0.99°. Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/$s^2$; SR Research, 2010b).

Eye-movement data were imported off-line into MATLAB using the EDFConverter tool. The first fixation, always located at the center of the display as a result of the pretrial fixation period, was eliminated from analysis.

### *Attention maps*

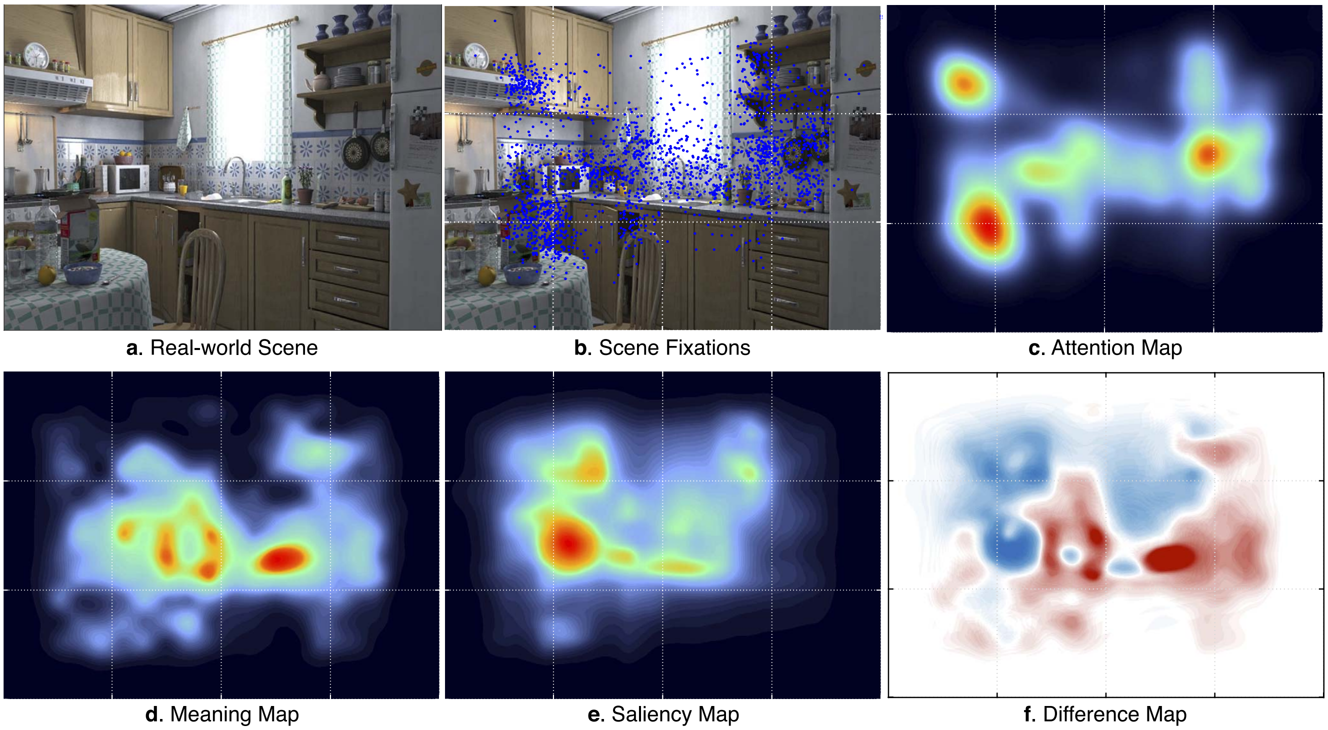
The distribution of attention over a scene is a function of the locations and durations of eye fixations (Henderson, 2003). Although maps created from fixation locations alone (Henderson & Hayes, 2017) and from the duration-weighted fixations were similar, they were not identical (see also Henderson, 2003). An example of the difference can be seen in Figure 3 by comparing fixation-density maps based on location alone (Figure 3a) to maps of location weighted by duration (Figure 3b). The difference in the two maps is shown in Figure 3c, with regions of greater difference shown with hotter colors. Note that the scale of the difference map is smaller than that of the original density maps. As can be seen, some regions changed their relative attentional weighting when duration was considered. For the present analyses, we therefore created attention maps from fixation density weighted by fixation duration.

To create duration-weighted attention maps, a duration weight was generated for every fixation following the initial (experimenter-defined) fixation. Because average fixation durations vary reliably and systematically across subjects (Castelhano & Henderson, 2008a; Henderson & Luke, 2014; Rayner, Li, Williams, Cave, & Well, 2007), duration weights were based on subject-normalized values. We first generated each subject's fixation-duration distribution across all 40 scenes. We then defined two parameters for these distributions, an upper-bound 95th-percentile cutoff (any values in the 95th percentile received a weight value of 1.0) and a lower-bound minimum weight cutoff of 0.1 (any value below the 0.1 percentile received a weight value of 0.1, to avoid weights of 0). Each fixation was therefore weighted from 0.1 to 1.0 based on its place in the overall distribution. Fixation-weighted values were accumulated across all subjects adding the weight to each location, producing a weighted fixation-frequency matrix for each scene. Finally, a Gaussian low-pass filter with a circular boundary and a cutoff frequency of −6 dB was applied to the matrix for each scene, to account for foveal acuity and eye-tracker error. The Gaussian low-pass function is from the MIT Saliency Benchmark code (https://github.com/cvzoya/saliency/blob/master/code_forMetrics/antonioGaussian.m). With a cutoff frequency fc = 6, the window size is approximately 2° of visual angle. An example of a resulting duration-weighted attention map is shown in Figure 3b.

## Results

We can take meaning maps and saliency maps as predictions concerning how viewers will distribute their attention over scenes. To investigate how well meaning maps and saliency maps predict the distribution of attention, it is important to assess the degree of association between the maps themselves. For the scenes used here, the correlation between meaning and salience was 0.80 averaged across the 40 scenes (Henderson & Hayes, 2017). This correlation is consistent with the suggestion that attention effects that have previously been attributed to salience could be due to meaning (Henderson et al., 2007; Henderson et al., 2009; Nuthmann & Henderson, 2010). At the same time, meaning and salience did not share 36% of their variance, and we can ask how well this unshared variance in each predicts attention.

The critical empirical question was how well the two types of prediction maps capture the distribution of attention. To investigate this question, we used linear correlation (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016) to determine the degree to which meaning maps (Figure 4d) and saliency maps (Figure 4e) for a scene (Figure 4a) statistically predicted the spatial distribution of attention (Figure 4b), as represented by the duration-weighted attention maps (Figure 4c). This method allows us to assess the degree to which meaning maps and saliency maps account for shared and unique variance in the attention maps. There are many ways in which prediction maps can be tested against attention maps, and no method is perfect (Bylinskii et al., 2016). We chose here a map-level analysis method that is sensitive, makes relatively few assumptions, is intuitive, can be visualized, generally balances the various positives and negatives of different analysis approach-


**a**. Real-world Scene

**b**. Scene Fixations

**c**. Attention Map

**d**. Meaning Map

**e**. Saliency Map

**f**. Difference Map

Figure 4. Example data used in the analyses. (a) Real-world scene. (b) Viewers' fixations superimposed on the scene as blue dots. (c) Duration-weighted attention map derived from the fixations. (d) Meaning map. (e) Saliency map for the example scene. (f) Difference between the meaning and saliency maps, with regions of greater meaning shown in red and greater saliency shown in blue. The meaning and saliency maps are on the same scale (0–1), and the difference map is on a 10% scale (0–0.10). Note that the guidelines in the scene were not shown to subjects and are presented here to facilitate comparison across panels.

es, and allows us to tease apart variance due to salience and meaning.

Figure 5 presents the primary data for each of the 40 scenes. Each data point shows the relationship ($R^2$ value) between the meaning map and the observed attention map for each scene (red), and between the saliency map and the observed attention map for each scene (blue). The top half of Figure 5 shows the squared linear correlations. On average, across the 40 scenes, meaning accounted for 50% of the variance in fixation

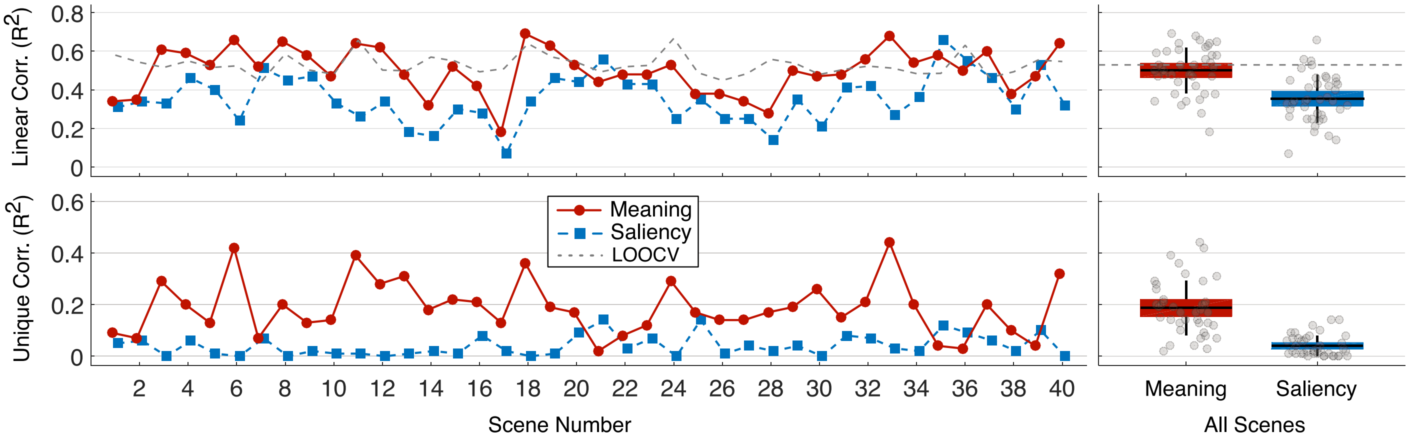


Figure 5. Squared linear correlation and semipartial correlation by scene and across all scenes. The line plots show the linear correlation (top) and semipartial correlation (bottom) between duration-weighted fixation density and meaning and salience by scene. The scatter box plots on the right show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and one standard deviation (black vertical line) for meaning and salience across all 40 scenes. The dotted lines in the top panels show consistency across subjects based on leave-one-out cross-validation.

density ($M = 0.50$, $SD = 0.12$) and salience accounted for 35% ($M = 0.35$, $SD = 0.12$). A two-tailed $t$ test revealed that this difference was statistically significant, $t(78) = 5.38$, $p < 0.0001$, 95% confidence interval (CI) [0.09, 0.20].

We used leave-one-out cross-validation to estimate the upper limit on salience and meaning performance given subject variability in attention (Torralba et al., 2006). The cross-validation analysis sets an expected maximum in the ability of meaning and salience to account for attention. Specifically, we computed for each scene a duration-weighted group attention map as described earlier for 64 subjects and a test map for the 65th subject. The linear correlation of the group and test maps was computed, and this was repeated for all 65 subjects. Mean correlations by scene and across scenes were then generated. The results are shown as dotted lines in the top panels of Figure 5, 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, cross-validation $R^2$ was 0.53 ($SD = 0.05$). Meaning-map performance was not statistically different from this theoretical maximum, as revealed by a two-tailed $t$ test, $t(78) = 1.37$, $p = 0.17$, 95% CI [−0.01, 0.07]. In comparison, saliency maps produced poorer performance than the theoretical maximum, $t(78) = 8.17$, $p < 0.0001$, 95% CI [0.13, 0.22]. These results show that meaning maps accounted for attention about as well as possible, given the reliability of the subject data, whereas saliency maps performed significantly below this level.

To examine the unique variance in attention explained by meaning and salience when controlling for their shared variance, we computed squared semipartial correlations (bottom half of Figure 5). Across the 40 scenes, meaning accounted for a significant 19% additional variance in the attention maps after controlling for salience ($M = 0.19$, $SD = 0.11$), whereas salience accounted for a nonsignificant 4% additional variance after controlling for meaning ($M = 0.04$, $SD = 0.04$). A two-tailed $t$ test confirmed that this difference was statistically significant, $t(78) = 8.22$, $p < 0.0001$, 95% CI [0.11, 0.18]. These results show that meaning explained the distribution of attention over scenes better than salience.

It has sometimes been proposed that during scene viewing, attention is initially guided by image salience, but as viewing progresses over time, meaning begins to play a greater role (Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999; Mannan, Ruddock, & Wooding, 1996; Parkhurst et al., 2002). To test this proposal, we conducted temporal time-step analyses. Linear correlation and semipartial correlations were conducted based on a series of attention maps, with each map generated from each sequential eye fixation (first, second, third fixation, etc.) in each scene. This method allowed us to test whether the relative importance of meaning and salience in predicting attention changed over time. The results are shown in Figure 6. For the linear correlations, the relationship was stronger between the meaning and attention maps for all time steps (top of Figure 6) and was highly consistent across the 40 scenes. Meaning accounted for 33.0%, 32.1%, and 29.7% of the variance in the first three fixations, whereas salience accounted for only 9.5%, 15.2%, and 16.6% of the variance in the first three fixations. Two-sample two-tailed $t$ tests were performed for all 38 time points, and $p$ values were corrected for multiple comparisons using the false-discovery-rate (FDR) correction (Benjamini & Hochberg, 1995). This procedure confirmed the advantage for meaning over salience at all 38 time points (FDR $< 0.05$).

The improvement in $R^2$ for the meaning maps over saliency maps observed in the overall analyses was again found to hold across all 38 time steps in the partial correlations (bottom of Figure 6; FDR $< 0.05$), with meaning accounting for 26.1%, 21.7%, and 17.4% of the unique variance in the first three fixations, whereas salience accounted for 2.7%, 4.6%, and 4.2%. Counter to the salience-first hypothesis but consistent with results based on unweighted fixations (Henderson & Hayes, 2017), meaning accounted for more variance in attention in both the correlation and semipartial-correlation analyses than did salience from the very first fixation. These results indicate that meaning begins guiding attention as soon as a scene appears (Rider, Coutrot, Pellicano, Dakin, & Mareschal, 2018).

## Central-region knockout analyses

It is commonly found in eye-tracking studies that viewers tend to concentrate their fixations near the center and rarely fixate the outside borders of a real-world scene (Borji et al., 2013; Henderson et al., 2007; Tatler, 2007). As noted under Method, in creating the final meaning maps we used a multiplicative center-bias operation to down-weight the scores in the periphery and consequently up-weight the center, as is commonly done with saliency maps. However, to further ensure that the advantage of meaning maps over saliency maps in predicting the distribution of attention was not due to a center-bias advantage for the meaning maps, we also conducted additional analyses in which the data from the central 7° of each map (attention, meaning, and saliency) were removed. Differences in the success of meaning and saliency maps in this analysis therefore cannot be due to differences in the ability of meaning maps to predict central fixations, since they are no longer included. The results of these analyses were qualitatively and quantitatively very similar to those of the complete analyses.

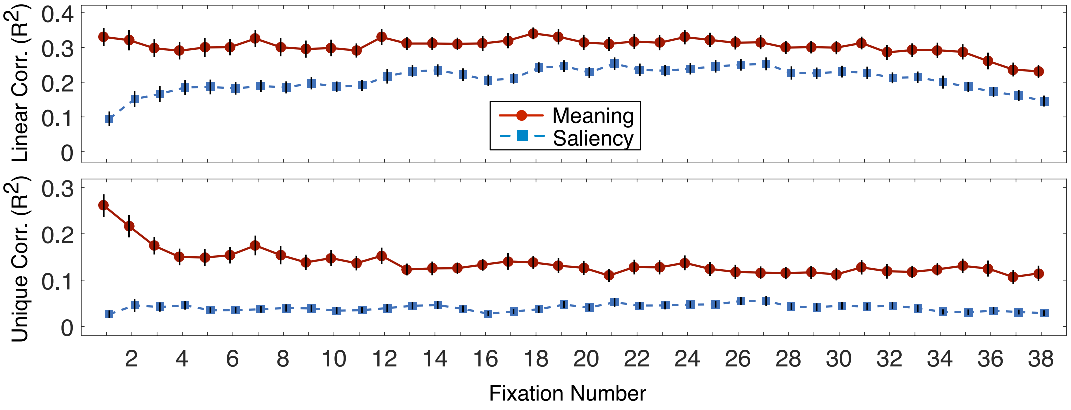


Figure 6. Squared linear correlation and squared semipartial correlation as a function of fixation number. The top panel shows the squared linear correlation between duration-weighted fixation density and meaning and salience as a function of fixation order across all 40 scenes. The bottom panel shows the corresponding semipartial correlation as a function of fixation order across all 40 scenes. Error bars represent standard error of the mean.

Figure 7 presents the correlation data used to assess the degree to which meaning maps and saliency maps accounted for shared and unique variance in the attention maps for each scene excluding the central 7°. Each data point shows the $R^2$ value for the prediction maps (meaning and saliency) and the observed attention maps for saliency (blue) and meaning (red). The top of Figure 7 shows the squared linear correlations. On average, across the 40 scenes excluding scene centers, meaning accounted for 46% of the variance in fixation density ($M = 0.46$, $SD = 0.11$) and saliency accounted for 34% ($M = 0.34$, $SD = 0.13$). A two-tailed $t$ test revealed that this difference was statistically significant, $t(78) = 4.39$, $p < 0.0001$, 95% CI [0.06, 0.17].

To examine the unique variance in attention explained by meaning and salience excluding the central 7° and when controlling for their shared variance, we computed squared semipartial correlations. These correlations, shown in the bottom of Figure 7, revealed that across the 40 scenes, meaning captured more than three times as much unique variance ($M = 0.17$, $SD = 0.10$) as saliency ($M = 0.05$, $SD = 0.05$). A two-tailed $t$ test confirmed that this difference was statistically significant, $t(78) = 6.78$, $p < 0.0001$, 95% CI [0.08, 0.16]. These results confirm those of the complete analysis and indicate that meaning was better able than salience to explain the distribution of attention over scenes even when the central 7° of maps was removed.

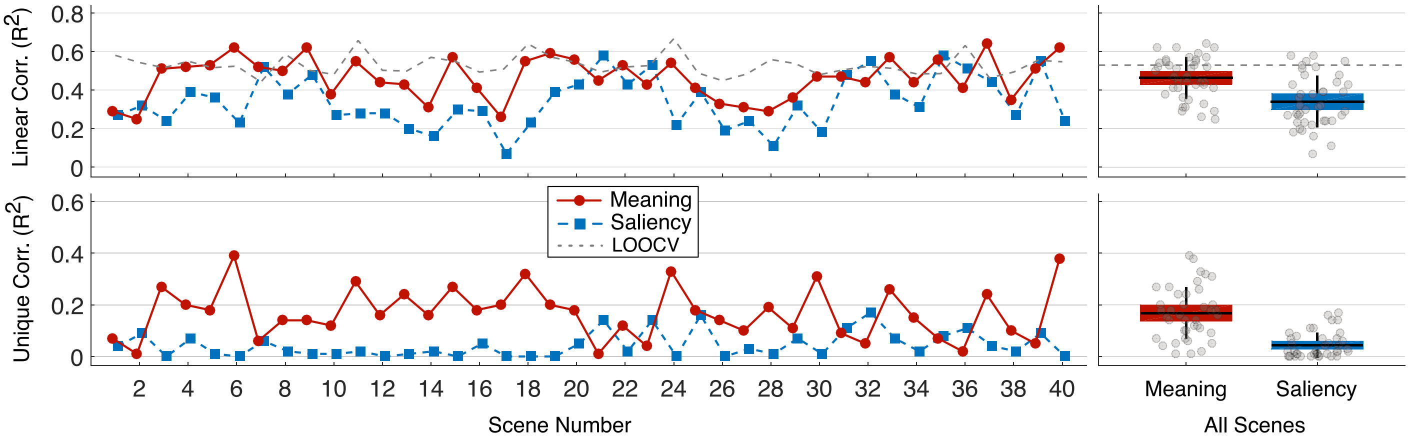


Figure 7. Squared linear correlation and semipartial correlation by scene and across all scenes with 7° center removed. The line plots show the linear correlation (top) and semipartial correlation (bottom) between duration-weighted fixation density and meaning and salience by scene after removing the central 7° from each scene. The scatter box plots on the right show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and one standard deviation (black vertical line) for meaning and salience across all 40 scenes. The dotted lines in the top panels show consistency across subjects based on leave-one-out cross-validation.

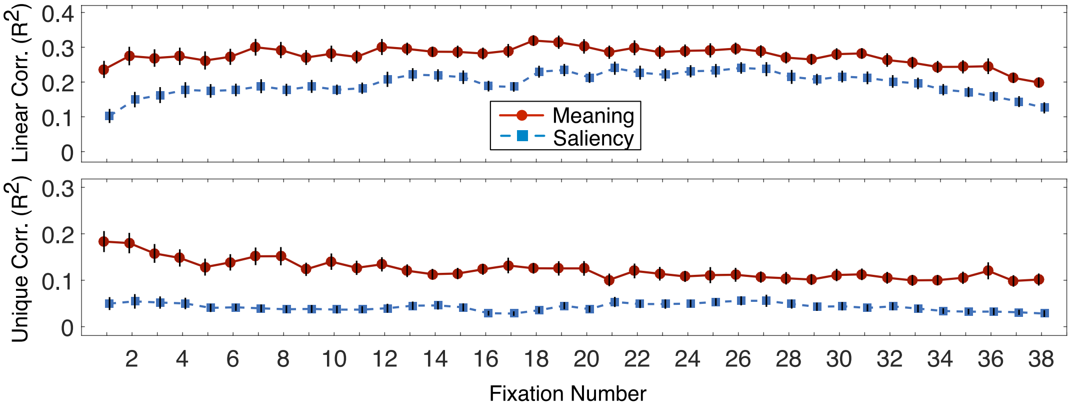


Figure 8. Squared linear correlation and squared semipartial correlation as a function of fixation number with 7° center removed. The top panel shows the squared linear correlation between fixation density and meaning (red) and salience (blue) as a function of fixation order averaged over all 40 scenes. The bottom panel shows the corresponding semipartial correlation as a function of fixation order averaged over all 40 scenes. Error bars represent standard error of the mean.

To test whether the overall advantage of meaning over salience early in viewing was due to meaning at the center, we conducted the fixation-series analysis excluding the central 7° of maps. Figure 8 shows the temporal time-step analyses with the central 7° of maps removed. Linear correlation and semipartial correlation were conducted as in the main time-step analyses, based on a series of attention maps generated from each sequential eye fixation in each scene. Under the same testing and FDR correction as in the main analyses, 34 of 38 time points were significantly different in both the linear and semipartial analyses (FDR $< 0.05$), excluding fixations 21, 25, 27, and 28. Importantly for assessing initial control of attention during scene viewing, meaning accounted for 22.9%, 27.0%, and 26.7% of the variance in the first three fixations in the linear-correlation analysis (top of Figure 8), whereas salience accounted for only 10.2%, 14.9%, and 16.2%. Critically, when the correlation among the two prediction maps was controlled for with semipartial correlations, the advantage for the meaning maps observed in the overall analyses was also found to hold across all time steps, as shown in the bottom of Figure 8 (FDR $< 0.05$). Meaning accounted for 17.9%, 17.8%, and 15.7% of the unique variance in the first three fixations, whereas salience accounted for 5.2%, 5.6%, and 5.4%. Consistent with the overall correlation and semipartial-correlation analyses, meaning produced an advantage over salience from the very first fixation even when the central 7° region of each map was removed from analysis. These results indicate that when overt attention leaves the center of a scene, meaning guides even those earliest shifts of overt attention. These results are especially strong evidence for the control of attention by meaning, because removing the central 7° should disadvantage the meaning maps, given that photographers tend to center meaningful information (Tatler, 2007). Nevertheless, the meaning maps continued to outperform the saliency maps in accounting for both overall variance and unique variance in the attention maps.

## Saccade-amplitude analyses

It could be that meaning controls attention as it is guided within objects and nearby scene regions, but that salience controls attention as it is guided from one scene region to another. If this is true, then meaning should be more highly related to attentional selection following shorter saccades, whereas image salience should be more highly related to attention following longer saccades. To investigate this prediction, we conducted an analysis in which we examined how meaning and salience related to attention following saccades of different amplitudes.

Figure 9 presents the distribution of saccade amplitudes in the present study. The average amplitude was 3.5°, but as is typically observed in scene viewing, saccade amplitude varied considerably (Henderson & Hollingworth, 1999). Once again, we used correlation analyses to assess the degree to which meaning maps and saliency maps accounted for shared and unique variance in the attention maps for fixations following saccades of different amplitudes. For these analyses, saccade amplitudes were binned by percentile. Each data point shows the $R^2$ value for the observed attention maps for saliency (blue) and meaning (red) at each saccade-amplitude ventile. The middle of Figure 9 shows the squared linear correlations, and the bottom of Figure 9 shows the unique variance accounted for by meaning and salience. The $R^2$ values for meaning and salience differed for all amplitudes except the very longest ventile in both figures (FDR $< 0.05$). These results confirm those of the complete analysis and indicate that meaning was better able than salience to

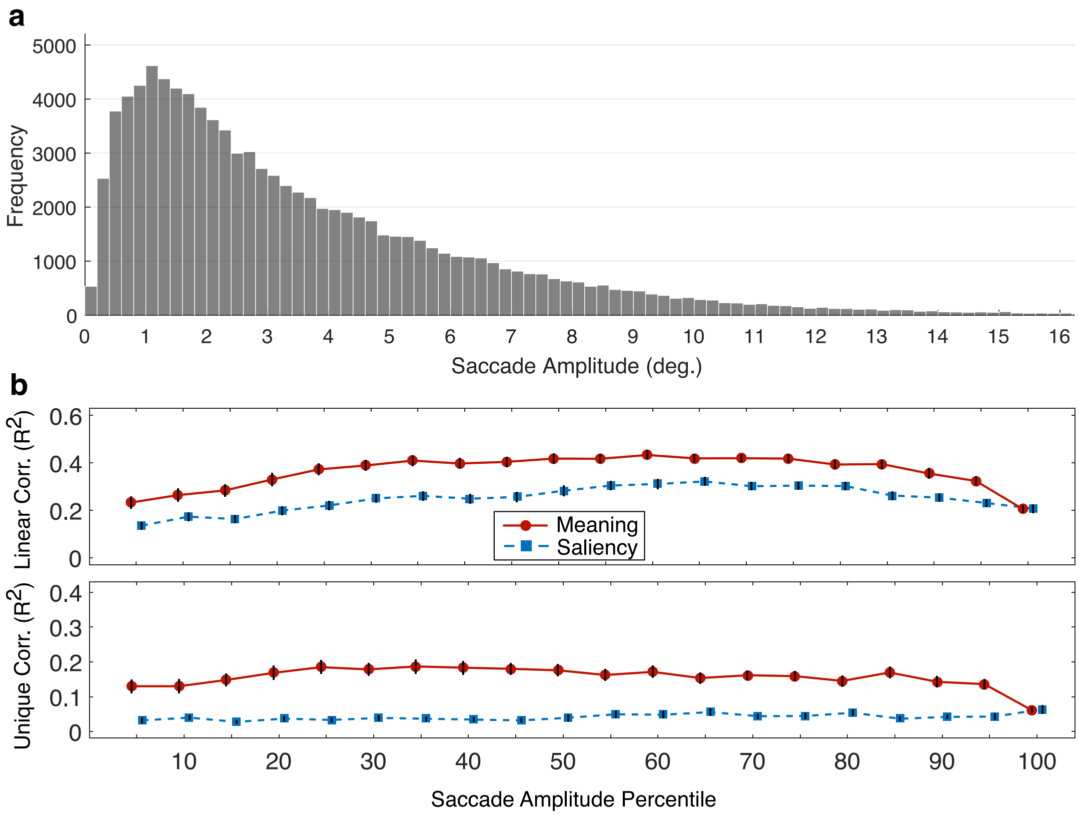


Figure 9. Squared linear correlation and squared semipartial correlation as a function of saccade amplitude to fixation. (a) The distribution of saccade amplitudes observed in the experiment. (b) The squared linear correlations between duration-weighted fixation density for meaning and salience as a function of the saccade-amplitude percentiles prior to fixation. (c) The corresponding semipartial correlations as a function of saccade amplitude. Data points are averages across all 40 scenes. Error bars represent standard error of the mean.

explain the distribution of attention over scenes even when attention was not limited to the object or scene region at the current point of attention.

## General discussion

Image salience as instantiated by computationally derived saliency maps currently provides a central theoretical framework and empirical paradigm for understanding how attention is guided through real-world scenes. Yet human viewers are known to be highly sensitive to the semantic content of the visual world that they perceive, suggesting that attention may be directed by semantic content rather than image salience. Until recently it has been difficult to directly contrast the influence of image salience and meaning. To address this difficulty, we developed a new method for identifying and representing the spatial distribution of meaning in any scene (Henderson & Hayes, 2017). The resulting meaning maps quantify the spatial distribution of semantic content across scenes in the same format that saliency maps quantify the spatial distribution of image salience. Meaning maps therefore provide a method for disentangling the distribution of meaning from the distribution of image salience. In the present study, we used meaning maps to test the relative importance of meaning and salience during scene viewing by testing meaning maps and saliency maps against observed duration-weighted attention maps.

The results showed that both meaning maps and saliency maps were able to account for considerable variance in attention maps, suggesting that they both offered good predictions concerning attention. However, meaning maps and saliency maps are themselves strongly correlated (Henderson & Hayes, 2017). When these correlations were statistically controlled, meaning maps accounted for additional unique variance in the duration-weighted distribution of attention over scenes. On the other hand, the variance due to visual salience was completely accounted for by meaning, such that saliency maps accounted for no additional unique variance in the attention maps when the variance accounted for by meaning was controlled. These results suggest that meaning plays the primary role in directing attention through scenes.

A similar advantage of meaning over salience was observed throughout the viewing period, with unique variance accounted for by meaning beginning with the first subject-determined fixation. Contrary to salience-first models, these results suggest that meaning influences attentional guidance more strongly than salience both early and later during scene viewing. The results indicate that meaning begins guiding attention as soon as a scene appears, and suggest that viewers are able to determine very quickly (within the first glimpse) where meaningful regions within the current scene are to be found and to direct their attention based on that assessment.

The strong role of meaning in guiding attention in scenes can be accommodated by a theoretical perspective that places explanatory primacy on scene semantics. For example, in the *cognitive-relevance* model (Henderson et al., 2007; Henderson et al., 2009), the priority of an object or scene region for attention is determined solely by its meaning in the context of the scene and the current goals of the viewer, not by image features or salience. It is meaning that determines attentional priority, with image properties used only to generate perceptual objects and other perceptually based potential saccade targets. Critically, then, attentional priority is assigned to potential attentional targets based not on image salience but rather on knowledge representations. The visual stimulus is relevant in that it is used to generate perceptual objects and other targets for attention, and processes related to salience may be relevant in determining whether a perceptual object is generated, but the image features themselves provide a flat (that is, unranked) landscape of potential attentional targets rather than one ranked by salience (Henderson et al., 2007). Instead, knowledge representations provide the attentional-priority ranking to the targets based on their meaning (Henderson, 2003; Henderson et al., 2007; Henderson et al., 2009).

It is important to note that the cognitive-relevance model does not require that meaning be assigned simultaneously across the entire scene to all perceptually mapped potential saccade targets. That is, the model does not require a strong late-selection view of scene perception in which all objects and scene regions are fully identified before they are attended. There are two reasons for this. First, when a scene is initially encountered, the gist of the scene can be quickly apprehended (Biederman, 1972; Castelhano & Henderson, 2008b; Fei-Fei, Iyer, Koch, & Perona, 2007; Potter, 1975) and can guide attention at the very earliest points of scene viewing (Castelhano & Henderson, 2003; Henderson & Hollingworth, 1999; Oliva & Torralba, 2006; Võ & Henderson, 2010). Apprehending the gist allows access to schema representations that provide constraints on what objects are likely to be present and where they are likely to be located (Henderson, 2003; Henderson & Hollingworth, 1999; Torralba et al., 2006). Information retrieved from memory schemas can be combined with low-quality visual information from the periphery to assign tentative meaning to perceptual objects and other scene regions. These initial representations provide a rich set of priors that can be used to generate predictions for guiding attention to regions that have not yet been identified (Henderson, 2017). Second, most saccades during scene viewing are relatively short, with an average amplitude of about 3.5° in the present study. The implication is that attention is frequently guided from the current location to the next location based on information that is relatively close to the fovea, where identity and meaning can easily be ascertained. Extraction of meaning from nearby regions cannot be the entire story for attentional guidance, given that meaning continues to dominate salience even for fixations following longer saccades, as shown in the present study, but it does suggest that for the many shorter shifts of attention, meaning is at least partly derived from a spatially local semantic analysis of the scene. For longer saccades, it is likely that guidance is based on scene representations retrieved from memory, as already described.

The present results at first glance appear to be at odds with past studies that have shown correlations between visual salience and attention. How can we account for the results of these earlier studies? One explanation can be found in the strong correlation between meaning and visual salience. We have hypothesized in the past that this correlation is likely to be high (Henderson et al., 2007). Meaning maps provide a method for testing this hypothesis, and robust support was found for it, with strong correlation between meaning and salience (Henderson & Hayes, 2017). Given this correlation, salience can do a reasonably good job of predicting meaning-driven attention. From an engineering perspective, this might be sufficient. However, from the perspective of the neurocognitive study of human vision, in which the goal is to provide a theoretical account of how the brain guides attention, a focus on image salience will be misleading. Instead, the present results along with previous results (Henderson & Hayes, 2017) strongly suggest that meaning, not visual salience, is the causal factor that guides attention.

## Limitations and future directions

We note several limitations and caveats of this study and our earlier meaning-map investigation (Henderson & Hayes, 2017). First, we have so far used a single viewing task. It has been shown that attention as indexed by eye movements differs over the same scene depending on the task (Castelhano, Mack, & Henderson, 2009; Henderson et al., 1999; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011; Yarbus, 1967), and it could be that under other task instructions, image salience would play a greater role than meaning. While this is a possibility, the memorization task used here is a relatively unstructured free-viewing task in which viewers are not explicitly or implicitly directed to meaningful scene regions. Therefore, this task would not seem to favor meaning-based over image-based attentional guidance. Nevertheless, we cannot rule out the possibility that salience might play a more important role in other tasks, and it will be important to assess the relative influence of meaning and salience in guiding attention in different viewing tasks.

Second, although meaning was the stronger predictor of attention on average and for the majority of scenes (36 out of 40) tested here, salience did perform better for four scenes. The question arises why these four scenes showed the opposite pattern. One possibility is that there may simply be statistical noise in one or more of the maps (meaning, saliency, or attention) for a given scene that occasionally leads to a random reversal of the true pattern. Another possibility is that there is some systematic difference in the scenes that show the reversed pattern. We were not able to discern any particular regularities across those scenes, but a future direction for study will be to compare different classes of scenes (e.g., indoor vs. outdoor; natural vs. artificial) to determine whether meaning and salience play greater or lesser roles for specific types of scenes.

Third, in the present study we defined meaning in a context-free manner, in the sense that each scene patch was rated for meaning without regard to the scene it came from. Meaning could instead be defined in a context-dependent manner, with the meaning of a scene region assessed in terms of its scene context. Similarly, meaning could vary as a function of the viewer's task. So far we have focused on context-free meaning as a first step, but it will be important to determine how meaning changes as the context changes, and in turn how context-dependent meaning influences attention. One way to determine context-dependent meaning is to ask participants to indicate directly (e.g., via mouse click) which regions in a scene they find most interesting (Onat, Açık, Schumann, & König, 2014). However, in this type of task subjects might click on regions that their attention has been drawn to, potentially confounding visual salience and meaning and leading to some circularity in using these clicks to predict future attention. Alternatively, consistent with the present approach, we might ask subjects to rate independent experimenter-defined scene patches but within the context of the entire scene or a specific task.

Fourth, we have chosen here to compare meaning to the class of saliency models that are inspired and motivated by neurobiologically plausible assumptions about the nature of visual computation in the human visual system (Borji & Itti, 2013; Itti et al., 1998; Itti & Koch, 2001). This class of saliency model continues to inspire a vast amount of research across many disciplines. Within this class of model, we have used the GBVS implementation because it is typically the best performer (Walther & Koch, 2006). Indeed, in our own comparisons of saliency models, GBVS outperforms other similar models on our data set. However, it should be noted that another class of model based on learning within deep neural networks has recently been advanced as a competitor to traditional saliency models (Vig, Dorr, & Cox, 2014). For example, DeepGaze II, the current top performer in this class, learns where people attend in scenes from training sets of fixations over object features and then predicts fixations on new scenes (Kümmerer, Wallis, Gatys, & Bethge, 2017). Interesting issues for future research include comparison of predictions from current deep neural networks and meaning maps, and extending deep neural networks to include meaning. However, an important consideration from the perspective of understanding human neurocognitive processes is whether these models trade neurobiological plausibility and transparency for engineering expediency.

## Conclusion

In this study we employed recently developed methods for comparing the relationship between the spatial distribution of meaning and image salience to the spatial distribution of attention in scene viewing (Henderson & Hayes, 2017). We investigated the relative importance of meaning and salience on the guidance of attention in scenes as indexed by attention maps based on duration-weighted fixations. We found that the spatial distribution of meaning was better able than image salience to account for the guidance of attention, both overall and when controlling for the correlation of meaning and salience. Furthermore, we found that the advantage of meaning over image salience appeared from the very beginning of scene viewing, held over both shorter and longer shifts of attention, and persisted when the central region of each scene was removed from analysis. This pattern of results is consistent with a cognitive-relevance theory of scene viewing in which attentional priority is assigned to scene regions based on semantic information rather than visual salience.

*Keywords: attention, scene perception, eye movements*

## Acknowledgments

We thank the members of the UC Davis Visual Cognition Research Group for their feedback and comments. Research reported in this article was supported by the National Science Foundation under award number BCS-1636586, and by the National Eye Institute of the National Institutes of Health under award 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.

Commercial relationships: none.
Corresponding author: John M. Henderson.
Email: johnhenderson@ucdavis.edu.
Address: Center for Mind and Brain, University of California, Davis, CA, USA.

## References

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300, http://www.jstor.org/stable/2346101.

Biederman, I. (1972, July 7). Perceiving real-world scenes. *Science*, *177*(4043), 77–80, https://doi.org/10.1126/science.177.4043.77.

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*(23), 2577–2587, https://doi.org/10.1016/j.visres.2010.08.016.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207, https://doi.org/10.1109/TPAMI.2012.89.

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision*, *14*(13):3, 1–32, https://doi.org/10.1167/14.13.3. [PubMed] [Article]

Borji, A., Sihite, D. N., & 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, https://doi.org/10.1109/TIP.2012.2210727.

Buswell, G. T. (1935). *How people look at pictures*. Chicago, IL: University of Chicago Press.

Buswell, G. T. (1936). *How People Look at Pictures. Psychological Bulletin*. Chicago, IL: University of Chicago Press. https://doi.org/10.1037/h0053409.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). *What do different evaluation metrics tell us about saliency models?* Retrieved from http://arxiv.org/abs/1604.03605

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*(9):4, 898–914, https://doi.org/10.1167/6.9.4. [PubMed] [Article]

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, https://doi.org/10.1167/3.9.67. [Abstract]

Castelhano, M. S., & Henderson, J. M. (2008a). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *62*(1), 1–14, https://doi.org/10.1037/1196-1961.62.1.1.

Castelhano, M. S., & Henderson, J. M. (2008b). 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., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3):6, 1–15, https://doi.org/10.1167/9.3.6. [PubMed] [Article].

Einhäuser, W., & Nuthmann, A. (2016). Salient in space, salient in time: Fixation probability predicts fixation duration during natural scene viewing. *Journal of Vision*, *16*(11):13, 1–17, https://doi.org/doi:10.1167/16.11.13. [PubMed] [Article].

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

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1):10, 1–29, https://doi.org/doi:10.1167/7.1.10. [PubMed] [Article].

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (pp. 545–552), Cambridge, MA: MIT Pres.

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity, *Journal of Vision*, *17*(5):23, 1–17, https://doi.org/10.1167/17.5.23. [PubMed] [Article]

Hayhoe, M. M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661305000598

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1):6, 49–63, https://doi.org/10.1167/3.1.6. [PubMed] [Article]

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

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

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

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. Van Gompel, M. H. Fischer, S. Murray, Wayne, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford, UK: Elsevier, https://doi.org/10.1016/B978-008044980-7/50027-6.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action: Eye Movements and the Visual World* (pp. 1–58). New York, NY: Psychology Press. https://doi.org/10.4324/9780203488430.

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., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271, https://doi.org/10.1146/annurev.psych.50.1.243.

Henderson, J. M., & Luke, S. G. (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, https://doi.org/10.1037/a0036330.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856. Retrieved from papers2://publication/doi/10.3758/PBR.16.5.850

Henderson, J. M., Nuthmann, A., & Luke, S. G. (2013). Eye movement control during scene viewing: Immediate effects of scene luminance on fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 318–322, https://doi.org/10.1037/a0031224.

Henderson, J. M., & Pierce, G. L. (2008). Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin & Review*, *15*(3), 566–573, https://doi.org/10.3758/PBR.15.3.566.

Henderson, J. M., & Smith, T. J. (2009). How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition*, *17*(6–7), 1055–1082, https://doi.org/10.1080/13506280802685552.

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

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203, https://doi.org/10.1038/35058500.

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. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=730558

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227, https://doi.org/10.1007/978-94-009-3833-5_5.

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. *2017 IEEE International Conference on Computer Vision* (pp. 4799–4808).

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

Laubrock, J., Cajar, A., & Engbert, R. (2013). Control of fixation duration during scene viewing by interaction of foveal and peripheral processing. *Journal of Vision*, *13*(12):11, 1–20, https://doi.org/10.1167/13.12.11. [PubMed] [Article]

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547–552, https://doi.org/10.3758/BF03210264.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*(3), 165–188. Retrieved from papers2://publication/uuid/D31BB05D-7550-4C20-B1A4-18A5DDF1086F

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*(8):17, 1–15, https://doi.org/10.1167/11.8.17. [PubMed] [Article]

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin & Review*, *24*(2), 370–392, https://doi.org/10.3758/s13423-016-1124-4.

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

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *117*(2), 382–405, https://doi.org/10.1037/a0018924.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*(B), 23–36, https://doi.org/10.1016/S0079-6123(06)55002-2.

Onat, S., Açık, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS ONE*, *9*(4),e93254. Retrieved from http://dx.plos.org/10.1371/journal.pone.0093254.g008

Parkhurst, D., Law, K., & Niebur, E. (2002). Modelling the role of salience in the allocation of visual selective attention. *Vision Research*, *42*(1), 107–123.

Potter, M. (1975, March 14). Meaning in visual search. *Science*, *187*(4180), 965–966, https://doi.org/10.1126/science.1145183.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19449261&retmode=ref&cmd=prlinks

Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research*, *47*(21), 2714–2726, https://doi.org/10.1016/j.visres.2007.05.007.

Rider, A. T., Coutrot, A., Pellicano, E., Dakin, S. C., & Mareschal, I. (2018). Semantic content outweighs low-level saliency in determining children's and adults' fixation of movies. *Journal of Experimental Child Psychology*, *166*, 293–309, https://doi.org/10.1016/j.jecp.2017.09.002.

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

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

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

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

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

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.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136, https://doi.org/http://dx.doi.org/10.1016/0010-0285(80)90005-5.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, *43*(3), 333–346, https://doi.org/10.1016/S0042-6989(02)00498-4.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition*, *17*:6-7, 812–834, https://doi.org/10.1080/13506280902771278.

Vig, E., Dorr, M., & Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. *Proceedings of the IEEE Computer Society Conference on Computer*

*Vision and Pattern Recognition* (pp. 2798–2805), https://doi.org/10.1109/CVPR.2014.358.

Võ, M. L. H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3):14, 1–13, https://doi.org/10.1167/10.3.14. [PubMed] [Article]

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.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407, https://doi.org/10.1016/j.neunet.2006.10.001.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 1–8, https://doi.org/10.1038/s41562-017-0058.

Yarbus, A. L. (1967). *Eye movements and vision*. New York, NY: Plenum Press, https://doi.org/10.1016/0028-3932(68)90012-2.

## Appendix: Patch-density parameter estimation

The optimal meaning-map grid density for each patch size was estimated by simulating the recovery of known image properties (i.e., luminance and entropy). For the sake of simplicity and visualization, the simulation procedure will be described in terms of luminance recovery, but the same procedure was also applied to edge density and entropy recovery.

The first step in the recovery simulation was to generate the ground-truth luminance image for each scene for a given patch size, which sets an upper limit on the luminance resolution that can be recovered. The ground-truth luminance image for each scene was computed by taking the scene luminance image and convolving it with a circular mean mask for a given patch size. Then the patch-density grid (simulating patch ratings) was systematically varied from 50 to 1,000 patches (fine patches) and 40 to 200 (coarse patches), and recovery of the ground truth was performed for each grid. The recovery procedure consisted of taking the mean of each patch from the original luminance image and then using thin-plate interpolation to interpolate between the patches across each grid. If the patch density was low enough that the entire image was not tiled, then the background was set to the mean value across all the patch samples in the grid.

Figure A1 shows an example of the recovery procedure for the scene shown in Figure 1a for patch densities of 88 (a) and 300 (b). As can be seen by comparing the ground truth (left) to the interpolated recovery (right), a patch density of 300 provides an excellent estimate of the ground truth. Figure A2 shows luminance, edge density, and entropy recovery ($R^2$) for the fine patch size (a) and the coarse patch size (b) as a function of patch density. Recovery improvement plateaus at a patch density of 300 patches for the fine patch size and 108 patches for the coarse patch size.

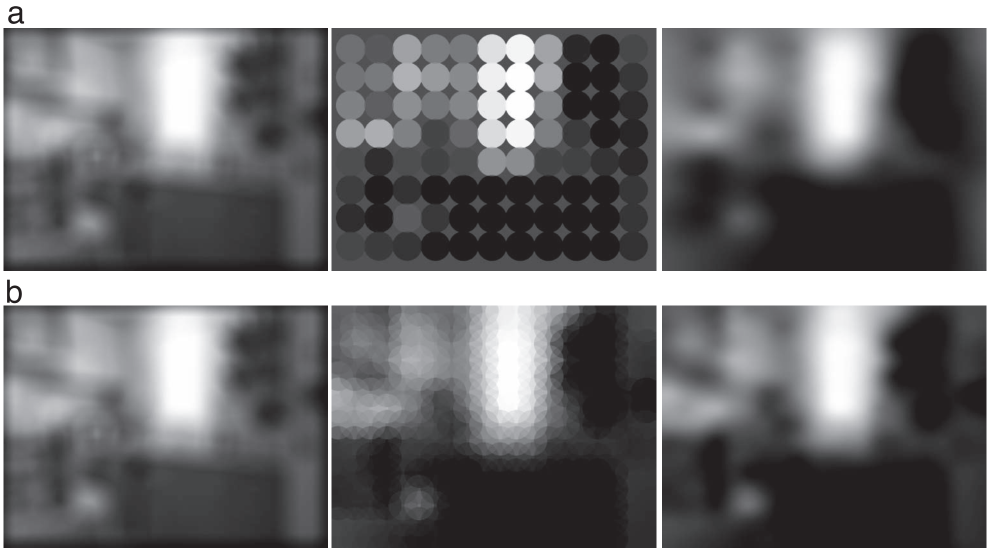


Figure A1. Example of scene luminance recovery. From left to right, the ground-truth luminance, simulated fine-patch rating density, and interpolated recovery images are shown for patch densities of (a) 88 and (b) 300. A comparison of the ground truth and recovery indicates that a patch-density value of 300 provided excellent recovery.

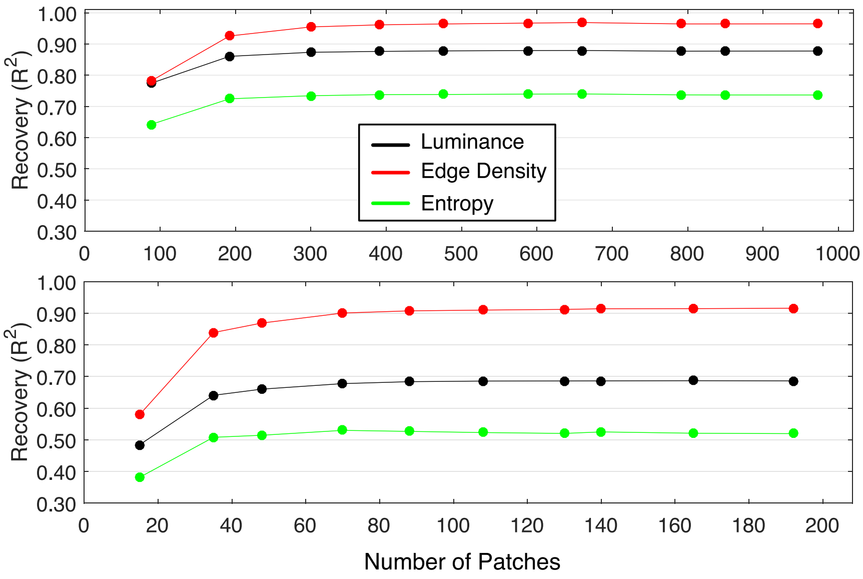


Figure A2. Ground truth recovery as a function of patch density for the fine and coarse patch sizes. The top panel shows the ground-truth recovery ($R^2$) across all 40 scenes for luminance, edge density, and entropy for the fine patch size. The bottom panel shows the corresponding ground-truth recovery ($R^2$) for the coarse patch size. Error bars represent standard error of the mean.

It is worth noting that the recovery procedure makes two assumptions. First, it assumes that meaning can be interpolated from subsampling similarly to luminance, edge density, and entropy. Second, it assumes that our rating task provides an accurate estimate of meaning at each patch-sample location. A priori, we did not know whether these assumptions about meaning or our rating task were satisfied. While we still do not know whether the selected patch densities or rating task are optimal for measuring meaning, the accuracy of the meaning-map prediction results suggests that the recovery simulations using image features provided reasonable sample density values for each patch size, and that the rating task provided reasonably accurate estimates of patch meaning.