Meaning Guides Attention in Real-World Scenes:

Evidence from Eye Movements and Meaning Maps

John M. Henderson[1,2] and Taylor R. Hayes[1]

[1]Center for Mind and Brain

[2]Department of Psychology

University of California, Davis

Running Head: Attentional Guidance in Scenes

Keywords: Attention, scene perception, eye movements

Correspondence:
John M. Henderson
Center for Mind and Brain
267 Cousteau Place
University of California
Davis, CA 95618
johnhenderson@ucdavis.edu

Abstract

We compared the influences of meaning and salience on attentional guidance in scenes. 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’ spatial distribution of attention over scenes, with attention operationalized as duration-weighted fixation density. The results showed that both meaning and salience predicted the distribution of attention, but that when the correlation between meaning and salience was statistically controlled, meaning accounted for unique variance in attention but salience did not. This pattern was observed for early as well as late fixations, for fixations following short as well as long saccades, and for fixations including or excluding the centers of the scenes. The results strongly suggest that meaning guides attention in real world scenes. We discuss the results from the perspective of the cognitive relevance theory of attentional guidance in scenes.

Introduction

We can only attend to 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, 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 scenes 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 focus on the attraction of attention by primitive visual features and feature differences (Treisman & Treisman, 1980; see Wolfe & Horowitz, 2017). When applied to real world scenes, the most influential instantiation of this type of theory is based on visual salience, which proposes that visual saliency maps are generated by pooling contrasts in semantically uninterpreted features from image dimensions such as luminance, color, and edge orientation (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Harel, Koch, & Perona, 2006; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002). On this view, attentional guidance is fundamentally a reaction to image features in the scene, with attention captured or “pulled” to visually salient scene regions (Henderson, 2007). The apeal of image guidance theory based on visual saliency is due in part to the fact that visual salience is both neurobiologically inspired and computationally tractable (Henderson, 2017).

Image guidance theory can be contrasted with cognitive guidance theory. On 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, cognitive relevance model (Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 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 task and goals (Buswell, 1935; Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Henderson, 2003, 2007, 2017; Henderson & Hollingworth, 1999; Rothkopf, Ballard, & Hayhoe, 2007; Tatler, Hayhoe, Land, & Ballard, 2011; Võ & Wolfe, 2013; Yarbus, 1967). Cognitive relevance proposes that the representations used to assign task relevance and meaning 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).

There is considerable evidence that image salience often does a poor job of accounting for attention in real-world scene viewing (Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007, 2009; Tatler et al., 2011; Underwood, Foulsham, & Humphrey, 2009). Indeed, most proponents of image guidance acknowledge that meaning must also play some role in attentional guidance. Nevertheless, much of the research on attentional guidance in real-world scenes 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 quantifying image features than it is to quantify meaning (Henderson, 2017). To investigate meaning and to compare its influence to salience, it is necessary to represent both constructs so that comparable quantitative predictions can be generated from them.

To provide a method for directly contrasting the influences of meaning and salience on the guidance of attention, we have 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 that region would be to recognize (Antes, 1974) or how informative it is (Mackworth & Morandi, 1967). In both studies, the eye movements of a different group of subjects were measured while they viewed the rated images. In general, viewers looked more at the more highly rated regions. We modified and extended these methods to develop meaning maps for real world scenes. We used crowd-sourced responses in which we asked naïve subjects 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 are taken and raters were asked to indicate how meaningful each patch was (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 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 spatially distributed non-uniformly across scenes, with some scene regions relatively rich in semantic content and others relatively sparse.

Because meaning maps are represented in the same format as saliency maps, they can be directly compared to saliency maps. 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 scene. They 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 image 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 saliency.

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 without taking into account fixation durations (Henderson & Hayes, 2017). The fixation location analysis was an important first step because the research assessing saliency maps to date has similarly focused on fixation location (Borji et al., 2014, 2013; Harel et al., 2006; Itti & Koch, 2001; Parkhurst et al., 2002). However, the durations of fixations 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, fixation is directed to that entity for more time (Henderson, Nuthmann, & Luke, 2013; Henderson, Weeks, & Hollingworth, 1999; Henderson & Pierce, 2008; Henderson & Smith, 2009; Laubrock, Cajar, & Engbert, 2013; 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 taking into account how long attention is focused at each location.

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 method, meaning maps, 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 in Henderson and Hayes (2017).

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 only allowed to participate in the study once. Subjects were paid $0.50 cents per assignment and all subjects provided informed consent.

Stimuli. Stimuli were 40 digitized 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 spatial scales of 3° and 7° (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 patches (see Appendix), suggesting that this method is sufficiently sensitive to underlying scene structure. The full patch stimulus set consisted of 12000 unique 3° patches and 4320 unique 7° patches for a total of 16320 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 they thought it was. Subjects were first given examples of two low-meaning and two high-meaning scene patches to make sure they understood

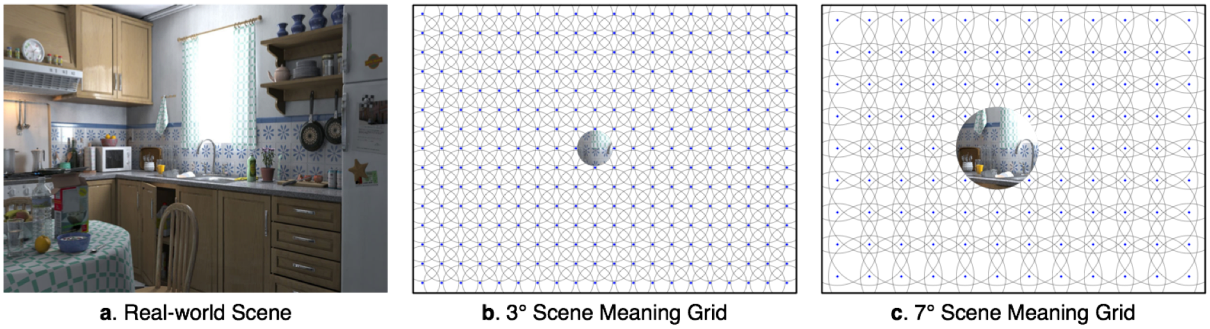

Figure 1. Real-world scene and corresponding tiled patch grids. (a) Example real-world scene. (b) Overlapping circular patches used for meaning rating at 3° and (c) at 7° spatial scales. The blue dots in (b) and (c) denote the center of each circular patch and the image circles show examples of the content captured by the 3° and 7° scales for the example scene

the rating task, and then 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 3 times by 3 independent raters for a total of 48960 ratings. However, due to the high degree of overlap across patches, each patch contained rating information from 27 independent raters for each 3° patch and 63 independent raters for each 7° patch. Figure 2 shows the distribution of ratings and the highest and lowest rated non-overlapping patches across all scenes at the two patch sizes. The lowest rated patches tended to come from the edges of the pictures,

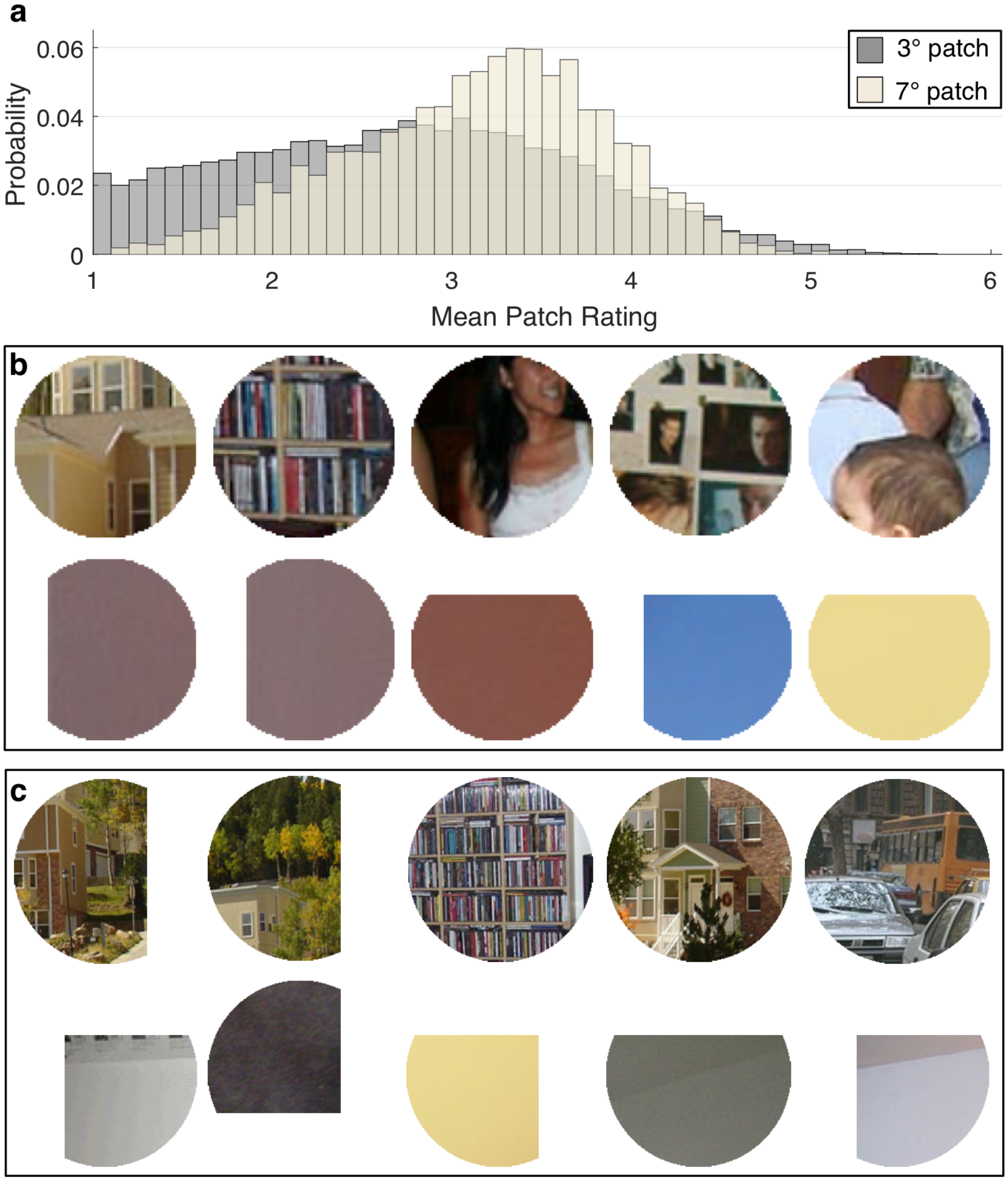

Figure 2. Rating distributions and example high and low patches. (a) Distribution of ratings for 3 and 7 patches across all raters and scenes. (b) Example highest and lowest rated non-overlapping patches for 3° and (c) 7° patches.

which accounts for their truncated shapes.

Meaning maps were generated from the ratings by averaging, smoothing, and then combining 3° and 7° maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average 3° and 7° rating map for each scene. The average 3° and 7° rating maps were then smoothed using thin-plate spline interpolation (Matlab 'fit' using the 'thinplateinterp' method). Finally, the smoothed 3° and 7° maps were combined using a simple average, i.e., (3° map + 7° map)/2). This procedure was used to create a meaning map for each scene. The final map was blurred using a Gaussian kernel followed 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 (Borji et al., 2013; Henderson et al., 2007; Tatler, 2007). This center bias operation is also commonly applied to saliency maps.

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 same center bias operation described for the meaning maps was applied to the saliency maps to down-weight the periphery.

Histogram Matching. Meaning and saliency maps were normalized to a common scale using image histrogram 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.

Eyetracking 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 subjects were naive concerning the purposes of the experiment and provided informed consent. The eye movement data from each subject was inspected for excessive artifacts caused by blinks or loss of calibration due to incidental movement by examining the mean percent of signal across all trials using Matlab. Fourteen subjects with less than 75% signal were removed, leaving 65 subjects for analysis who tracked very well (mean signal = 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 the initial study of meaning maps (Henderson & Hayes, 2017).

Apparatus. Eye movements were recorded with an EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01) sampling at 1000 Hz (SR Research, 2010b). Subjects sat 90 cm away from a 21" monitor, so that scenes subtended approximately 33°x25° of visual angle. 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. The memory test was not administered. Each trial began with fixation on a cross at the center of the display for 300ms. Following central fixation, each scene was presented for 12s 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; SR Research, 2010b).

Eye movement data were imported offline into Matlab using the EDFConverter tool. The first fixation, always located at the center of the display as a result of the pretrial fixation 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. 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.

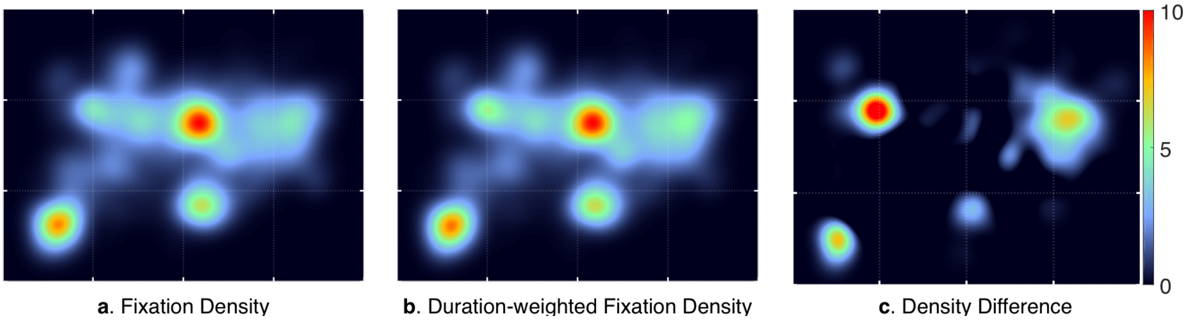

Figure 3. *Duration-weighted fixation density*. Example (a) fixation density and (b) duration-weighted fixation density, for all fixations on one scene. (c) The density difference depicting the absolute-value difference in the two densities, with hotter regions representing greater difference.

To create duration-weighted attention maps, a duration weight was generated for every 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 2 parameters for these distributions, an upper bound 95th percentile cutoff (any values in the 95 percentile received a weight value of 1.0) and the 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 -6dB was applied to the matrix for each scene to account for foveal accuity/eye tracker error. An example of a resulting duration-weighted attention map is shown in Figure 4c.

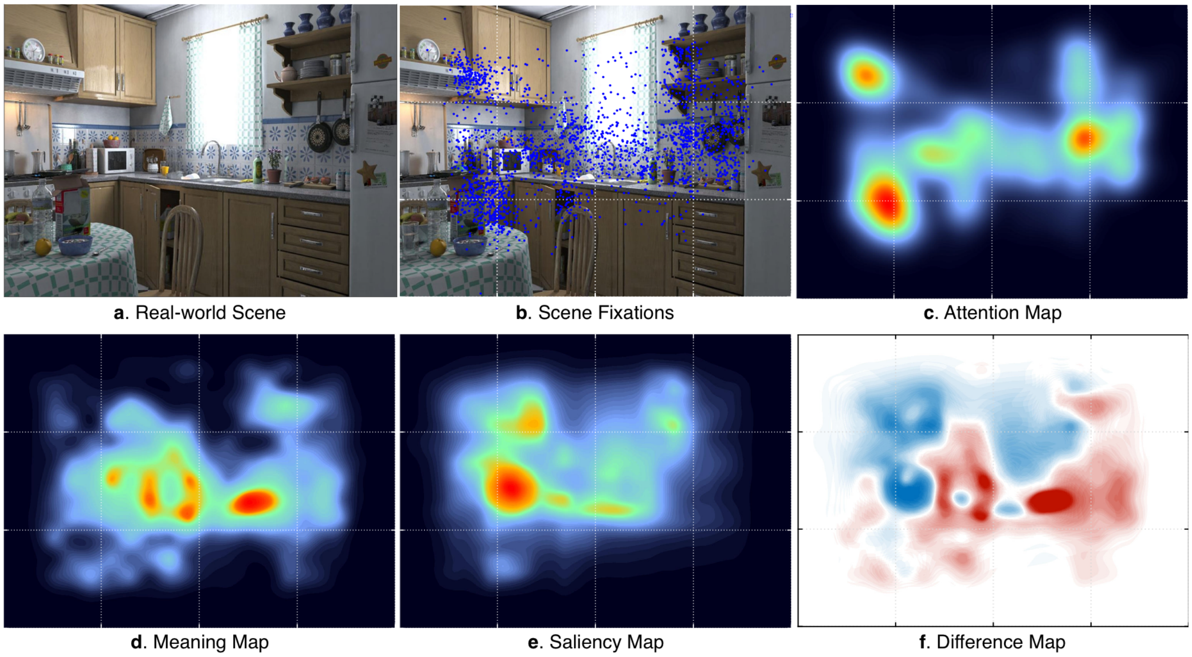

Figure 4. Example data used in the analyses. (a) Real-world scene, (b) viewers' fixations superimposed on the scene as blue dots, and (c) the duration-weighted attention map derived from the fixations. (d) Meaning map and (e) saliency map for the example scene, and (f) the difference between the meaning and saliency maps, with regions of greater meaning shown in read and greater saliency shown in blue.

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 meaning maps and saliency 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, 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 in the pesent study was how well the two types of prediction maps (meaning and saliency 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) statistically predicted the spatial distribution of attention (Figure 4b) as captured 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.

Figure 6 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 density (M=0.50, SD=0.12) and saliency account for 35% of the variance in fixation density (M=0.35, SD=0.12). A two-tailed t-test revealed this difference was statistically significant, $t(78) = 5.38$, $p < .0001$, 95% CI [0.09,0.20].

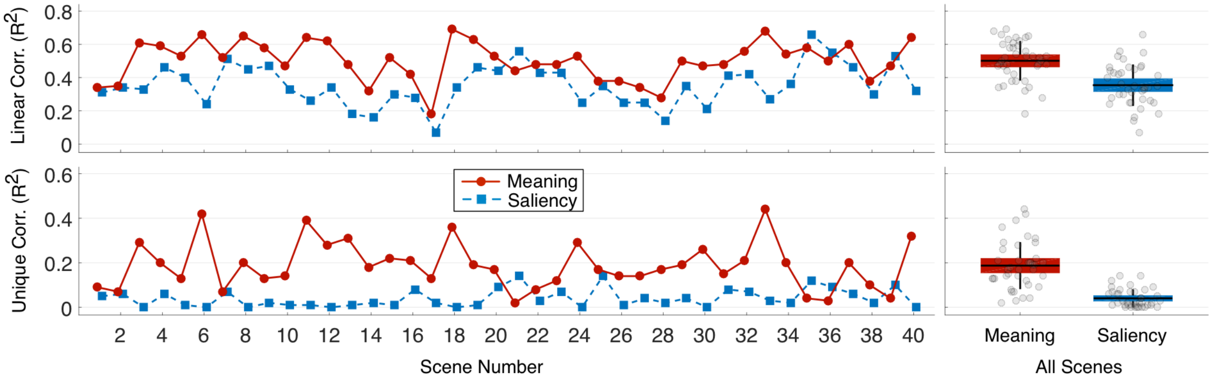

Figure 5. Squared linear correlation and semi-partial correlation by scene and across all scenes. The line plots show the linear correlation (top) and semi-partial correlation (bottom) between duration-weighted fixation density and meaning and salience by scene. The scatter box plots on the right show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and 1 standard deviation (black vertical line) for meaning and salience across all 40 scenes.

To examine the unique variance in attention explained by meaning and salience when controlling for their shared variance, we computed squared semi-partial 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 saliency maps accounted for a non-significant 4% additional variance after controlling for meaning as saliency (M=0.04, SD=0.04). A two-tailed t-test confirmed that this difference was statistically significant, $t(78) = 8.22$, $p < .0001$, 95% CI [0.11, 0.18]. These results show that meaning explained the distribution of attention over scenes better than salience.

It has been been proposed that attention is initially guided by image salience, but that as viewing progresses over time, meaning begins to play a greater role (Anderson, Donk, & Meeter, 2016; Anderson, Ort, Kruijne, Meeter, & Donk, 2015; see also Henderson & Hollingworth, 1999). To test this proposal, we conducted temporal time-step analyses. Linear correlation and semi-partial correlations were conducted based on a series of attention maps, with each map generated from each sequential eye fixation (i.e., 1st, 2nd, 3rd fixation, etc.) in each scene. This 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 3 fixations, whereas salience accounted for only 9.5%, 15.2%, and 16.6% of the variance in the first 3 fixations, respectively. 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 (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 3 fixations, whereas salience accounted for 2.7%, 4.6%, and 4.2% of the unique variance in the first 3 fixations, respectively. In conclusion, counter to the salience-first hypothesis but consistent with results based on unweighted fixations reported (Henderson & Hayes, 2017), in both the correlation and semi-partial correlation analyses meaning accounted for more variance in attention than salience from the very first fixation. These results indicate that meaning begins guiding attention as soon as a scene appears.

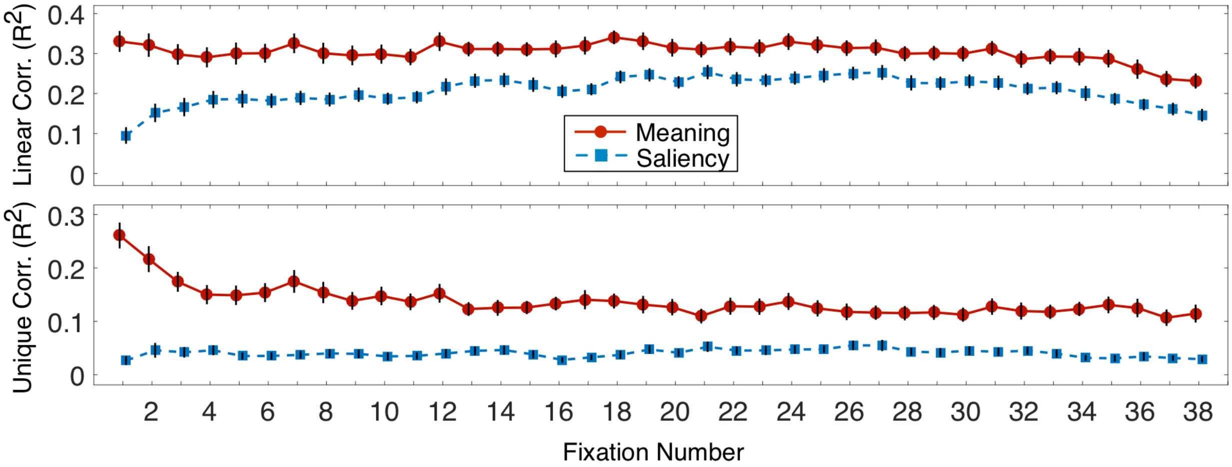

Figure 6. Squared linear correlation and squared semi-partial correlation as a function of fixation number. The top panel shows the squared linear correlation between duration-weighted fixation density and meaning and salience as a function of fixation order across all 40 scenes. The bottom panel shows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

Central Region Knock-Out Analyses

It is commonly found in eyetracking studies that viewers tend to concentrate their fixations near the center of a real world scene and rarely fixate the outside borders of a scene (Borji et al., 2013; Henderson et al., 2007; Tatler, 2007). As noted in the methods, in creating the final meaning maps, we used a multiplicative center bias operation to down-weight the scores in the periphery, 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 can not be due to differences in the ability of meaning maps to predict central fixations. The results of these analyses were qualitatively and quantitatively very similar to the complete analyses.

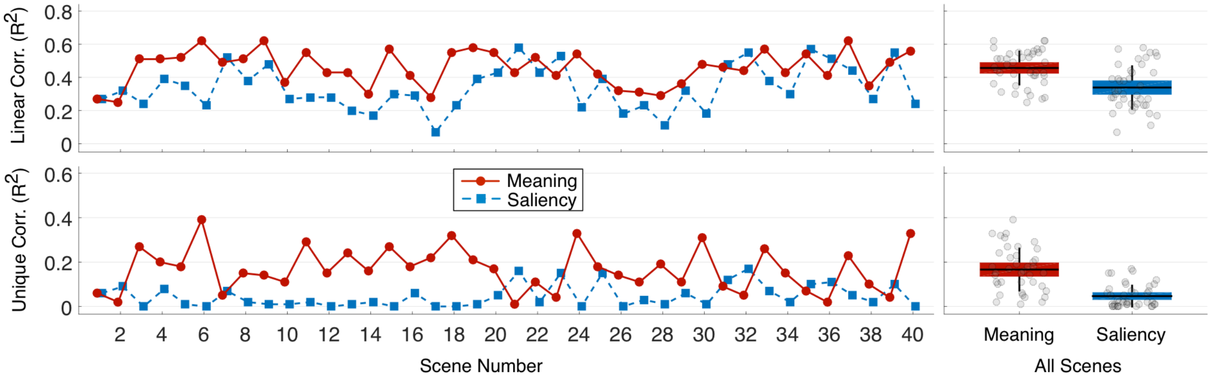

Figure 7. Squared linear correlation and semi-partial correlation by scene and across all scenes with 7° center removed. The line plots show the linear correlation (top) and semi-partial correlation (bottom) between duration-weighted fixation density and meaning and salience by scene after removing the 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 1 standard deviation (black vertical line) for meaning and salience across all 40 scenes.

Figure 7 presents the linear 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 9 shows the squared linear correlations. On average across the 40 scenes, meaning accounted for 46% of the variance in fixation density (M=0.46, SD=0.11) and saliency account for 34% of the variance in fixation density (M=0.34, SD=0.13). A two-tailed t-test revealed this difference was statistically significant, $t(78) = 4.39$, $p < .0001$, 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 semi-partial correlations. These correlations, shown in the bottom of Figure 7, revealed that across the 40 scenes, meaning captured more than 3 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* < .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.

To test whether the overall advantage of meaning over salience early in viewing ws 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 semi-partial 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. Using the same testing and false discovery rate correction as in the main analyses, 34 of 38 time points were significantly different in both the linear and semi-partial analyses (*FDR* < 0.05), excluding fixations 21, 25, 27, and 28. Importantly for assessing initial control of attention during scene viewing, in the linear correlation analysis (top of Figure 8), meaning accounted for 22.9%, 27.0%, and 26.7% of the variance in the first three fixations, whereas salience accounted for only 10.2%, 14.9%, and 16.2% of the variance in the first three fixations. Critically, when controlling for the correlation among the two prediction maps with semi-partial 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 Supplementary Figure 8 (*FDR* < 0.05). Meaning accounting for 17.9%, 17.8%, and 15.7% of the unique variance in the first 3 fixations, whereas salience accounted for 5.2%, 5.6%, and 5.4% of the unique variance in the first three fixations, respectively. Consistent with the overall correlation and semi-partial correlation analyses, meaning produced an advantage over salience from the very first fixation even when the central region of each map was removed from the 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

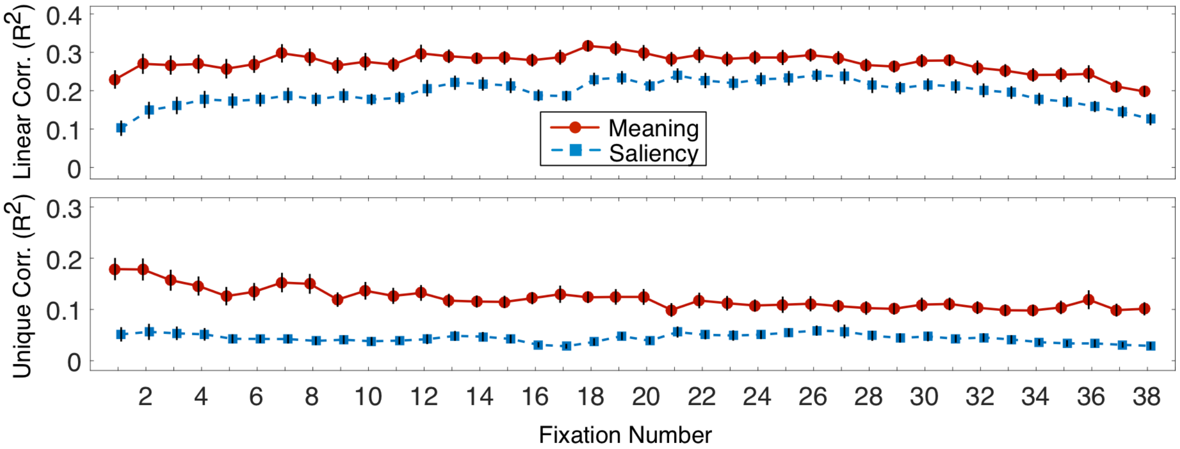

Figure 8. Squared linear correlation and squared semi-partial correlation as a function of fixation number with 7° center removed. The top panel shows the squared linear correlation between fixation density and meaning (red) and salience (blue) as a function of fixation order averaged over all 40 scenes.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

especially strong evidence for the control of attention by meaning because removing the central 7° should disadvantage the meaning maps because photographers tend to center meaningful information in photographs (Tatler, 2007). Nevertheless, the meaning maps continued to outperform the saliency maps in both the overall variance and unique variance accounted for in the attention maps.

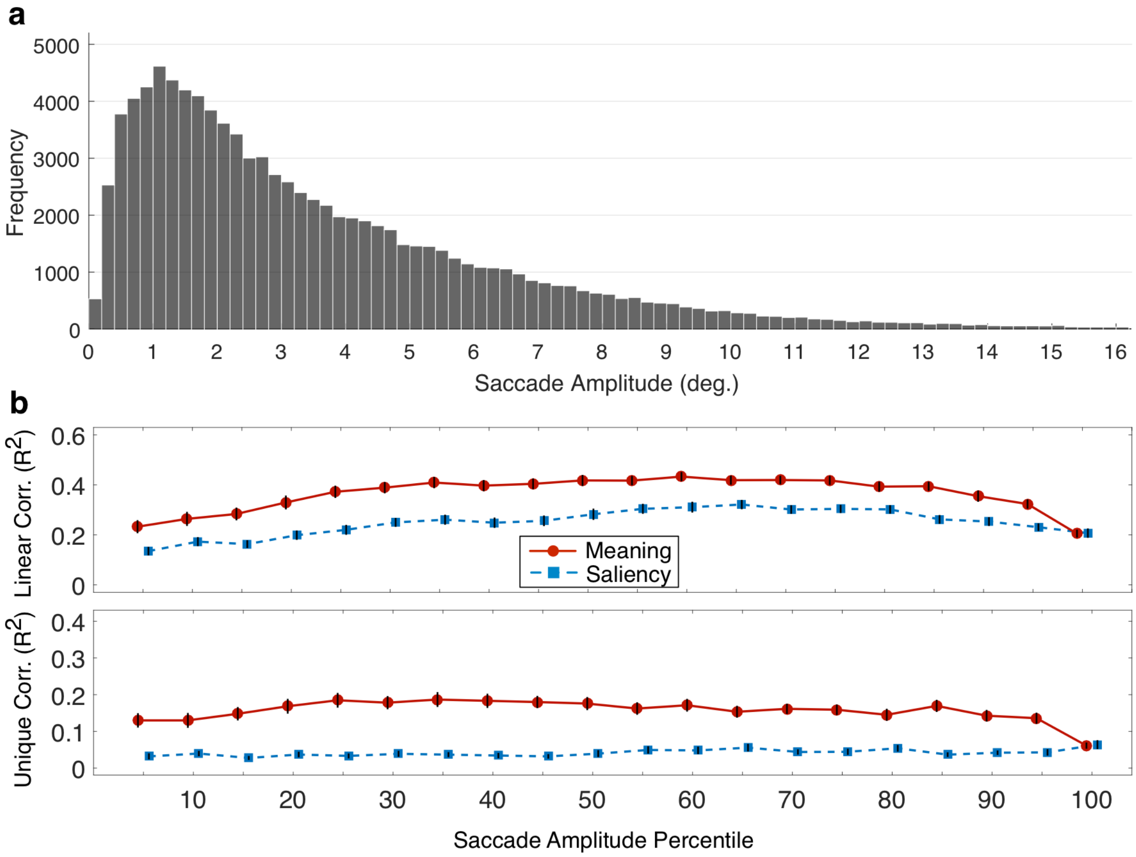

Figure 9. Squared linear correlation and squared semi-partial correlation as a function of saccade amplitude to fixation. (a) The distribution of saccade amplitudes observed in the experiment. (b) The squared linear correlations between duration-weighted fixation density for meaning and salience as a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

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 of 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 degrees, but as 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 data 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 decile. 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 decile in both figures (*FDR* < .05). 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 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 contents 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. Recently we introduced a new method for identifying and representating 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 dominance of meaning over sailence 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, on the *cognitive relevance* model (Henderson et al., 2007, 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, and not by its visual features or salience. On this model, meaning determines attentional priority, with image properties used only to generate perceptual (“proto-”) objects and other perceptually based potential saccade targets. Critically, then, attentional priority is assigned to potential attentional targets not based image saliency, but rather based on knowledge representations (e.g., knowledge about what objects are likely to be present and where those objects are likely to be found). In this model, the visual stimulus is relevant in that it is used to generate perceptual (“proto”) 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 a landscape ranked by salience. Instead, on this model, knowledge representations provide the attentional priority ranking to the targets based on their meaning (Henderson, 2003; Henderson et al., 2007, 2009).

It is important to note that the cognitive relevance model does not require 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; Potter, 1975) and can guide attention at the very earliest points of scene viewing (Castelhano & Henderson, 2003, 2008b; Henderson & Hollingworth, 1999; Oliva & Torralba, 2006). Apprehending the gist allows access to schema representations that provide constraints on what objects are likely to be present and where those objects are likely to be located (Henderson, 2003; Henderson & Hollingworth, 1999). Information retrieved from memory schemas can be combined with low-quality peripheral 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 and can be used to generate predictions for guiding attention to regions that have not yet been identified (Henderson, 2017). Second, as shown in the present study as well as many others (Henderson et al., 1999), most saccades during scene viewing are relatively short, with the 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 visual information that is relatively close to the fovea, where identity and meaning can easily be ascertained. Extraction of meaning from nearby 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 described above.

The present results at first glance appear to be at odds with past studies that have shown strong correlations between visual salience and attention. How can we account for this discrepancy? 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 hypothesis, and strong support was found for it, with a correlation of 0.80 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 study of human vision in which the goal is to provide a theoretical account of how the brain guides attention, a focus on 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 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), and it could be that under other task instructions, image salience would play a greater role than meaning. Furthermore, comparing the results from Henderson and Hayes (2017) and the present results show that the results were similar for attention maps created from duration-weighted and unweighted fixation densities. It may be that differences in these attention maps would be more pronounced for other types of tasks. While this is a possibility, the memorization task 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 image-based over meaning-based attentional guidance. Nevertheless, we can not rule out the possibility that salience might play a more important role in other tasks, and it will be important to assess the relative influences of meaning and salience in guiding attention across viewing tasks.

Second, although meaning was the stronger predictor of attention for the majority of scenes (36 out of 40 scenes) and overall on average, salience did perform better for 4 of the scenes. The question arises why these latter 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 noise-based reversal of the actual 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 in those particular scenes, but a future direction for study will be to compare different classes of scenes (e.g., indoor versus outdoor; natural versus man-made) 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. Here we 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.

Conclusion

In this study we employed recently developed methods for comparing the relationship between the spatial distribution of meaning and image salience and 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 stronger influence of meaning persisted when the central region of each scene was removed from the analyses, appeared from the very beginning of scene viewing, and held over both shorter and longer shifts of attention. 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 value rather than visual salience.

References

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

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 9. Retrieved from http://jov.arvojournals.org/article.aspx?doi=10.1167/15.5.9

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62–70.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. http://doi.org/10.2307/2346101

Biederman, I. (1972). Perceiving Real-World Scenes. *Science*. http://doi.org/10.1126/science.177.4043.77

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. Retrieved from http://www.journalofvision.org/lookup/doi/10.1167/14.13.3

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. http://doi.org/10.1109/TIP.2012.2210727

Buswell, G. T. (1935). *How People Look at Pictures*. University of Chicago Press Chicago. Retrieved from papers2://publication/uuid/EA9683B1-9AB3-46F7-B4C3-1EBE6E3D645D

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *arXiv*, 1–23. 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), 898–914. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17083283&retmode=ref&cmd=prlinks

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), 67a. http://doi.org/10.1167/3.9.67

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. Retrieved from http://doi.apa.org/getdoi.cfm?doi=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. Retrieved from http://doi.apa.org/getdoi.cfm?doi=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). http://doi.org/10.1167/9.3.6

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 2.1-19. http://doi.org/10.1167/8.2.2

Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. *Advances in Neural Information Processing Systems.*, 1–8. http://doi.org/10.1.1.70.2254

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity, *17*, 1–17. http://doi.org/10.1167/17.5.23.doi

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), 49–63. http://doi.org/10.1167/3.1.6

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661303002481

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219–222. Retrieved from http://www.psy.ed.ac.uk/people/jhender9/Henderson_CDir_2007.pdf

Henderson, J. M. (2017). Gaze Control as Prediction. *Trends in Cognitive Sciences*, *21*(1), 15–23. http://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 Ltd. http://doi.org/10.1016/B978-008044980-7/50027-6

Henderson, J. M., & Ferreira, F. (2004). *Scene perception for psycholinguists*. *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. http://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*(October), 743–747. http://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. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=10074679&retmode=ref&cmd=prlinks

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. http://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.

http://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 and Review*, *15*(3), 566–573. http://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. Retrieved from papers2://publication/uuid/DD6DE50F-58CD-413E-9301-76065FA9F07F

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.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11256080&retmode=ref&cmd=prlinks

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *20*(11), 1254–1259. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=730558

Land, M. F., & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25–26), 3559–3565. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11718795&retmode=ref&cmd=prlinks

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), 1–20. Retrieved from http://www.journalofvision.org/lookup/doi/10.1167/13.12.11

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

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), 1–15. Retrieved from http://www.journalofvision.org/lookup/doi/10.1167/11.8.17

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

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. Retrieved from http://doi.apa.org/getdoi.cfm?doi=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*(Chapter 2), 23–36. http://doi.org/10.1016/S0079-6123(06)55002-2

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). Meaning in visual search. *Science*, *187*(4180), 965–966. Retrieved from http://www.sciencemag.org/cgi/content/abstract/187/4180/965

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. Retrieved from papers2://publication/doi/10.1016/j.visres.2007.05.007

Rothkopf, C. a, Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 16.1-20. http://doi.org/10.1167/7.14.16

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. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=18217799&retmode=ref&cmd=prlinks

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. Retrieved from http://www.journalofvision.org/lookup/doi/10.1167/11.5.5

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766–786. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17014302&retmode=ref&cmd=prlinks

Treisman, A. M., & Treisman, A. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, *12*, 97–136. Retrieved from http://www.cse.psu.edu/~rtc12/CSE597E/papers/treismanFeatIntegration.pdf

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during …. *Visual Cognition*. Retrieved from http://www.ingentaconnect.com/content/psych/pvis/2009/00000017/F0020006/art00003

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. http://doi.org/10.1016/j.cognition.2012.09.017

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

Yarbus, A. L. (1967). *Eye movements and vision*. Plenum Press. http://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 (i.e., 3° and 7°). Then, the patch density grid (simulating patch ratings) was systematically varied from 50 to 1000 patches (3°) and 40 to 200 (7°) 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.

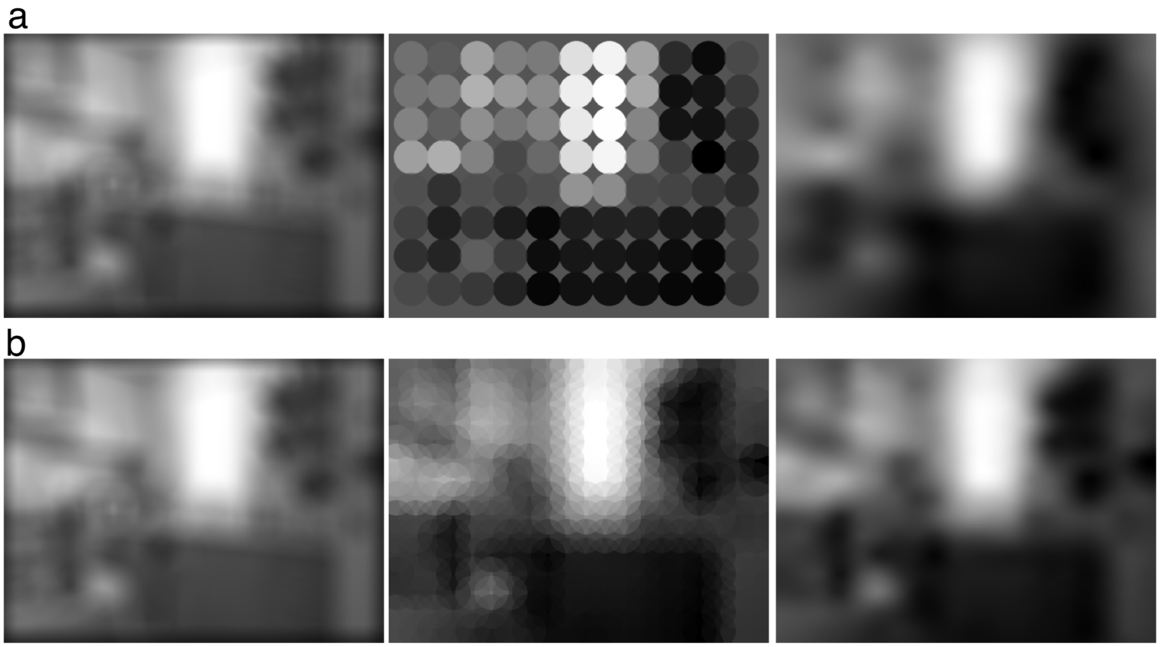

Figure A1. Example of scene luminance recovery. From left to right, the 3° ground truth luminance, simulated rating density, and interpolated recovery images are shown for a patch density of 88 (a) and a patch density of 300 (b). A comparison of the ground truth and recovery indicates that a patch density value of 300 provided excellent recovery.

Figure A1 shows an example of the recovery procedure for the scene shown in Figure 1a for a patch density 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 3° patch size (a) and the 7° patch size (b) as a function of patch density. Recovery improvement plateaus at a patch density of 300 patches for the 3° patch size and 108 patches for the 7° patch size.

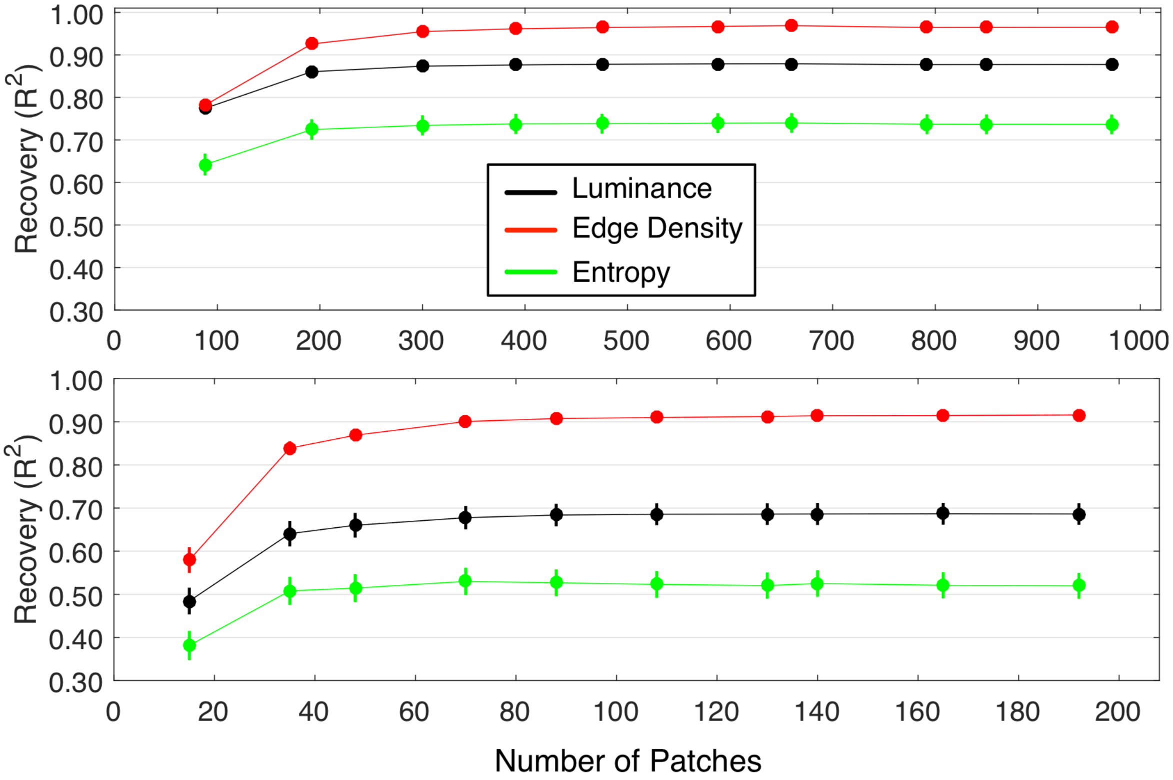

Figure A2. *Ground truth recovery as a function of patch density for 3° and 7° patch sizes*. The top panel shows the ground truth recovery (R^2) across all 40 scenes for luminance, edge density, and entropy for the 3° patch size. The bottom panel shows the corresponding ground truth recovery (R^2) for the 7° 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 sub-sampling like luminance 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 can not judge whether the selected patch densities or rating task are optimal for measuring meaning, the accuracy of the meaning map prediction results suggests the recovery simulations using luminance and entropy provided reasonable sample density values for each patch size and the rating task provided reasonably accurate estimates of patch meaning.