HHS Public Access

Author manuscript

Acta Psychol (Amst). Author manuscript; available in PMC 2020 July 11.

Published in final edited form as:

Acta Psychol (Amst). 2019 July ; 198: 102889. doi:10.1016/j.actpsy.2019.102889.

The Role of Meaning in Attentional Guidance during Free Viewing of Real-world Scenes

Candace E. Peacock[1,2], **Taylor R. Hayes**[1], **John M. Henderson**[1,2]

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

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

Abstract

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

Keywords

attention; scene perception; eye movements; meaning; context; saliency

---

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

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

Additional Information

The authors declare no competing financial interests. Additionally, these data have not previously been published.

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

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

To address this challenge, Henderson and Hayes (2017) introduced *meaning maps* as a semantic analog of saliency maps. Specifically, meaning maps were designed to capture the spatial distribution of semantic features in a scene in the same format that saliency maps use to capture the spatial distribution of image features. The key idea of a meaning map is that it represents the spatial distribution of semantic informativeness over a scene in the same format as a saliency map represents the spatial distribution of image salience. Inspired by two classic scene viewing studies (Antes, 1974; Mackworth & Morandi, 1967), meaning maps are created using crowd-sourced ratings given by large numbers of naive subjects. These subjects rate the meaningfulness of individual scene patches taken from dense arrays of objectively defined circular overlapping patches at two spatial scales (Figure 1). Meaning maps are then constructed for each scene by averaging these ratings and smoothing the results (Figure 2). Meaning maps represent the spatial distribution of meaning across the scene, providing a means for directly comparing meaning and image salience and their relationships with attentional guidance. Research based on meaning maps has shown that meaning is a better predictor of attentional guidance than image salience across several

active viewing tasks including scene memorization, aesthetic judgment (Henderson & Hayes, 2017, 2018), and scene description (Henderson, Hayes, Rehrig, & Ferreira, 2018).

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

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

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

Method

Eye-tracking

Participants.—Thirty-two University of California, Davis, undergraduate students with normal to corrected-to-normal vision participated in the experiment (24 females, average age = 20.91). All participants were naive to the purpose of the study and provided verbal consent. The eye movement data were inspected for excessive artifacts due to blinks or loss of calibration. Following Henderson and Hayes (2017), we averaged the percent signal ([number of good samples / total number of samples] × 100) for each trial and participant using custom MATLAB code. The percent signal for each trial was then averaged for each participant and compared to an *a priori* 75% criterion for signal. Outlier removal was then conducted by trial and participant. If a trial had less than 75% signal, it was excluded from analysis. Furthermore, if a participant's average percent signal was less than 75%, that entire participant was excluded from analysis. In total, no individual trials were excluded based on these criteria. Because two participants had lower than 75% signal, their data were excluded from analyses, resulting in a total of 30 participants/datasets analyzed. The number of participants used in the current study (N = 30) was derived from previous meaning map studies using 30 participants (Henderson et al., 2018; Peacock et al., 2019).

Apparatus.—Eye movements were recorded using an EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01° rms) sampling at 1000 Hz (SR Research, 2010b). Participants sat 85 cm away from a 21" monitor, so that scenes subtended approximately 26.5° × 20° of visual angle at 1024×768 pixels. Head movements were minimized by using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software (SR Research, 2010a). Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s^2; SR Research, 2010b). Eye movement data were imported offline into Matlab using the EDFConverter tool. The first fixation, always located at the center of the display as a result of the pretrial fixation marker, was eliminated from analysis. Additionally, fixations that landed off the screen, and any fixations that were less than 50ms and greater than 1500ms were eliminated as outliers. Occasionally, saccade amplitudes are not segmented correctly by EyeLink's standard algorithm, resulting in large values. Given this, saccade amplitudes > 25° were also excluded. Fixations corresponding to these saccades were included as long as they met the other exclusion criteria. This outlier removal process resulted in loss of 5.84% of the data across all subjects.

Stimuli.—Stimuli consisted of 20 digitized photographs (1024×768 pixels) of indoor and outdoor real-world scenes. Scenes were luminance matched across the scene set by converting the RGB image of the scene to LAB space and scaling the luminance channel of all scenes from 0 to 1. Luminance matching was done to ensure that there were no overly bright or dark scenes in the experiment and does not change the relative ranking of salience within a scene. All instruction, calibration, and response screens were luminance matched to the average luminance (M = 0.45) of the scenes.

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

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

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

Map Creation

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

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

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

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

Contextualized meaning maps.—Contextualized meaning maps were generated using the identical method as the context-free meaning maps (Henderson and Hayes, 2017) with the following exceptions. For contextualized maps, we instructed participants to rate how 'meaningful' a patch was based on how informative or recognizable it was in the context of the larger scene (Figure 1). Additionally, for each rating, the patch was circled in green in the context scene. Other than these changes, the rating methods were identical. Importantly, the patches were identical to those used for context-free mapping, allowing direct comparison of the meaning mapping methods. In their raw form without center bias added, the resulting contextualized maps were significantly correlated with the context free maps, ($M = 0.67$, $SD = 0.09$): $t(19) = 34.69$, $p < 0.001$, 95% CI= [0.63, 0.71]. This correlation increased with center bias applied ($M = 0.88$, $SD = 0.05$): $t(19) = 76.59$, $p < 0.001$, 95% CI = [0.85, 0.90].

Saliency maps.—Saliency maps for each scene were generated using the Graph-Based Visual Saliency (GBVS) toolbox with default settings (Harel et al., 2006). GBVS is a prominent saliency model that combines maps of low-level image features to create saliency maps (Figure 2).

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

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

Results

Whole Scene Analyses

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

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

Because our primary research question concerned the ability of meaning and salience to independently account for variance in fixations, we used semi-partial correlations. Semi-

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

Context-free meaning vs. image salience.—For the squared linear correlation, context-free meaning explained 39% of the variance in fixation density ($M = 0.39$, $SD = 0.14$) and image salience explained 24% of the variance ($M = 0.24$, $SD = 0.14$), $t(19) = 7.08$ $P < 0.001$, 95% CI = [0.10, 0.19] (Figure 3). For the semi-partial correlations, context-free meaning explained a unique 16% of the variance in fixation density controlling for salience ($M = 0.16$, $SD = 0.07$): $t(19) = 9.52$, $p < 0.001$, 95% CI = [0.13, 0.20], whereas salience uniquely explained only a unique 2% of the variance in fixation density controlling for meaning ($M = 0.02$, $SD = 0.03$): $t(19) = 2.37$, $p = 0.03$, 95% CI = [0.002, 0.03].

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

Contextualized meaning vs. context-free meaning.—For our secondary question, we investigated whether contextualized and context-free meaning maps would produce similar results. For the squared linear correlation, contextualized meaning explained 40% of the variance in fixation density ($M = 0.40$, $SD = 0.14$) and context-free meaning explained 39% of the variance in fixation density ($M = 0.39$, $SD = 0.14$), $t(19) = 1.44$, $p = 0.17$, 95% CI = [−0.007, 0.04] (Figure 3). When the variance explained by context-free meaning was statistically controlled, contextualized meaning uniquely explained 4% of the variance in fixation density ($M = 0.04$, $SD = 0.03$): $t(19) = 5.74$, $p < 0.001$, 95% CI = [0.03, 0.05]. When the variance explained by contextualized meaning was statistically controlled, context-free meaning uniquely explained 2% of the variance in fixation density ($M = 0.02$, $SD = 0.02$): $t(19) = 4.27$, $p = 0.0004$, 95% CI = [0.02, 0.03]. These results demonstrate that the two types of meaning largely account for the same variance in the distributions of fixations over scenes.

Ordinal Fixation Analyses

It has been suggested that when a scene first appears, attention might initially be guided by image salience, with meaning playing a larger role as viewing unfolds (Anderson & Donk, 2017; Anderson et al., 2015; Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999). This hypothesis predicts that the correlation between image salience and fixation density maps should be greater for earlier than later fixations, with salience dominating meaning in the earliest fixations. Alternatively, it could be that meaning guides attention from scene onset due to rapid gist apprehension (Oliva & Torralba, 2006; Potter et al., 2014)

and the use of schema to activate memory representations of where likely objects will be located in the scene (Henderson, 2003; Henderson & Hollingworth, 1999; Torralba et al., 2006) for attentional prioritization. This hypothesis predicts that meaning should account for attentional guidance at the earliest moments of scene viewing. To test these competing hypotheses in the free viewing task, we conducted an ordinal fixation analysis for the first three fixations, in which density maps were generated for each sequential fixation for each scene. The analyses focused on the first three of these fixations (1st fixation, 2nd fixation, and 3rd fixation) and proceeded as in the main analyses, with p-values corrected for multiple comparisons using the Bonferroni correction.

Context-free meaning vs. image salience.—For the squared linear correlations, context-free meaning accounted for 38%, 31%, and 20% of the variance in the first three fixations whereas salience accounted for 10%, 15%, and 11% of this variance (Figure 4), with all three ordinal fixation meaning versus salience comparisons significant (fixation 1: $t(19) = 7.71$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.20, 0.36]; fixation 2: $t(19) = 5.48$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.10, 0.22]; fixation 3: $t(19) = 3.06$, Bonferroni-corrected $p = 0.02$, 95% CI = [0.03, 0.15]). For the semi-partial correlations, meaning accounted for 30%, 19%, and 12% of the unique variance in the first three fixations (fixation 1: $t(19) = 8.92$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.23, 0.37]; fixation 2: $t(19) = 7.06$, Bonferroni-corrected $p < 0.001$, 95% CI = [0.14, 0.25]; fixation 3: $t(19) = 4.65$, Bonferroni-corrected $p = 0.001$, 95% CI = [0.07, 0.17]) and image salience accounted for 2%, 3%, and 3% of this variance (fixation 1: $t(19) = 3.00$, Bonferroni-corrected $p = 0.04$, 95% CI = [0.005, 0.03]; fixation 2: $t(19) = 4.51$, Bonferroni-corrected $p = 0.001$, 95% CI = [0.02, 0.05]; fixation 3: $t(19) = 4.06$, Bonferroni-corrected $p < 0.004$, 95% CI = [0.02, 0.05]), (Figure 4).[1]

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

General Discussion

The current study was designed to assess several questions related to understanding the roles of meaning and image salience in predicting attentional guidance in real-world scenes. The main question was whether the previously observed advantage for meaning over image salience would extend to a free viewing task that does not impose any specific top-down task constraints. Earlier comparisons of meaning and image salience have used explicit viewing tasks (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019). However, it is common to use free viewing with no explicit task in the saliency literature, and indeed saliency model benchmarks are based on free viewing (Bylinskii et al., 2015), so it was important to extend the previous results to free viewing. The present results using free

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

viewing were consistent with previous studies that have compared meaning and image salience: meaning accounted for significantly more of the overall variance in the spatial distribution of fixation density than salience. Furthermore, when the variance explained by meaning was statistically controlled, salience explained no more of the unique variance in fixation density, but when the variance explained by salience was controlled, meaning continued to explain substantial unique variance. In addition, contrary to the idea that image salience plays a major role during early scene viewing, these results held for the earliest fixations.[2] This pattern of results replicates the previous findings and extends them to free viewing. We note that when we weighted the fixations by duration to produce duration-weighted fixation density maps, all of the results held, consistent with Henderson and Hayes (2018).

The current results are consistent with the results of recent research suggesting that meaning continues to strongly influence attentional guidance in scenes even when meaning is not directly relevant to the viewer's task. Peacock et al. (2019) used tasks in which participants were cued to either count bright patches within a scene or rate scenes for their overall brightness. Despite the fact that salience was task-relevant and meaning was task-irrelevant, meaning continued to guide attention. The convergence of the findings reported by Peacock et al. (2019) and the current study suggests that the influence of meaning over salience on attention is robust and not readily influenced by tasks designed to minimize attention to meaning nor tasks designed to reduce top-down influences of task on attention.

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

An interesting issue is why the contextualized and context-free meaning maps were similarly related to attention. One possibility is that local scene meaning ratings are not significantly affected by global scene context in the absence of object-scene inconsistencies. Indeed, studies using scenes containing local object manipulations find that objects contain greater meaning when they are inappropriate (versus appropriate) to the global context of the scene which influences attention (Henderson et al., 1999; Loftus & Mackworth, 1978; Võ &

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

Henderson, 2009). The literature has also shown that scene gist (Oliva & Torralba, 2006; Potter et al., 2014) and schema representations guide our expectations of what local objects will appear in a scene given its global context (Henderson, 2003; Henderson & Hollingworth, 1999; Torralba et al., 2006). Given the importance of global scene context on attention, future conceptions of contextualized meaning maps will need to be made with scenes containing object-scene inconsistencies to fully predict how global scene context influences the meaning of local elements and attention to those elements.

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

Conclusion

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

Acknowledgements

We thank Conner Daniels, Praveena Singh, and Diego Marquez for assisting in data collection.

Funding

This research was supported by the National Eye Institute of the National Institutes of Health [grant number R01EY027792]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

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

Anderson NC, Ort E, Kruijne W, Meeter M, & Donk M (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. Journal of Vision, 75(5), 1–22. https://doi.Org/10.1167/15.5.9

Antes JR (1974). The time course of picture viewing. Journal of Experimental Psychology, 703(1), 62–70. 10.1037/h0036799

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

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

Bylinskii Z, Judd T, Oliva A, Torralba A, & Durand F (2019). What do different evaluation metrics tell us about saliency models?. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3), 740–757. [PubMed: 29993800]

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

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

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

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

Chen Y, & Zelinsky GJ (2019). Is there a shape to the attention spotlight? Computing saliency over proto-objects predicts fixations during scene viewing. Journal of Experimental Psychology: Human Perception and Performance, 45(1), 139. [PubMed: 30596438]

Clarke AD, & Tatler BW (2014). Deriving an appropriate baseline for describing fixation behaviour. Vision Research, 102, 41–51. [PubMed: 25080387]

De Graef P, Christiaens D, & d'Ydewalle G (1990). Perceptual effects of scene context on object identification. Psychological Research, 52(4), 317–329. [PubMed: 2287695]

Einhauser W, Rutishauser U, & Koch C (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. Journal of Vision, 8(2, 1–19.

Foulsham T, & Underwood G (2007). How does the purpose of inspection influence the potency of visual salience in scene perception?. Perception, 36(8), 1123–1138. [PubMed: 17972478]

Greene MR, & Fei-Fei L (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. Journal of Vision, 14(1, 14–14. 10.1167/14.E14

Hayhoe M, & Ballard D (2014). Modeling task control of eye movements. Current Biology, 24(13), R622–R628. [PubMed: 25004371]

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

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

Henderson JM (2003). Human gaze control during real-world scene perception. Trends in Cognitive Sciences, 7(11), 498–504. https://doi.Org/10.1016/i.tics.2003.09.006 [PubMed: 14585447]

Henderson JM (2007). Regarding scenes. Current Directions in Psychological Science, 76(4), 219–222.

Henderson JM (2017). Gaze control as prediction. Trends in Cognitive Sciences, 27(1), 15–23.

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

Henderson JM, & Hayes TR (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. Nature Human Behaviour, 7, 743–747. 10.1038/s41562-017-0208-0

Henderson JM, & Hayes TR (2018). Meaning guides attention in real-world scenes: evidence from eye movements and meaning maps. Journal of Vision, 18(6, 1–18. 10.1089/jmf.2012.0243

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

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

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

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

Itti L, & Koch C (2001). Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging, 70(1), 161–169.

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

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

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

Loftus GR, & Mackworth NH (1978). Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance, 4(4), 565. [PubMed: 722248]

Mackworth NH, & Morandi AJ (1967). The gaze selects informative details within pictures. Perception and Psychophysics, 2(11), 547–552.

Navalpakkam V, & Itti L (2005). Modeling the influence of task on attention. Vision Research, 45(2, 205–231. [PubMed: 15581921]

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

Oliva A, & Torralba A (2006). Building the gist of a scene: The role of global image features in recognition In Progress in Brain Research (Vol. 155, pp. 23–36). Elsevier. [PubMed: 17027377]

Parkhurst D, Law K, & Niebur E (2002). Modeling the role of salience in the allocation of overt visual attention. Vision Research, 42(1), 107–123. 10.1016/S0042-6989(01)00250-4 [PubMed: 11804636]

Peacock CE, Hayes TR, & Henderson JM (2019). Meaning guides attention during scene viewing even when it is irrelevant. Attention, Perception, and Psychophysics. 10.3758/sl3414-018-1607-7

Potter MC, Wyble B, Hagmann CE, & McCourt ES (2014). Detecting meaning in RSVP at 13 ms per picture. Attention, Perception, and Psychophysics, 76(2), 270–279.

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

Rothkopf CA, Ballard DH, & Hayhoe MM (2007). Task and context determine where you look. Journal of Vision, 7(14), 16–16.

Spotorno S, Tatler BW, & Faure S (2013). Semantic consistency versus perceptual salience in visual scenes: Findings from change detection. Acta Psychologica, 142(2, 168–176. [PubMed: 23333876]

Stirk JA, & Einderwood G (2007). Low-level visual saliency does not predict change detection in natural scenes. Journal of Vision, 7(10), 3–3.

Tatler BW, Hayhoe MM, Land MF, & Ballard DH (2011). Eye guidance in natural vision: Reinterpreting salience. Journal of Vision, 77(5), 5–5.

Torralba A, Oliva A, Castelhano MS, & Henderson JM (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review, 113(4, 766. [PubMed: 17014302]

Turano KA, Geruschat DR, & Baker FH (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. Vision Research, 43(3, 333–346. [PubMed: 12535991]

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

Wu CC, Wick FA, & Pomplun M (2014). Guidance of visual attention by semantic information in real-world scenes. Frontiers in Psychology, 5, 54. [PubMed: 24567724]

Yarbus AL (1967). Eye movements during perception of complex objects In Eye movements and vision (pp. 171–211). Springer, Boston, MA.

Highlights

- Meaning predicts overall attention better than image salience during free viewing.
- This result held for early attention, contrary to the ‘saliency first’ hypothesis.
- Contextualized meaning maps were introduced to replicate and extend context-free meaning maps.
- Context-free and contextualized meaning are significantly correlated and similarly predict attention.

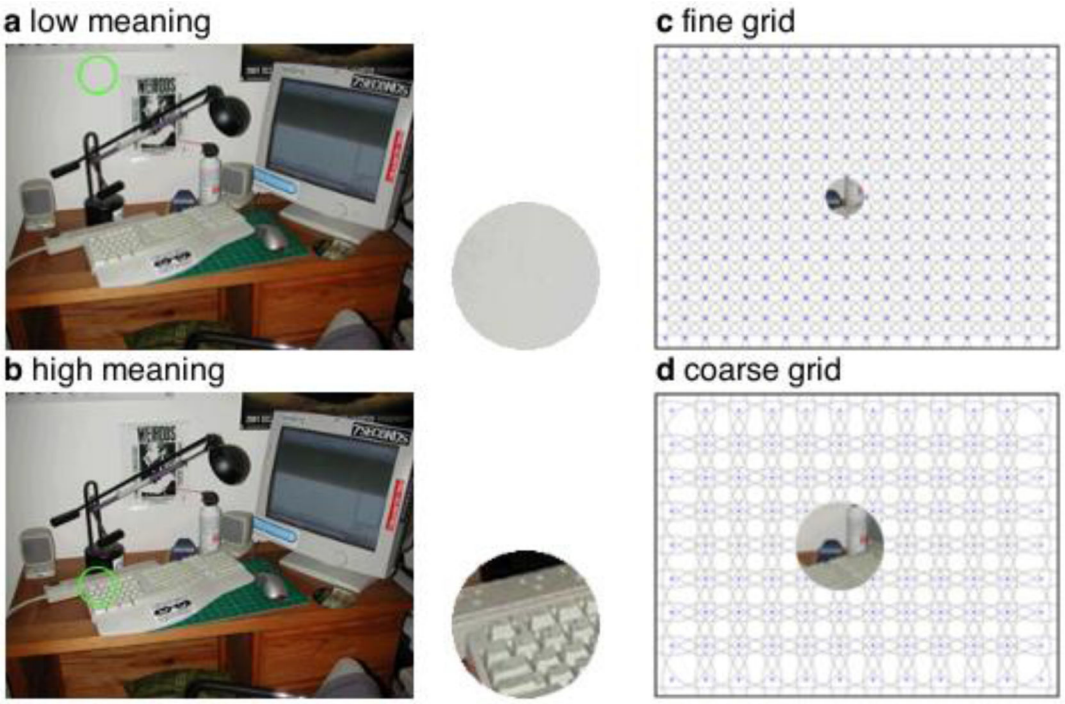

Figure 1. Rating Patch Examples.

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

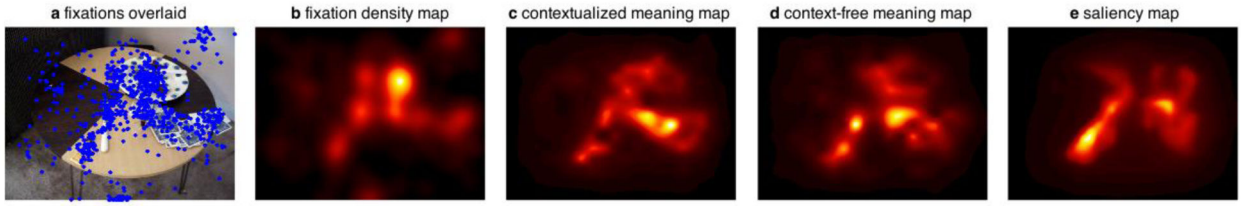

Figure 2. Map Examples.

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

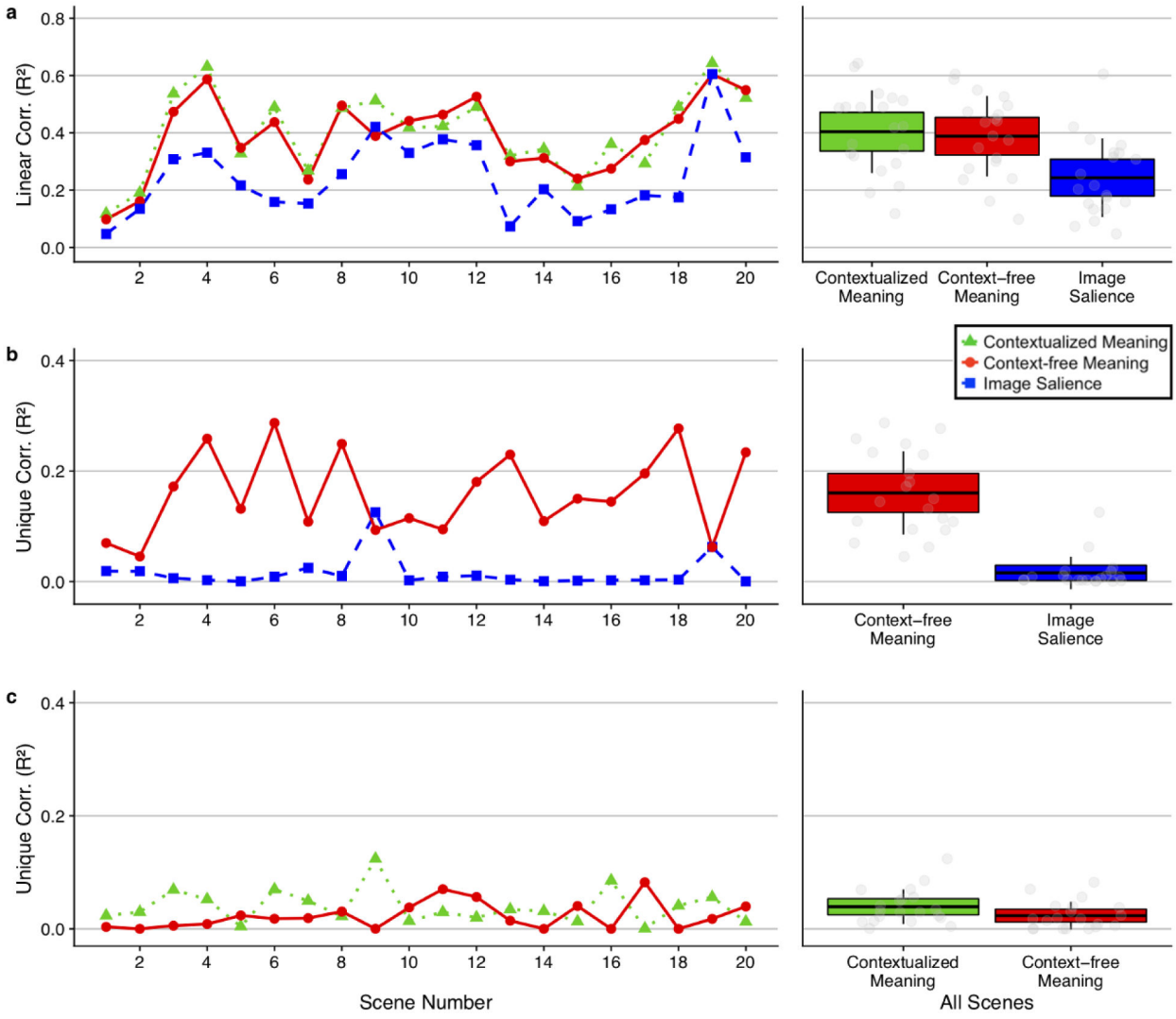

Figure 3. Squared Linear and Semi-Partial Correlations by Scene.

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

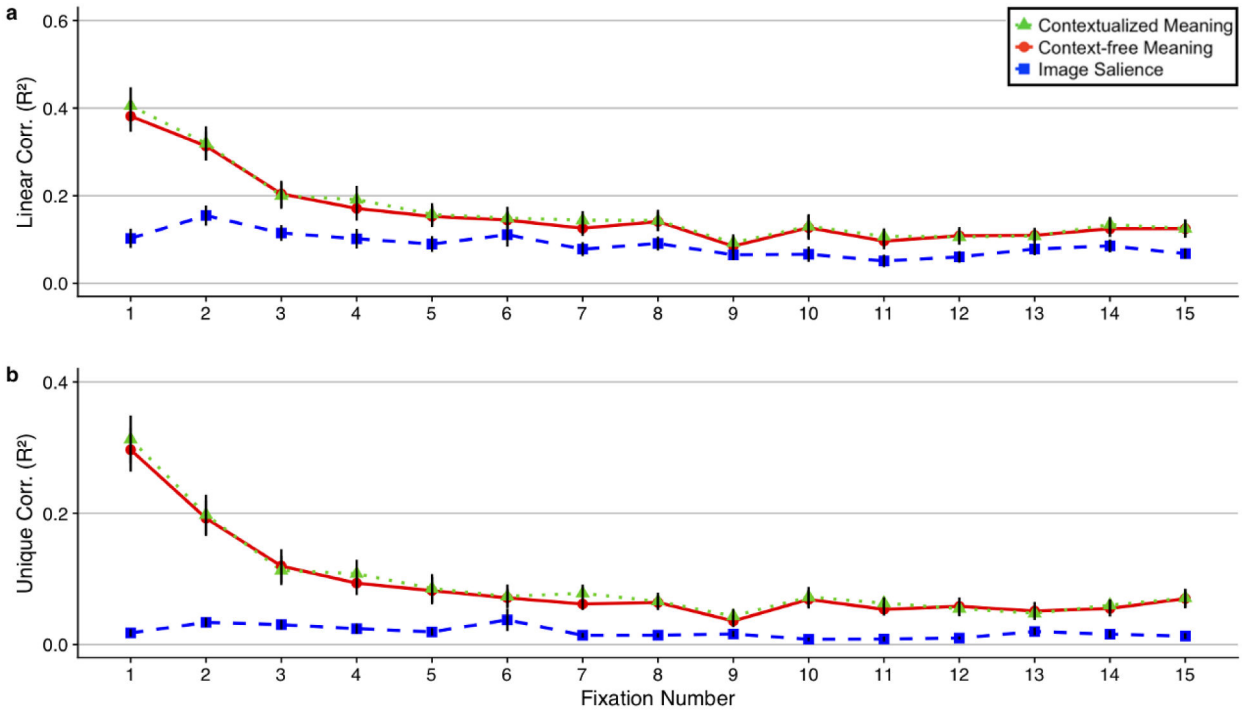

Figure 4. Ordinal Fixation Analysis.

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