Published in final edited form as:

J Cogn Neurosci. 2020 October ; 32(10): 2013–2023. doi:10.1162/jocn_a_01599.

Neural Correlates of Fixated Low- and High-level Scene Properties during Active Scene Viewing

John M. Henderson[1], **Jessica E. Goold**[1], **Wonil Choi**[2], **Taylor R. Hayes**[1]

[1]University of California, Davis

[2]Gwangju Institute of Science and Technology

Abstract

During real-world scene perception, viewers actively direct their attention through a scene in a controlled sequence of eye fixations. During each fixation, local scene properties are attended, analyzed, and interpreted. What is the relationship between fixated scene properties and neural activity in the visual cortex? Participants inspected photographs of real-world scenes in an MRI scanner while their eye movements were recorded. Fixation-related fMRI was used to measure activation as a function of lower- and higher-level scene properties at fixation, operationalized as edge density and meaning maps, respectively. We found that edge density at fixation was most associated with activation in early visual areas, whereas semantic content at fixation was most associated with activation along the ventral visual stream including core object and scene-selective areas (lateral occipital complex, parahippocampal place area, occipital place area, and retrosplenial cortex). The observed activation from semantic content was not accounted for by differences in edge density. The results are consistent with active vision models in which fixation gates detailed visual analysis for fixated scene regions, and this gating influences both lower and higher levels of scene analysis.

INTRODUCTION

Visual perception and visual cognition are active processes in which saccadic eye movements play a central role (Henderson, 2013; Rayner, 2009; Henderson & Ferreira, 2004; Findlay & Gilchrist, 2003; Yarbus, 1967; Buswell, 1935; Dodge, 1903). In natural perception, the eyes move from location to location to enable the acquisition of information as it is needed in real time (Hayhoe, 2017; Henderson, 2003, 2011; Rayner, 1998; Yarbus, 1967; Buswell, 1935). During active visual scene processing, the information acquired during each eye fixation is combined to produce a complete scene representation (Hollingworth, 2005; Hollingworth & Henderson, 2002). Indeed, what we recognize, understand, and remember about a scene is tightly tied to where we look (Hayhoe, 2017; Henderson, 2011). Current theoretical approaches and computational models of active scene viewing therefore attempt to account for how attention and, particularly, eye movements are related to scene variables associated with eye fixations (Kümmerer, Wallis, Gatys, &

Reprint requests should be sent to John M. Henderson, Center for Mind and Brain, 267 Cousteau Place, University of California, Davis, CA 95618, or via johnhenderson@ucdavis.edu.

Bethge, 2017; Nuthmann, Smith, Engbert, & Henderson, 2010; Torralba, Oliva, Castelhano, & Henderson, 2006; Itti & Koch, 2001; Koch & Ullman, 1985).

Real-world scenes contain a large amount of information about local scene elements and relationships that cannot be processed from a brief central glimpse (Fei-Fei, Iyer, Koch, & Perona, 2007; Hollingworth & Henderson, 2002; Henderson & Hollingworth, 1999b). To access this local information, viewers must actively direct attention to specific scene regions via eye movements. Previous eye-tracking studies have shown a tight link between fixation locations and cognitive processes (Hayhoe, 2017; Henderson, 2011; Rayner, 2009). However, the nature of the associations between fixations in scenes and activation in the cortical systems supporting visual analysis during active scene perception remains largely an open question (Malcolm, Groen, & Baker, 2016; Peelen & Kastner, 2014). To address this question, here, we tested the hypothesis that variation in scene content across fixations is related to variation in activation in visual areas of cortex. We specifically hypothesized that increases in lower-level visual feature content at fixation would be associated with greater activation in early visual areas, whereas increases in higher-level scene content at fixation would be associated with greater activation in higher-level visual areas including scene-related areas.

We tested these hypotheses by combining fMRI with high-resolution eye tracking in fixation-related (FIRE) fMRI. Specifically, we coregistered fMRI with eye tracking while participants silently and freely viewed photographs of real-world scenes (Figure 1A). We then employed FIRE fMRI analysis to measure neural activation associated with the content at each fixated location. FIRE fMRI has been shown to reveal underlying neural activity associated with eye fixations in reading (Schuster, Hawelka, Himmelstoss, Richlan, & Hutzler, 2020; Carter, Foster, Muncy, & Luke, 2019; Hsu, Clariana, Schloss, & Li, 2019; Desai, Choi, Lai, & Henderson, 2016; Henderson, Choi, Lowder, & Ferreira, 2016; Henderson, Choi, Luke, & Desai, 2015; Schuster, Hawelka, Richlan, Ludersdorfer, & Hutzler, 2015; Richlan et al., 2014) as well as object perception (Marsman, Renken, Haak, & Cornelissen, 2013; Marsman, Renken, Velichkovsky, Hooymans, & Cornelissen, 2012). However, little work to date has extended the method to active scene perception (Henderson & Choi, 2015).

To investigate these issues, we quantified the content of each fixated scene region at two levels of representation. For the analysis of low-level visual features, we examined in a whole-brain analysis the edge density at fixation (Figure 1D). We investigated edge density because edges are well known to be associated with activity in the visual cortex, and it has been suggested that they may also be related to activation in scene-related areas along the ventral visual stream in fMRI studies (Watson, Hymers, Hartley, & Andrews, 2016; Kauffmann, Ramanoël, Guyader, Chauvin, & Peyrin, 2015; Musel et al., 2013; Rajimehr, Devaney, Bilenko, Young, & Tootell, 2011; although see Henderson, 2011). Edge density is also related to fixation behavior in scenes (Henderson, Chanceaux, & Smith, 2009; Baddeley & Tatler, 2006; Mannan, Ruddock, & Wooding, 1996).

For the analysis of high-level properties of scene content, we capitalized on “meaning map” representations introduced by Henderson and Hayes (2017). Meaning maps represent the spatial distribution of semantic features across a scene (Figure 1C), with meaningful regions operationalized as those that are informative and recognizable (Antes, 1974; Mackworth & Morandi, 1967). To create meaning maps, we used crowdsourced responses given by large numbers of naive participants who rated the informativeness and recognizability of thousands of context-free scene patches. Meaning defined in this way is associated with the distribution of attention during free viewing of scenes (Peacock, Hayes & Henderson, 2019; Henderson & Hayes, 2017, 2018), visual search in scenes (Hayes & Henderson, 2019), memory for scenes (Bainbridge, Hall, & Baker, 2019), and linguistic descriptions of scenes (Ferreira & Rehrig, 2019; Henderson, Hayes, Rehrig, & Ferreira, 2018).

For our questions concerning higher-level scene properties, we conducted whole-brain analyses as well as conjunction analyses on four cortical areas related to object and scene perception chosen a priori based on previous studies: lateral occipital complex (LOC), parahippocampal place area (PPA), transverse occipital sulcus (TOS; also called occipital place area [OPA]), and retrosplenial cortex (RSC; also called medial place area; Epstein & Baker, 2019; Çukur, Huth, Nishimoto, & Gallant, 2016; Malcolm et al., 2016; MacEvoy & Epstein, 2011; Peelen, Fei-Fei, & Kastner, 2009; Walther, Caddigan, Fei-Fei, & Beck, 2009; Epstein & Higgins, 2007; Epstein & Kanwisher, 1998). If, in addition to previously demonstrated computations for global scene properties, scene-selective areas are also sensitive to higher-level properties of locally attended scene regions, then activation in scene-related areas should also increase systematically with increases in meaning across fixations.

A critical aspect of the FIRE fMRI method is that it provides information concerning voxels that change their activation with changes in features at fixation. For this reason, FIRE analysis does not identify regions that activate to the stimulus globally but instead reflects activation to locally fixated features beyond any activation produced by the presence of the global stimulus across all fixations. Because of this characteristic, FIRE fMRI analysis can provide information about which attended scene properties modulate activation beyond any overall activation produced globally by scenes. Importantly, when comparing different fixated properties, only the property values included in the FIRE analysis as regressors change from one analysis to another. The fixations themselves are identical across analyses, so all other aspects of viewing (e.g., the participant’s motivation, fatigue, time in the scanner) are controlled.

In summary, this study investigated neural activation associated with the properties of fixated scene regions when participants freely viewed photographs of real-world scenes via eye movements.

METHODS

Participants

Forty-three right-handed participants (14 men), aged 18–34 ($M = 21.48$) years, were recruited from the Columbia, South Carolina, community. They were all native speakers of English and reported normal or corrected-to-normal vision. All participants gave informed consent, were screened for MRI safety, and were given $10 per hour for participation in compliance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) and by approval of the University of South Carolina Institutional Review Board for human participants. Three participants were removed from analysis, one because of a technical problem with the scanner and the others because of inattention during the experiment, leaving 40 participants for the analysis.

Stimuli

Images were 40 full-color photographs depicting a wide variety (20 natural and 20 human constructed) of real-world scenes. The scenes did not include faces.

Apparatus

Images were presented using an Avotec Silent Vision 6011 projector in its native resolution (1024 × 768) at a refresh rate of 60 Hz. Eye movements were recorded via an SR Research Eyelink 1000 long-range MRI eye-tracker sampling at 1000 Hz. Viewing was binocular, and eye movements were recorded from one eye.

Procedure

Scenes were presented in two functional runs, with each run containing 10 natural and 10 human-constructed scenes presented in a random order. Scenes were shown individually, and participants were instructed to view them silently. Each scene was presented for 12 sec, with a 6-sec central fixation marker on a gray screen between scenes. Each run lasted about 6 min. These runs were presented with separate runs containing text that were not relevant for this study.

Eye-movement Data Acquisition—A 13-point calibration procedure was implemented in the scanner before each functional run 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°. A central fixation marker was presented on the screen during the 6-sec interval between each trial, and participants were instructed to fixate that marker. Eye movements were recorded throughout the runs.

MRI Data Acquisition—MR data were collected on a Siemens Medical Systems 3-T Trio. A 3-D T1-weighted MPRAGE radio frequency–spoiled rapid flash scan in the sagittal plane and a T2/PD-weighted multislice axial 2-D dual fast/turbo spin-echo scan in the axial plane were used. The multiecho whole-brain T1 scans had a 1-mm isotropic voxel size (repetition time [TR] = 2530 msec, flip angle = 7°). Functional runs were acquired using gradient-echo EPI images with TR = 1850 msec, echo time = 30 msec, flip angle = 75°, field of view = 208 mm, and matrix = 64 × 64. Volumes consisted of thirty-four 3-mm axial slices, resulting in a 3.3 × 3.3 × 3 mm voxel size.

Eye Movement and fMRI Coregistration—The fMRI and eye-tracking data were synchronized so that fixation onset from the eye tracker could be aligned with the fMRI data. This was accomplished by aligning the onset of the trial run with the onset of the functional scan. Times of experiment onset, block onset, and fixation onset were saved in the eye-movement record by Experiment Builder (SR Research). Timing was obtained by recording a transistor–transistor logic pulse from the scanner to the experimental control computer running Experiment Builder, making it possible to coregister eye movement and fMRI events for later analysis.

fMRI Analysis—The AFNI software package (Cox, 1996) was used for image analysis. Within-participant analysis involved slice timing correction, spatial coregistration (Cox & Jesmanowicz, 1999), and registration of functional images to the anatomy (Saad et al., 2009). Voxel-wise multiple linear regression was performed with the program *3dREMLfit*, using reference functions representing each condition convolved with a standard hemodynamic response function. Reference functions representing the six motion parameters were included as covariates of no interest. In addition, the signals extracted from cerebrospinal fluid and white matter (segmented using 3dSeg) were also included as noise covariates of no interest.

To examine the effects of fixated scene features, an amplitude-modulated (parametric) regressor was used that contained the onset times (from the onset of each run) of each fixation and the feature values (edge density and meaning map value) at the fixated location. There are multiple fixations within each TR. We take advantage of the fact that the timings of the fixations within each TR, as well as the feature values of the fixated locations within each TR, vary from TR to TR. This variation, combined with the large number of TRs, provides sufficient power to extract information from the low-temporal-resolution fMRI data based on the high-temporal-resolution eye-tracking data. The ideal hemodynamic response resulting from this regressor was subsampled to match the time resolution of EPI images. A binary regressor coding the onset of all fixations was used.

The individual statistical maps and the anatomical scans were projected into standard stereotaxic space (Talairach & Tournoux, 1988) and smoothed with a Gaussian filter of 5-mm FWHM. In the random effects analyses, group maps were created by comparing activations against a constant value of 0. The group maps were thresholded at voxel-wise $p < .001$ and corrected for multiple comparisons by removing clusters with a below-threshold size to achieve a map-wise corrected alpha < .05. Using the (recently updated) program 3dREMLfit with 1000 iterations, the cluster threshold was determined through Monte Carlo simulations that estimate the chance probability of spatially contiguous voxels exceeding the voxel-wise *p* threshold, that is, of false-positive noise clusters. The analysis was restricted to a mask that excluded areas outside the brain as well as deep white matter areas and the ventricles. Data and analysis scripts related to these analyses are available from the authors upon reasonable request.

Eye-Movement Analysis—Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s^2). Fixation was defined as pauses between saccades of 50–1500 msec that were not part of a blink. Eye-movement data were imported into MATLAB (The MathWorks) using the EDF converter tool. The first fixation, located at the center of the display because of the 6-sec blank period, was eliminated from analysis. Basic eye-movement measures were examined to ensure that participants were moving their eyes naturally in the scanner. Data analysis was based on 52,246 fixations across participants and scenes, with a mean of 33 fixations per scene (*SD*

= 6.98 fixations), a mean fixation duration of 318 msec (*SD* = 69 msec), and a mean saccade amplitude of 2.84° (*SD* = 0.66°). Eye-movement measures were comparable to those typically obtained for similar scenes viewed outside the scanner (Cronin, Hall, Goold, Hayes, & Henderson, 2020; Castelhano, Mack, & Henderson, 2009).

Scene Feature Definitions

Low-level Content: Edge Density—Edge density is a measure of the edges present in an image. We calculated edges using the Canny edge detection algorithm in MATLAB, which returns a value of 1 for an edge and 0 otherwise, using parameter settings for fine edges with low and high thresholds of 0.10 and 0.27, respectively, and sigma = 1. For each fixated region, edge density was defined as the total count of edge pixels within that region (Henderson et al., 2009).

High-level Semantic Content: Meaning Maps—Local scene meaning was represented by meaning maps (Henderson & Hayes, 2017). Meaning maps capture the spatial distribution of semantic features in scenes. To generate meaning maps, each scene photograph was decomposed into a series of highly overlapping, tiled circular patches at fine and course spatial scales (Figure 2). The two scales and numbers of patches were chosen based on simulations showing that ground-truth visual properties of scenes can be recovered from them (Henderson & Hayes, 2017). Patches were rated by workers on Amazon Mechanical Turk (MTurk). MTurk workers each rated a randomly selected subset of individually presented patches taken from the set of scenes to be rated. MTurk workers were recruited from the United States, had a hit approval rate of 99% and 500 hits approved, were only allowed to participate in the study once, and were paid $0.50 cents per assignment. All workers provided informed consent. Each worker rated 300 random patches. Workers were instructed to assess the meaningfulness of each patch based on how informative or recognizable it was. They were first given two low-meaning and two high-meaning scene patches as examples and then rated the meaningfulness of scene patches on a 6-point Likert scale from “very low” to “very high.” Patches were presented in a random order and without scene context. Each unique patch was rated three times by three independent raters for a total of 48,960 ratings. Because of the high degree of overlap across patches, each fine patch contained rating information from 27 independent raters, and each coarse patch contained rating information from 63 independent raters. Meaning maps for each scene were generated by averaging ratings by pixel over patches and raters and smoothing the results to create fine and coarse maps, averaging those maps, and smoothing the resulting maps using a Gaussian kernel (Figure 1C). More details can be found in previous work (Henderson & Hayes, 2017, 2018). For the FIRE fMRI analysis, meaning was defined as the average meaning map value within a 1° circle centered at each fixation location corresponding to the scene region falling on the fovea.

RESULTS

Of primary interest was FIRE activation that increased as lower- and higher-level scene content increased at attended locations. These analyses also included fixation onset along with parametric regressors for fixation content. Fixation onset accounted for the average level of activation across all fixations and served as a measure of global activation from the continuously present scene. Onset activation was not of particular interest here, although it served as the baseline (intercept) against which to assess the edge density and semantic content regressors.

Beginning with edge density, the results of the parametric edge density regressor showed that the density of edges at fixation was strongly associated with activation in occipital visual areas (Figure 3, Table 1). Additional activation was seen bilaterally in the cingulate cortex and, in the right hemisphere, in the supramarginal gyrus/inferior parietal lobe and in the frontal gyrus. No activation was observed in the inferior temporal lobes along the ventral visual stream.

In contrast to edge density, the results of the parametric meaning regressor showed that meaning map value at fixation was positively associated with activation further along the ventral visual stream (Figure 4A, Table 2), including the bilateral superior, middle, and inferior occipital gyri; precuneus; lingual gyrus; middle temporal gyrus; parahippocampal gyrus; and fusiform gyrus as well as right angular gyrus. Bilateral activation was also observed frontally in the superior, middle, and inferior frontal gyri and insula, with additional frontal activation in the left precentral and left medial frontal gyri.

It could be that more semantically informative scene regions are also more visually complex. To examine the influence of meaning at attended locations while controlling for visual complexity, in a second analysis, we used a partial regression approach in which edge density was included as a parametric regressor along with meaning. This analysis allowed us to look for activation uniquely related to meaning at fixated locations while statistically controlling for edges at the same locations. Consistent with the main analysis, meaning remained positively associated with activation in the ventral visual stream when controlling for edges (Figure 4B, Table 3). Notably, these partial regression results were replicated using a variety of other image features as controls not reported here, including high-pass spatial frequencies, image entropy, and clutter. Together, these results suggest that much of the activation observed along the ventral stream was produced by higher-level features at fixation and not by simple visual features.

To investigate whether the activation related to meaning along the ventral stream included cortical regions previously associated with real-world scenes, we generated functional regions from a meta-analysis of previous studies using the 2019 release of Neurosynth (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). Specifically, we identified four regions related to scene processing based on the published literature: LOC, OPA/TOS, PPA, and RSC. We used Neurosynth to generate activation masks for these regions and then examined the degree to which activation associated with meaning controlling for edge density overlapped with each region. Results showed clear activation related to fixated semantic features in all four scene regions (Figure 5). Although not reported here, this pattern of results did not change when high-spatial-frequency content, image entropy, and image clutter were used instead of edge density as image feature controls.

In summary, the results showed that activation in the occipital cortex was associated with edge density at the fixated location. In comparison, activation along much of the ventral visual stream, and importantly within core areas of the scene processing network, was associated with moment-to-moment changes in fixated semantic content as assessed by meaning maps. Importantly, this latter activation was not explained by variation in edge density.

DISCUSSION

Natural perception of real-world scenes (and, indeed, of any complex visual stimulus) requires that the eyes be oriented to important local regions of the image so that those regions can be perceived, recognized, understood, and remembered (Hayhoe, 2017; Henderson, 2003, 2011; Rayner, 1998; Yarbus, 1967; Buswell, 1935). Scene representations are built up incrementally as fixation is moved from region to region through the scene and local detail is added to the scene representation (Henderson, 2017; Hollingworth, 2005; Henderson & Hollingworth, 1999b). How complete scene representations are generated by the brain during active visual perception in which viewers freely select regions for fixation and attention is largely unknown. In neuroimaging studies of scene perception, scenes are often presented for a brief period and/or viewers are asked to maintain fixation at a central location. Although these studies provide critical information concerning how global scene representations are generated by the brain, they do not speak to processes related to moment-to-moment changes in features at fixation as the eyes move through a scene during active viewing.

To investigate this question, we measured the relationship between variation in the content of fixated scene regions and activation in the ventral visual stream. We specifically tested the hypothesis that neural activity is associated with increases in content at each fixation position. Consistent with this hypothesis, the results showed that activation along much of the visual stream increased with moment-to-moment changes in fixated scene content, with activation in early visual areas reflecting fixated edge density and, in later areas, reflecting higher-level properties related to semantic content. Importantly, the latter activity was observed in core areas of the scene-specific cortical network.

During active scene viewing, although local features change from fixation to fixation, the global characteristics of the scene do not. If regions along the visual stream were only sensitive to global visual features or global scene semantics, then there would be no reason for activation in these regions to vary with fixation location. Instead, we observed that activation in both early and later cortical visual regions was sensitive to the content present in each fixation. With regard to higher visual areas along the ventral stream, the results provide evidence that, in addition to scene gist and global scene properties and structure, scene-selective cortical regions are involved in processing local high-level properties at fixation as the eyes naturally traverse a scene.

The finding that the activation in higher-order visual areas along the ventral stream was associated with ratings of meaning implemented as meaning maps is consistent with the general hypothesis that the ventral stream codes for both high-level visual and conceptual scene properties (Epstein & Baker, 2019; Devereux, Clarke, & Tyler, 2018; Martin, Douglas, Newsome, Man, & Barense, 2018; Bonner, Price, Peelle, & Grossman, 2016). In future studies, it will be important to unpack what the meaning maps capture. For example, ratings of meaning may be based on a number of semantic dimensions that could be separated, and this type of analysis may lead to a more fine-grained assessment of the computations performed during fixations by specific cortical regions. Meaning ratings may also be related to the presence of objects, although the finding that the meaning-based effects persisted when controlling for edge density rules out an explanation based on a simple definition of visual object as a visually complex region.

The results demonstrate that FIRE fMRI can provide an additional source of evidence for extending and testing neural theories of scene processing, with an emphasis specifically on natural active viewing in which people freely move their eyes as they incrementally acquire information. In addition to providing a basis for investigating incremental scene processing in the brain, the FIRE fMRI approach provides two other key advantages. First, participants perform no secondary behavioral task during scanning; instead, they simply view scene photographs freely. Therefore, any observed effects cannot be attributed to task-specific response processes. Moreover, the obtained eye-movement data allow us to verify that participants are viewing scenes naturally and attentively although they are not required to perform any sort of behavioral task. On the basis of the present data as well as our previous work, we can confirm that eye movements for scene viewing while in an MRI scanner are similar to those observed outside the scanner (Choi & Henderson, 2015; Henderson et al., 2015; Choi, Desai, & Henderson, 2014).

Second, because participants view complex scenes that vary in their local characteristics over space, each fixation provides a data point, resulting in large amounts of data for each scene and each participant. The result is substantial statistical power. Importantly, because the same set of fixations contributes to data analysis for different regressors (in this case, edge density and meaning), all aspects of the stimulus, the participant’s motivation, level of fatigue, time in the run, practice, and so forth are controlled. That is, the BOLD data are the same in the regressor comparisons, and all that changes across comparisons is the regressor that is considered. This aspect of FIRE fMRI analysis provides an important advantage for comparing the influences of different scene properties in active vision.

At a general level, this study was motivated by an interest in understanding the relationship between overt attention and visual processing in the brain under naturalistic viewing conditions (Choi & Henderson, 2015; Henderson & Choi, 2015). This approach has become of interest across a number of other domains such as natural reading (Carter et al., 2019; Desai et al., 2016; Schuster et al., 2015; Altmann, Bohrn, Lubrich, Menninghaus, & Jacobs, 2014; Choi et al., 2014), event understanding (Aly, Chen, Turk-Browne, & Hasson, 2018; Baldassano, Hasson, & Norman, 2018), and auditory narrative comprehension (Brennan, 2016; Hale, Lutz, Luh, & Brennan, 2015). The research is in the spirit of recent calls for understanding both neural activity and neural models in the context of natural behavior (Kriegeskorte & Douglas, 2018; Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Yamins & DiCarlo, 2016). As Kriegeskorte and Douglas (2018) argue, the challenge is to build models of brain information processing that are consistent with both the types of complex cognitive tasks that the brain must support and the brain structures and functions needed to implement those processes. Our work specifically focuses on these issues in the context of active vision for naturalistic visual scenes. Overall, this study sets the stage for using FIRE fMRI to investigate a host of theoretical issues related to the neurocognition of natural active scene viewing.

In the present work, we specifically focused on real-world scenes because they include many aspects of the visual world that are likely highly functional in the control of attention but that do not exist in simple stimuli. These aspects include visual complexity, physical regularity and physical constraint, and semantic content. Furthermore, we focused on active vision in which viewers are allowed to orient their attention naturally over extended time and space with eye movements because these are the conditions under which attention normally operates. This approach contrasts with studies that investigate covert attention over briefly presented stimuli. Although those studies are clearly important and have established critical findings, the best way to determine whether the principles generated from them will scale up to natural active vision is to investigate that scalability directly.

The present results have important implications for understanding the cortical basis of scene perception. We do not fully perceive and understand a scene from only an initial brief glimpse. Instead, the first glimpse delivers general semantic and spatial gist (Greene & Oliva, 2009; Fei-Fei et al., 2007; Potter, 1975; Biederman, 1972). The gist provides a context within which to perceive and integrate information derived from attentively sampling local scene regions. This sampling allows incremental generation of a more complete scene representation in which details are filled in over time (Henderson, 2011; Hollingworth, 2005; Henderson & Hollingworth, 1999a). Complete understanding of a scene requires selecting important local scene regions for closer perceptual and conceptual analysis via shifts of overt attention (Hayhoe, 2017; Henderson, 2003, 2011; Rayner, 1998; Yarbus, 1967; Buswell, 1935). The present results suggest that cortical areas previously found to be scene-selective play a critical role in these incremental computations. A complete theory of the cortical processing of scenes will require an account of how initial global scene gist and incremental local scene details are combined into a unified representation as well as how the scene perception network supports this process.

Acknowledgments

This research was supported by the National Eye Institute of the National Institutes of Health under award number R01EY027792. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

REFERENCES

Altmann U, Bohrn IC, Lubrich O, Menninghaus W, & Jacobs AM (2014). Fact vs fiction—How paratextual information shapes our reading processes. Social Cognitive and Affective Neuroscience, P, 22–29. [PubMed: 22956671]

Aly M, Chen J, Turk-Browne NB, & Hasson U (2018). Learning naturalistic temporal structure in the posterior medial network. journal of Cognitive Neuroscience, 30, 1345–1365. [PubMed: 30004848]

Antes JR (1974). The time course of picture viewing. journal of Experimental Psychology, 103, 62–70. [PubMed: 4424680]

Baddeley RJ, & Tatler BW (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. Vision Research, 46, 2824–2833. [PubMed: 16647742]

Bainbridge WA, Hall EH, & Baker CI (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. Nature Communications, 10, 5.

Baldassano C, Hasson U, & Norman KA (2018). Representation of real-world event schemas during narrative perception. journal of Neuroscience, 38, 9689–9699. [PubMed: 30249790]

Biederman I (1972). Perceiving real-world scenes. Science, 177, 77–80. [PubMed: 5041781]

Bonner MF, Price AR, Peelle JE, & Grossman M (2016). Semantics of the visual environment encoded in parahippocampal cortex. journal of Cognitive Neuroscience, 28, 361–378. [PubMed: 26679216]

Brennan J. (2016). Naturalistic sentence comprehension in the brain. Language and Linguistics Compass, 10, 299–313.

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

Carter BT, Foster B, Muncy NM, & Luke SG (2019). Linguistic networks associated with lexical, semantic and syntactic predictability in reading: A fixation-related fMRI study. Neuroimage, 18P, 224–240.

Castelhano MS, Mack ML, & Henderson JM (2009). Viewing task influences eye movement control during active scene perception. journal of Vision, 9, 6. [PubMed: 19761321]

Choi W, Desai RH, & Henderson JM (2014). The neural substrates of natural reading: A comparison of normal and nonword text using eyetracking and fMRI. Frontiers in Human Neuroscience, 8, 1024. [PubMed: 25566039]

Choi W, & Henderson JM (2015). Neural correlates of active vision: An fMRI comparison of natural reading and scene viewing. Neuropsychologia, 75, 109–118. [PubMed: 26026255]

Cox RW (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research, 29, 162–173. [PubMed: 8812068]

Cox RW, & Jesmanowicz A. (1999). Real-time 3D image registration for functional MRI. Magnetic Resonance in Medicine, 42, 1014–1018. [PubMed: 10571921]

Cronin DA, Hall EH, Goold JE, Hayes TR, & Henderson JM (2020). Eye movements in real-world scene photographs: General characteristics and effects of viewing task. Frontiers in Psychology, 10, 2915. [PubMed: 32010016]

Çukur T, Huth AG, Nishimoto S, & Gallant JL (2016). Functional subdomains within scene-selective cortex: Parahippocampal place area, retrosplenial complex, and occipital place area. journal of Neuroscience, 36, 10257–10273. [PubMed: 27707964]

Desai RH, Choi W, Lai VT, & Henderson JM (2016). Toward semantics in the wild: Activation to manipulable nouns in naturalistic reading. journal of Neuroscience, 36, 4050–4055. [PubMed: 27053211]

Devereux BJ, Clarke A, & Tyler LK (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Scientific Reports, 8, 10636. [PubMed: 30006530]

Dodge R. (1903). Five types of eye movements in the horizontal meridian plane of the field of regard. American journal of Physiology, 8, 307–329.

Epstein RA, & Baker CI (2019). Scene perception in the human brain. Annual Review of Vision Science, 5, 373–397.

Epstein RA, & Higgins JS (2007). Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. Cerebral Cortex, 17, 1680–1693. [PubMed: 16997905]

Epstein RA, & Kanwisher N. (1998). A cortical representation of the local visual environment. Nature, 3P2, 598–601.

Fei-Fei L, Iyer A, Koch C, & Perona P. (2007). What do we perceive in a glance of a real-world scene? journal of Vision, 7, 10.

Ferreira F, & Rehrig G. (2019). Linearisation during language production: Evidence from scene meaning and saliency maps. Language, Cognition and Neuroscience, 34, 1129–1139.

Findlay JM, & Gilchrist ID (2003). Active vision: The psychology of looking and seeing. Oxford: Oxford University Press.

Greene MR, & Oliva A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cognitive Psychology, 58, 137–176. [PubMed: 18762289]

Hale JT, Lutz DE, Luh W-M, & Brennan JR (2015). Modeling fMRI time courses with linguistic structure at various grain sizes. In Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics (pp. 89–97). Denver, CO: Association for Computational Linguistics.

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

Hayhoe MM (2017). Vision and action. Annual Review of Vision Science, 3, 389–413.

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

Henderson JM (2011). Eye movements and scene perception. In Liversedge SP, Gilchrist ID, & Everling S. (Eds.), The Oxford handbook of eye movements (pp. 593–606). Oxford: Oxford University Press.

Henderson JM (2013). Eye movements. In Reisberg D. (Ed.), The Oxford handbook of cognitive psychology. New York: Oxford University Press.

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

Henderson JM, Chanceaux M, & Smith TJ (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. journal of Vision, 9, 32.

Henderson JM, & Choi W. (2015). Neural correlates of fixation duration during real-world scene viewing: Evidence from fixation-related (FIRE) fMRI. journal of Cognitive Neuroscience, 27, 1137–1145. [PubMed: 25436668]

Henderson JM, Choi W, Lowder MW, & Ferreira F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. Neuroimage, 132, 293–300. [PubMed: 26908322]

Henderson JM, Choi W, Luke SG, & Desai RH (2015). Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. Neuroimage, 119, 390–397. [PubMed: 26151101]

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

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

Henderson JM, & Hayes TR (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. journal of Vision, 18, 10.

Henderson JM, Hayes TR, Rehrig G, & Ferreira F. (2018). Meaning guides attention during real-world scene description. Scientific Reports, 8, 13504. [PubMed: 30202075]

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

Henderson JM, & Hollingworth A. (1999b). The role of fixation position in detecting scene changes across saccades. Psychological Science, 10, 438–443.

Hollingworth A. (2005). The relationship between online visual representation of a scene and long-term scene memory. journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 396–411. [PubMed: 15910127]

Hollingworth A, & Henderson JM (2002). Accurate visual memory for previously attended objects in natural scenes. journal of Experimental Psychology: Human Perception and Performance, 28, 113–136.

Hsu C-T, Clariana R, Schloss B, & Li P. (2019). Neurocognitive signatures of naturalistic reading of scientific texts: A fixation-related fMRI study. Scientific Reports, P, 10678. [PubMed: 31337859]

Itti L, & Koch C. (2001). Computational modelling of visual attention. Nature Reviews Neuroscience, 2, 194–203. [PubMed: 11256080]

Kauffmann L, Ramanoël S, Guyader N, Chauvin A, & Peyrin C. (2015). Spatial frequency processing in scene-selective cortical regions. Neuroimage, 112, 86–95. [PubMed: 25754068]

Koch C, & Ullman S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology, 4, 219–227. [PubMed: 3836989]

Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, & Poeppel D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. Neuron, 93, 480–490. [PubMed: 28182904]

Kriegeskorte N, & Douglas PK (2018). Cognitive computational neuroscience. Nature Neuroscience, 21, 1148–1160. [PubMed: 30127428]

Kümmerer M, Wallis TSA, Gatys LA, & Bethge M. (2017). Understanding low- and high-level contributions to fixation prediction. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 4799–4808). Venice, Italy: IEEE.

MacEvoy SP, & Epstein RA (2011). Constructing scenes from objects in human occipitotemporal cortex. Nature Neuroscience, 14, 1323–1329. [PubMed: 21892156]

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

Malcolm GL, Groen IIA, & Baker CI (2016). Making sense of real-world scenes. Trends in Cognitive Sciences, 20, 843–856. [PubMed: 27769727]

Mannan SK, Ruddock KH, & Wooding DS (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. Spatial Vision, 10, 165–188. [PubMed: 9061830]

Marsman JBC, Renken R, Haak KV, & Cornelissen FW (2013). Linking cortical visual processing to viewing behavior using fMRI. Frontiers in Systems Neuroscience, 7, 109. [PubMed: 24385955]

Marsman JBC, Renken R, Velichkovsky BM, Hooymans JMM, & Cornelissen FW (2012). Fixation based event-related fMRI analysis: Using eye fixations as events in functional magnetic resonance imaging to reveal cortical processing during the free exploration of visual images. Human Brain Mapping, 33, 307–318. [PubMed: 21472819]

Martin CB, Douglas D, Newsome RN, Man LLY, & Barense MD (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. eLife, 7, e31873. [PubMed: 29393853]

Musel B, Bordier C, Dojat M, Pichat C, Chokron S, Le Bas J-F, et al. (2013). Retinotopic and lateralized processing of spatial frequencies in human visual cortex during scene categorization. journal of Cognitive Neuroscience, 25, 1315–1331. [PubMed: 23574583]

Nuthmann A, Smith TJ, Engbert R, & Henderson JM (2010). CRISP: A computational model of fixation durations in scene viewing. Psychological Review, 117, 382–405. [PubMed: 20438231]

Peacock CE, Hayes TR, & Henderson JM (2019). Meaning guides attention during scene viewing, even when it is irrelevant. Attention, Perception, G Psychophysics, 81, 20–34.

Peelen MV, Fei-Fei L, & Kastner S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature, 460, 94–97. [PubMed: 19506558]

Peelen MV, & Kastner S. (2014). Attention in the real world: Toward understanding its neural basis. Trends in Cognitive Sciences, 18, 242–250. [PubMed: 24630872]

Potter MC (1975). Meaning in visual search. Science, 187, 965–966. [PubMed: 1145183]

Rajimehr R, Devaney KJ, Bilenko NY, Young JC, & Tootell RBH (2011). The “parahippocampal place area” responds preferentially to high spatial frequencies in humans and monkeys. PLoS Biology, P, e1000608.

Rayner K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372–422. [PubMed: 9849112]

Rayner K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. Quarterly journal of Experimental Psychology, 62, 1457–1506.

Richlan F, Gagl B, Hawelka S, Braun M, Schurz M, Kronbichler M, et al. (2014). Fixation-related fMRI analysis in the domain of reading research: Using self-paced eye movements as markers for hemodynamic brain responses during visual letter string processing. Cerebral Cortex, 24, 2647–2656. [PubMed: 23645718]

Saad ZS, Glen DR, Chen G, Beauchamp MS, Desai R, & Cox RW (2009). A new method for improving functional-to-structural MRI alignment using local Pearson correlation. Neuroimage, 44, 839–848. [PubMed: 18976717]

Schuster S, Hawelka S, Himmelstoss NA, Richlan F, & Hutzler F. (2020). The neural correlates of word position and lexical predictability during sentence reading: Evidence from fixation-related fMRI. Language, Cognition and Neuroscience, 35, 613–624.

Schuster S, Hawelka S, Richlan F, Ludersdorfer P, & Hutzler F. (2015). Eyes on words: A fixation-related fMRI study of the left occipito-temporal cortex during self-paced silent reading of words and pseudowords. Scientific Reports, 5, 12686. [PubMed: 26235228]

Talairach J, & Tournoux P. (1988). Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: An approach to cerebral imaging. New York: Thieme.

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, 766–786. [PubMed: 17014302]

Walther DB, Caddigan E, Fei-Fei L, & Beck DM (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. journal of Neuroscience, 2P, 10573–10581.

Watson DM, Hymers M, Hartley T, & Andrews TJ (2016). Patterns of neural response in scene-selective regions of the human brain are affected by low-level manipulations of spatial frequency. Neuroimage, 124, 107–117. [PubMed: 26341028]

Yamins DLK, & DiCarlo JJ (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience, 19, 356–365. [PubMed: 26906502]

Yarbus AL (1967). Eye movements and vision. New York: Plenum.

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, & Wager TD (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature Methods, 8, 665–670. [PubMed: 21706013]

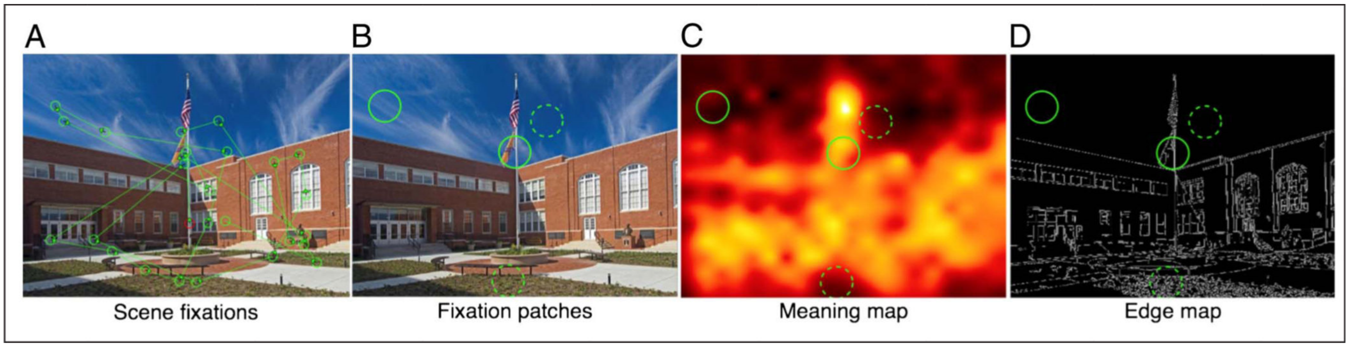

Figure 1.

Example scene with eye movements and analysis maps. (A) One participant’s eye movements. (B) Regions of analysis around four example fixation locations. (C) The four regions within the scene’s meaning map. (D) The same four regions within the scene’s edge map. In B and C, the solid green circles show a high- and low-meaning fixation. In B and D, the dotted green circles show a high- and low-edge-density region.

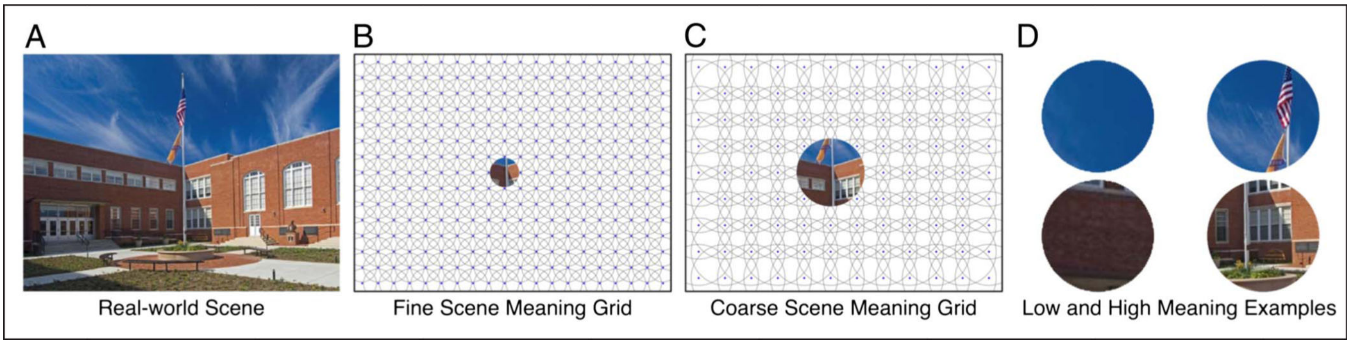

Figure 2.

Meaning maps. (A) A real-world scene. (B) Fine scale patches from the patch grids. (C) Coarse scale patches from the patch grids. (D) Examples of patches receiving low (left column) and high (right column) average meaning ratings.

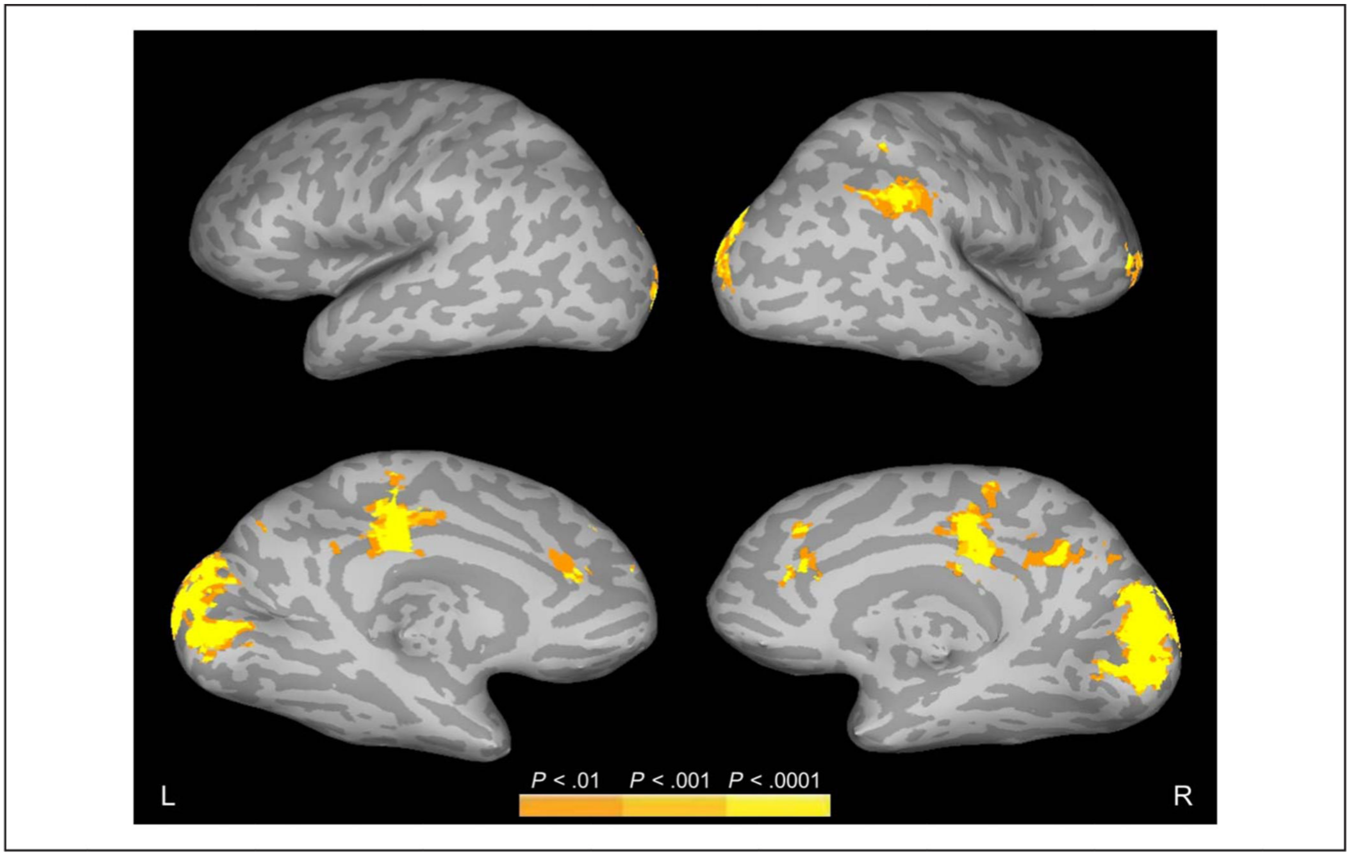

Figure 3.

Regions showing increased FIRE fMRI activity to increased edge density at fixation during active scene viewing. Data are shown on inflated brain images, with gyri shown as light gray and sulci shown as dark gray. The top and bottom rows show the lateral and medial views, respectively, of the left (L) and right (R) hemispheres.

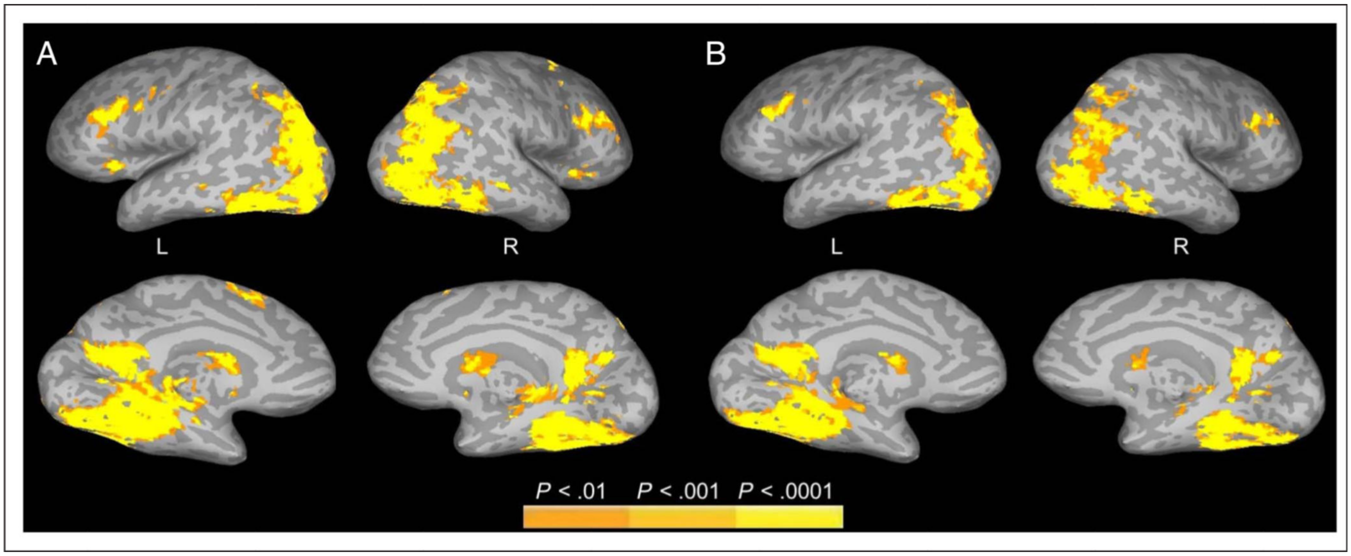

Figure 4.

Regions showing increased FIRE fMRI activity to increased meaning at fixation during active scene viewing. (A) FIRE activation associated with meaning. (B) FIRE activation associated with meaning controlling for edge density. Data are shown on inflated brain images, with gyri shown as light gray and sulci shown as dark gray. The top and bottom rows show the lateral and medial views, respectively, of the left (L) and right (R) hemispheres.

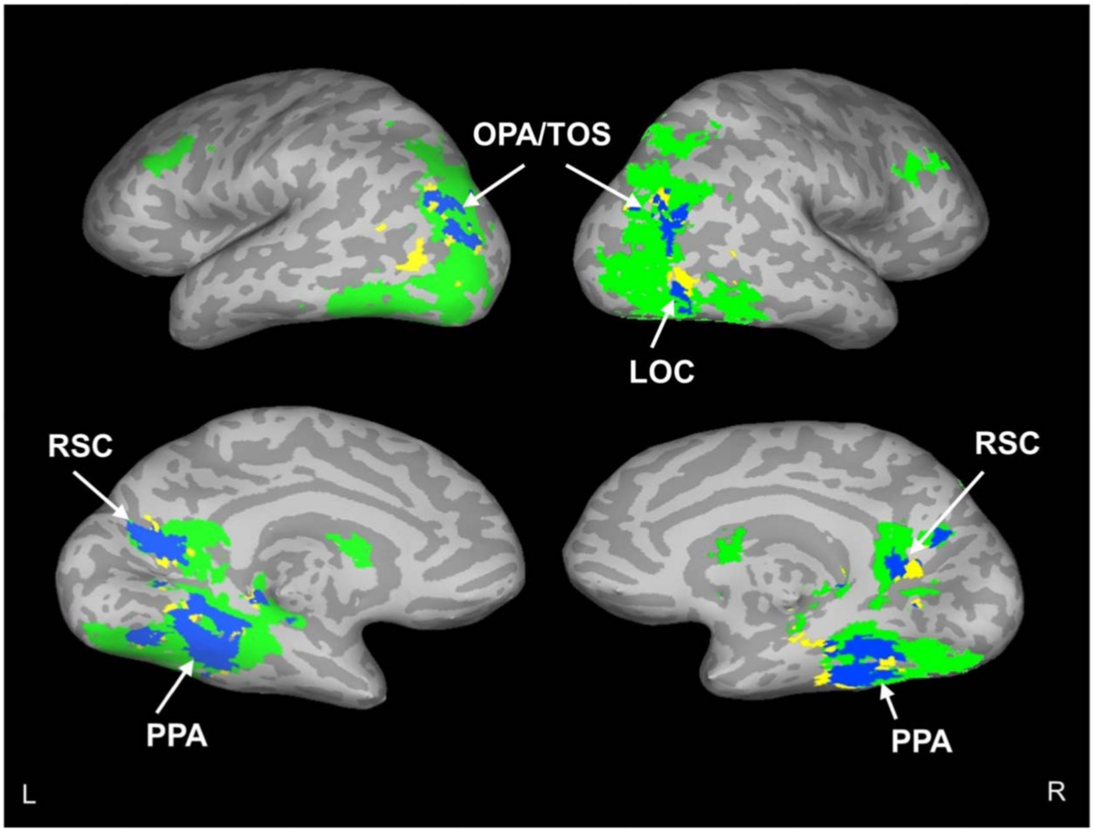

Figure 5.

FIRE fMRI activation in scene-related areas to increased meaning. Overlap (blue) of regions showing increased FIRE fMRI activity to increased local meaning (green; controlled for edges) and scene regions identified by Neurosynth (yellow). Labeled regions are PPA, LOC, RSC, and OPA/TOS. Data are shown on inflated brain images, with gyri shown as light gray and sulci shown as dark gray. The top and bottom rows show the lateral and medial views, respectively, of the left (L) and right (R) hemispheres.

Table 1.

Regions Showing Increased FIRE fMRI Activity to Increased Edge Density at Fixation during Active Scene Viewing

Volume	Max	x	y	z	Anatomical Structure
19251	6.573	4	−85	−3	R/L cuneus, R/L lingual gyrus
3645	5.222	52	−52	38	R supramarginal gyrus, R inferior parietal lobule
3321	5.165	−7	−25	38	R/L cingulate gyrus, R paracentral lobule
1296	4.689	−1	34	14	R/L anterior cingulate
891	5.059	10	−52	29	R cingulate gyrus, R precuneus
864	4.919	46	40	8	R middle frontal gyrus, R inferior frontal gyrus

Locations of peak activation are shown for each cluster with significant activity (Family-wise error [FWE] corrected at alpha < .05). Multiple peaks required separation by a minimum of 25 voxels. The volume of the cluster (μL), peak *z* score, Talairach coordinates, and anatomical structures are shown. L = left hemisphere; R = right hemisphere.

Table 2.

Regions Showing Increased FIRE fMRI Activity to Increased Meaning at Fixation during Active Scene Viewing

Volume	Max	x	y	z	Anatomical Structure
107541	8.099	−25	−40	−6	L parahippocampal gyrus, L fusiform gyrus, L lingual gyrus
	7.804	28	−37	−12	R fusiform, R parahippocampal gyrus, R culmen
	6.938	34	−70	−12	R fusiform, R middle occipital gyrus, R lingual gyrus, R declive, R inferior occipital gyrus
	6.346	−37	−82	2	L middle occipital gyrus, L inferior occipital gyrus
	6.284	31	−76	26	R superior occipital gyrus, R middle temporal gyrus, R cuneus, R precuneus, R middle occipital gyrus, R angular gyrus
	6.045	−28	−79	26	L superior occipital gyrus, L cuneus, L precuneus, L middle temporal gyrus, L middle occipital gyrus
4617	5.820	−43	16	29	L middle frontal gyrus, L precentral gyrus, L inferior frontal gyrus
3699	5.271	43	16	26	R middle frontal gyrus, R inferior frontal gyrus
2808	5.660	22	7	5	R caudate, R lentiform nucleus
2376	5.835	−13	4	14	L caudate, L lentiform nucleus
1485	5.194	−4	13	47	L superior frontal gyrus, L medial frontal gyrus
783	4.440	−25	25	5	L insula
756	4.926	28	16	56	R middle frontal gyrus, R superior frontal gyrus
675	5.151	34	22	2	R insula, R inferior frontal gyrus
567	4.857	−10	−76	−21	L declive, L pyramis

Locations of peak activation are shown for each cluster with significant activity (Family-wise error [FWE] corrected at alpha < .05). Multiple peaks required separation by a minimum of 25 voxels. The volume of the cluster (μL), peak *z* score, Talairach coordinates, and anatomical structures are shown. L = left hemisphere; R = right hemisphere.

Table 3.

Regions Showing Increased FIRE fMRI Activity to Increased Meaning at Fixation during Active Scene Viewing, Controlling for Edges

Volume	Max	x	y	z	Anatomical Structure
44577	7.141	25	−40	−12	R parahippocampal gyrus, R fusiform gyrus
	6.433	35	−70	−12	R fusiform gyrus, R declive, R inferior occipital gyrus
	5.473	34	−68	23	R middle temporal gyrus, R superior occipital gyrus, R middle occipital gyrus
35208	7.189	−25	−40	−6	L parahippocampal gyrus, L fusiform gyrus, L culmen, L lingual gyrus
	6.601	−38	−68	−10	L middle occipital gyrus, L inferior occipital gyrus, L fusiform gyrus
2484	5.030	−46	16	29	L middle frontal gyrus
1674	4.753	43	16	26	R middle frontal gyrus
1107	4.923	19	4	5	R caudate, R lentiform nucleus
1107	5.141	−13	4	14	L caudate, L lentiform nucleus

Locations of peak activation are shown for each cluster with significant activity (Family-wise error [FWE] corrected at alpha < .05). Multiple peaks required separation by a minimum of 25 voxels. The volume of the cluster (μL), peak *z* score, Talairach coordinates, and anatomical structures are shown. L = left hemisphere; R = right hemisphere.