Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2021, Vol. 32(8) 1262–1270
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797621994768
www.psychologicalscience.org/PS
SAGE

Looking for Semantic Similarity: What a Vector-Space Model of Semantics Can Tell Us About Attention in Real-World Scenes

Taylor R. Hayes[1] and John M. Henderson[1,2]
[1]Center for Mind and Brain, University of California, Davis, and [2]Department of Psychology, University of California, Davis

Abstract

The visual world contains more information than we can perceive and understand in any given moment. Therefore, we must prioritize important scene regions for detailed analysis. Semantic knowledge gained through experience is theorized to play a central role in determining attentional priority in real-world scenes but is poorly understood. Here, we examined the relationship between object semantics and attention by combining a vector-space model of semantics with eye movements in scenes. In this approach, the vector-space semantic model served as the basis for a concept map, an index of the spatial distribution of the semantic similarity of objects across a given scene. The results showed a strong positive relationship between the semantic similarity of a scene region and viewers' focus of attention; specifically, greater attention was given to more semantically related scene regions. We conclude that object semantics play a critical role in guiding attention through real-world scenes.

Keywords

scene perception, object semantics, attention, eye movements

Received 4/29/20; Revision accepted 12/23/20

Given the importance of visual attention for vision and visual cognition, a fundamental theoretical question concerns how attention is guided through a scene in real time. For the past 20 years or so, models based on image salience have provided the most influential approach to answering this question (Itti & Koch, 2001; Koch & Ullman, 1985; Parkhurst et al., 2002). These classic saliency models propose that attention is controlled by contrasts in primitive, presemantic image features such as luminance, color, and edge orientation (Treisman & Gelade, 1980; Wolfe, 1994; Wolfe & Horowitz, 2017). Although theories based on image salience can account for key data regarding attentional guidance, it is also clear that in meaningful real-world scenes, human attention is strongly influenced by cognitive knowledge structures that represent the viewer's understanding of the scene and of the world (Antes, 1974; Buswell, 1935; Henderson & Hayes, 2017; Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Wu et al., 2014; Yarbus, 1967).

Cognitive-guidance theory emphasizes the importance of scene semantics in directing attention—where attention is "pushed" by the cognitive system to scene regions that are recognizable, informative, and relevant (Henderson, 2007). In this view, low-level image features are primarily used to identify potential target objects in the scene, not to assign attentional priority to those objects. Instead, attentional priority is determined by stored semantic representations of the relationships between the scene category and the objects it contains, along with the goals of the viewer (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003; Henderson & Hollingworth, 1999). For example, when we look at the kitchen scene in Figure 1a, we rapidly extract the scene's gist as "kitchen," which then allows

Corresponding Author:
Taylor R. Hayes, University of California, Davis, Center for Mind and Brain
E-mail: trhayes@ucdavis.edu

us to draw on our associated semantic knowledge of objects that tend to be found in kitchens (e.g., table, stove, sink) and where those objects tend to be located (Hayes & Henderson, 2019; Oliva & Torralba, 2006). Given the central role that stored semantic knowledge plays in cognitive-guidance theory, it is critical to gain a more complete understanding of the relationship between scene semantics and the control of attention in real-world scenes.

How can we study the relationship between stored semantic knowledge and attention in complex scenes? One approach is to use human ratings of the semantic content of local scene regions to generate "meaning maps" that can then be tested against attention (Henderson & Hayes, 2017). The meaning-map approach has shown that the meaning of a scene region is one of the best predictors of where people look in scenes regardless of the task (for a review, see Henderson et al., 2019). However, the meaning-map approach does not say precisely what makes a local scene region meaningful, beyond its overall semantic density (Henderson & Hayes, 2018). One interesting possibility is that meaningful scene regions are those that contain objects that are more conceptually related to one another and the broader scene category.

In the present study, we used a computational approach based on a vector-space model of semantics to test the role of object semantics in real-world scenes. The theory behind this approach is that objects that conceptually cohere with each other and with the scene category are most likely to be informative about the specific nature of that scene. For the vector-space model, we used ConceptNet Numberbatch, which combines how words are used in written text with crowd-sourced basic knowledge about the world (Günther et al., 2019). Unlike meaning maps that estimate the semantic density of isolated local scene regions, the vector-space model creates a representation based entirely on the semantic similarity between objects globally across a scene. Moreover, these semantic representations are generated computationally rather than requiring human raters and are derived from data that are not based on scenes or even visual in nature. Here, these semantic vectors serve as an index of viewers' stored semantic knowledge gained from experience with the world. We can then directly compare semantic representations derived from the vector-space model with overt attention as indexed by eye movements.

The semantic relationships between objects in each scene were used to generate concept maps for 100 scenes across 100 different categories, which were then compared with the eye movements of 100 participants viewing those scenes. The results indicated that the more semantically related the objects in a scene region were to the other objects in the scene and the scene category, the more likely that scene region was to be attended. These findings highlight the important role that object semantics play in determining where we look. The results also provide interesting new avenues for using computational methods to understand the role of semantics in scene perception.

Statement of Relevance

Stored knowledge gained from experience with the world is thought to play a central role in how people guide attention to process real-world scenes. Here, we tested the role of object knowledge by combining a model of object similarity derived from almost a trillion words of human-generated text with eye tracking in real-world scenes. We found evidence that the more conceptually similar a regions' objects were to the other objects in the scene and the scene category (e.g., kitchen), the more likely that region was to be attended. This result is especially striking given that object similarity was modeled independently of any visual scene input. The results provide direct evidence that humans use their stored knowledge of objects to help selectively process complex visual scenes, a theoretically important finding with implications for models in a wide range of areas including cognitive science, linguistics, computer vision, and visual neuroscience.

Method

Participants

One hundred fourteen University of California, Davis, undergraduate students with normal or corrected-to-normal vision participated in the experiment in exchange for course credit. All participants were naive to the purposes of the experiment and provided verbal informed consent as approved by the University of California, Davis, Institutional Review Board.

We have previously used this eye-movement data set to study general eye-movement characteristics in scenes (Cronin et al., 2020). The ConceptNet and center-proximity results are presented here for the first time.

Stimuli

Participants viewed 100 typical real-world scene images. The 100 scenes were chosen to represent 100 unique scene categories (e.g., kitchen, park); half of the images were indoor scenes and half were outdoor scenes.

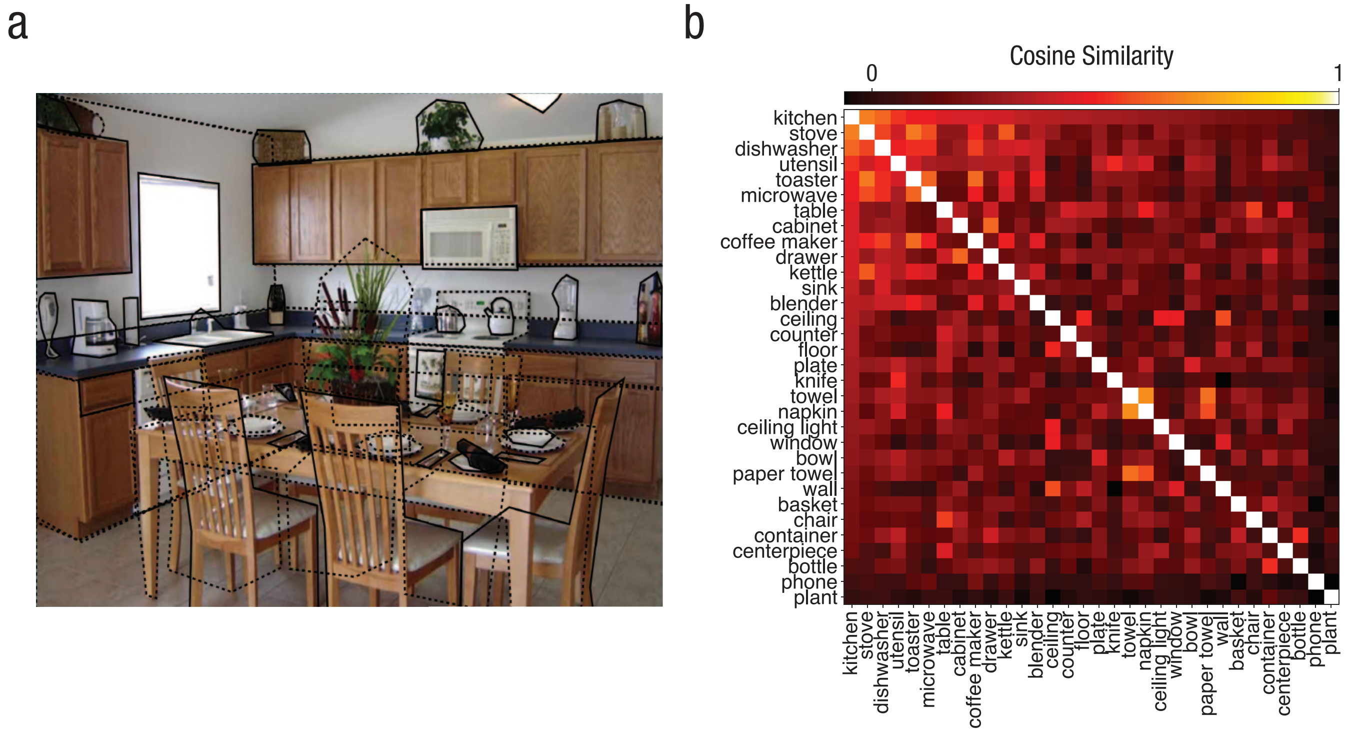

Fig. 1. An example scene segmentation (a) and its ConceptNet object-similarity matrix (b). The objects in the scene were first segmented (as indicated by the dashed lines) and labeled. Then the pairwise semantic similarity between the scene and the objects it contains was computed using ConceptNet Numberbatch. The heat map shows how semantically similar each object is to the scene category and all other scene objects.

Apparatus

Eye movements were recorded using an EyeLink 1000+ tower-mount eye tracker (spatial resolution 0.01°) sampling at 1,000 Hz (Version 1.5.2; SR Research, 2010). Participants sat 85 cm away from a 21-in. monitor and viewed scenes that subtended approximately 27° × 20° of visual angle. Head movements were minimized using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The display presentation was controlled with *Experiment Builder* software (Version 2.1.140; SR Research, 2017).

Eye-tracking calibration and data quality

A 9-point calibration procedure was performed at the start of each session to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49°, and a maximum error of less than 0.99°. Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9,500°/s^2, respectively). A drift correction was performed before each trial, and recalibrations were performed as needed.

The recorded data were examined for data artifacts from excessive blinking or calibration loss based on mean percentage signal across trials (Holmqvist et al., 2015). Fourteen participants with less than 75% signal were removed, leaving 100 participants who were tracked well (signal mean = 92.1%, *SD* = 5.31%).

Procedure

Each participant (*N* = 100) viewed 100 scenes for 12 s each while we recorded their eye movements. Each trial began with a fixation cross at the center of the display for 300 ms. For half the scenes, participants were instructed to memorize each scene in preparation for a later memory test. For the other half of the scenes, participants were instructed to indicate how much they liked each scene following the 12-s scene presentation. They made their judgments on a scale ranging from 1 to 3 by pressing the appropriate key. The scene set and presentation order of the two tasks were counterbalanced across participants. This procedure produced a large eye-movement data set that contained 334,725 fixations and an average of 3,347 fixations per participant.

Scene segmentation and labeling

To build a representation of the semantics of a scene, we first segmented and labeled each object in each scene (Fig. 1a). All objects that were present in the 100

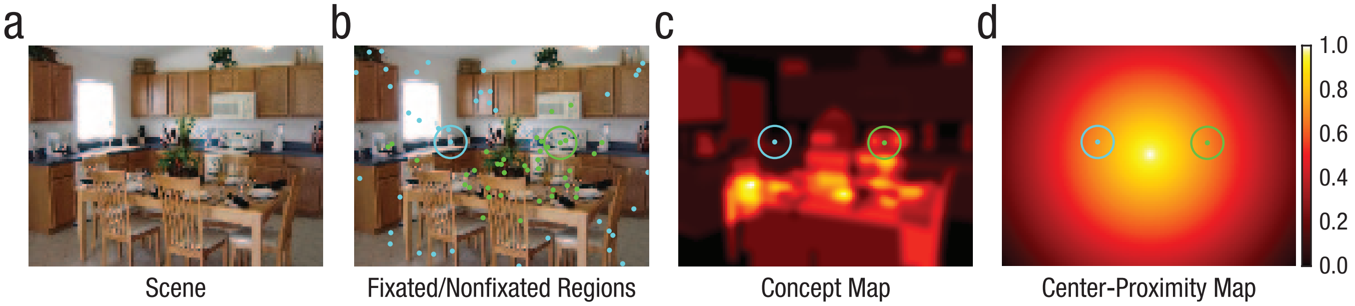

Fig. 2. Example scene (a) with the fixated and nonfixated regions for a single participant (b) and the corresponding values from the concept map (c) and center-proximity map (d). In (b), the green dots indicate the fixation locations, and the cyan dots indicate randomly sampled nonfixated regions that represent where the participant did not look. Together, these locations provide an account of which scene regions did and did not capture this participant's attention. The green and cyan circles mark the 3° window around one fixated and one nonfixated scene region, respectively. Each fixated and nonfixated location was then used to compute a mean ConceptNet (c) and center-proximity (d) map value across a 3° window centered on each location. The heat map in (c) reflects cosine similarity, and the heat map in (d) reflects scaled center proximity.

scenes were identified to form a set of all possible scene object labels. Then, from this global set of object labels, each object label was mapped to an individual object's spatial location within each scene using the Computer Vision Annotation Tool (https://github.com/opencv/cvat). In cases in which there were densely overlapping objects such as a stack of papers, the overlapping objects were grouped together and given a single label (e.g., "papers"). The scene segmentation defined the spatial locations of each object, and the object labels were used to compute the semantic similarities among the different objects and between each object and its scene category for each scene.

ConceptNet Numberbatch

ConceptNet Numberbatch (Version 17.06; Speer et al., 2017) was used to estimate the semantic similarity between object labels as vectors in a high-dimensional space. ConceptNet Numberbatch uses an ensemble approach combining the semantic vectors from Word 2vec (Mikolov et al., 2013) and GloVe (Version 1.2; Pennington et al., 2014)—which learn how words are associated with each other from large text corpora (i.e., Google News, 100 billion words; Common Crawl, 840 billion words)—with ConceptNet, a knowledge graph that draws on expert-created resources (WordNet, Fellbaum, 1998; Open Mind Common Sense, Singh et al., 2002; OpenCyc, Lenat & Guha, 1989) and crowdsourced knowledge (Auer et al., 2007; Kuo et al., 2009; von Ahn et al., 2006). The benefit of the ConceptNet Numberbatch ensemble approach is that it produces high-quality semantic representations that are better than any single component of the ensemble (e.g., Word2vec) on a number of important semantic benchmarks, such as SAT analogies (Speer et al., 2017).

Concept map

We then used the generated ConceptNet Numberbatch semantic vectors to compute how semantically related the objects in each scene were to one another and to the scene itself using all the pairwise-similarity values (Fig. 1b). Specifically, we computed the similarity between each pair of object-label vectors using cosine similarity (i.e., the normalized dot product of the two word vectors). The process for generating a scene concept map from the pairwise-similarity values consisted of three steps. First, for each object in a given scene, a mean similarity value was computed by averaging its similarity across all other within-scene objects and the scene category (i.e., the mean across the object's row or column in the similarity matrix in Fig. 1b). Second, each object's mean similarity value was then added to the spatial location or locations in which the segmented object (or objects) occurred in the scene. The final scene concept map was then smoothed using a Gaussian filter (MATLAB, The MathWorks, Natick, MA; imgaussfilt function, $\sigma = 10$).

This procedure produced a scene concept map that captured semantic-object similarity (i.e., how similar the objects at a given location are to everything else in the scene and the scene itself) while also representing the semantic density (i.e., objects on top of other objects) of each scene region (Fig. 2b). The concept maps could then be directly compared with where observers looked in each scene.

Center-proximity map

In addition to the concept map, we also generated a center-proximity map that served as a global representation of how far each fixated location in the scene image was from the scene center. Specifically, this map measured the inverted Euclidean distance from the center pixel of the scene to all other pixels in the scene image (Fig. 2d). The center-proximity map was used to explicitly control for observers' general bias to look more centrally than peripherally in scenes, independently of the underlying scene content (Hayes & Henderson, 2020; Tatler, 2007).

Fixated and nonfixated scene locations

To model the relationship between scene features and overt attention, we needed to compare where each participant looked in each scene with where they did not look (Nuthmann et al., 2017). Therefore, for each fixation, we computed the mean concept-map value (Fig. 2c) and center-proximity-map value (Fig. 2d) by taking the average over a 3° window around each fixation in each map (Fig. 2b, neon green locations). To represent scene features that were not associated with overt attention for each participant, we randomly sampled an equal number of scene locations where each particular participant did not look in each scene they viewed (Fig. 2b, cyan locations). The only constraint for the random sampling of the nonfixated scene regions was that the nonfixated 3° windows could not overlap with any of the 3° windows of the fixated locations. This procedure provided the concept-map values and center-proximity values that were and were not associated with attention for each individual scene viewed by each individual participant.

General linear mixed-effects (GLME) model

We applied a GLME model to our data using the *lme4* package (Version 1.1-27.1; Bates et al., 2015) in the R programming environment (Version 3.6.0; R Core Team, 2019). A mixed-effects modeling approach was chosen because it does not require aggregating the eye-movement data at the participant or scene level like analyses of variance or map-level correlations. Instead, both participant and scene could be explicitly modeled as random effects. Additionally, the GLME approach allowed us to control for the role of center bias by including the distance from the screen center (Fig. 2d) as both a fixed effect and an interaction term with the concept-map values. We used a GLME logit model to investigate which factors were predictive of whether a scene region was attended or not (Fig. 2). Specifically, whether a region was fixated (1) or not fixated (0) was the dependent variable, and the continuous concept-map value, continuous center-proximity value, and their interaction were treated as fixed effects. We included participant and scene as crossed random effects. There was no significant difference between the memorization and aesthetic-judgment tasks, so the data were collapsed over task.

Results

Using the concept maps, center-proximity maps, and eye-movement data, we tested the hypothesis that attention in scenes is guided by stored semantic knowledge. If this hypothesis is correct, then participants should be more likely to fixate on scene regions that are rich in conceptual information, all else being equal. This hypothesis should hold when we control for the overall tendency for participants to look more centrally regardless of scene content (Hayes & Henderson, 2020; Tatler, 2007) and for the random effects of different participants and scenes.

The fixation-location GLME results are shown in Figure 3 and Table 1. The results indicated a significant interaction between the probability of fixating a scene region and its concept-map and center-proximity values (Fig. 3a). As shown in Figure 3b, higher concept regions were more likely to be fixated than lower concept regions, and this effect was enhanced when regions were closer to the scene center and reduced when regions were farther away from the scene center. The isolated marginal effects of the concept-map and center-proximity predictors are shown in Figure 3c. The marginal effects indicated again that regions with higher concept-map values were more likely to be fixated, all else being equal (Fig. 3c). Importantly, the relationship between the concept maps and fixations could not be accounted for by differences in low-level visual salience, and the pairwise ConceptNet similarity structures were only partially explained (32%) by highly diagnostic visual features such as object shape (see the Supplemental Material available online). Together, our findings provide strong evidence that stored semantic knowledge is strongly associated with where we look in real-world scenes.

Discussion

One of the central tenets of cognitive-guidance theory is that we use stored knowledge structures gained from our previous experience with the world to guide our attention in real-world scenes. Whereas previous research has shown that task relevance (Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2009; Neider & Zelinsky, 2006; Rothkopf et al., 2007; Tatler et al., 2011; Torralba et al., 2006; Yarbus, 1967) and information density

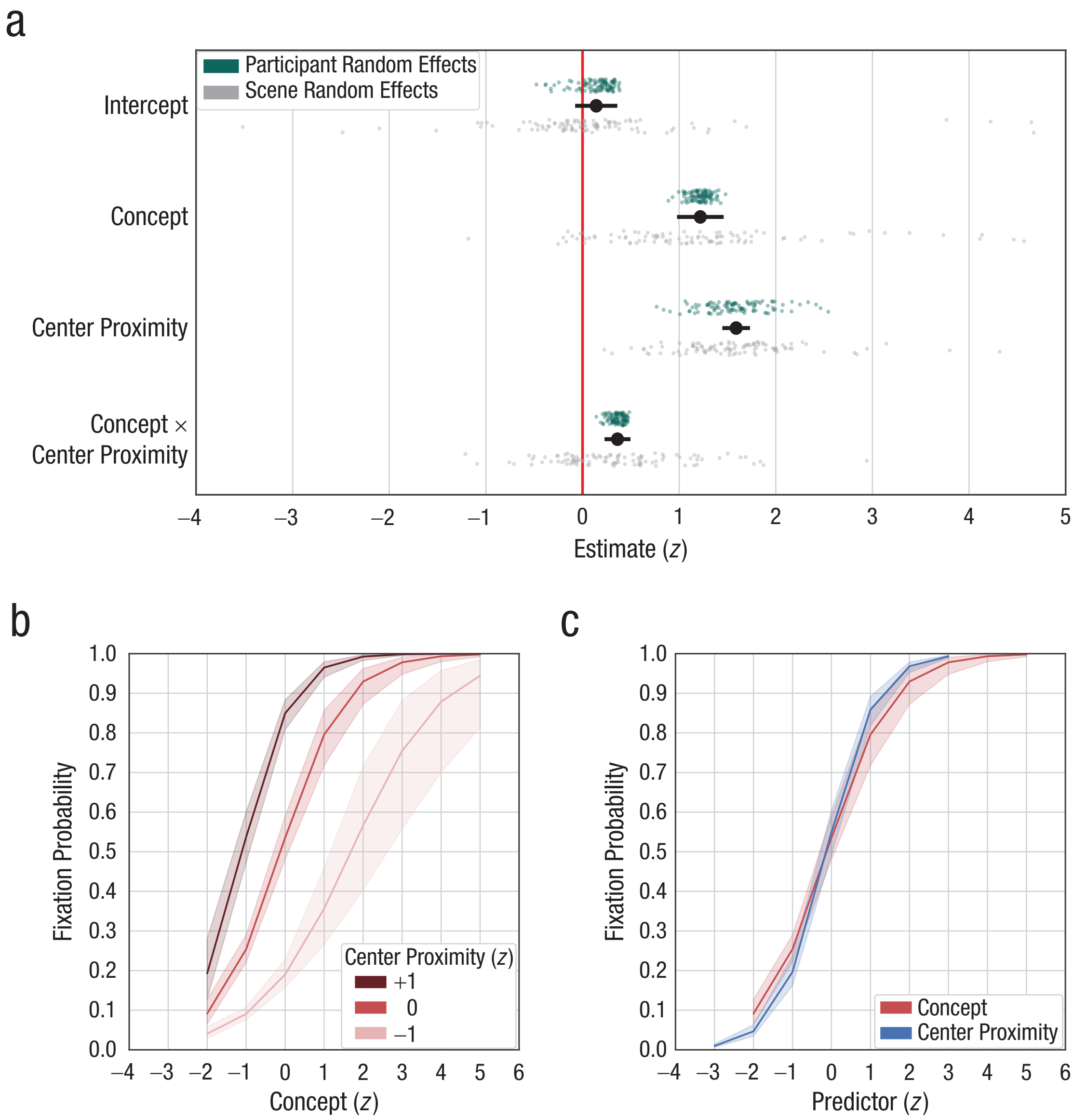

Fig. 3. Fixation-location results from the general linear mixed-effects model. Whether a scene region was fixated or not was the dependent variable, and the concept-map value, center-proximity value, and their interaction were included as fixed effects. In (a), the black dots show estimates for the intercept and fixed effects (error bars indicate 95% confidence intervals). Participants (green dots) and scenes (gray dots) were both accounted for in the model as random effects (intercept and slope). The line plots show the probability of fixating a scene (b) as a function of the interaction between concept and center-proximity effects and (c) as a function of predictor and the marginal effects of concept and center proximity. Error bands reflect 95% confidence intervals.

(Antes, 1974; Buswell, 1935; Henderson & Hayes, 2017, 2018; Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Wu et al., 2014; Yarbus, 1967) are related to attention in scenes, the specific relationship between scene semantics and attention is largely unexplored. To address this question, here we used a semantic vector-space model based on text corpora as an index of stored semantic knowledge to directly test the hypothesis that attention is driven by semantics in real-world scenes. We found that the more the objects in a scene region semantically cohered with the scene category and the other objects in the scene, the more likely that region was to be fixated. This result supports cognitive-guidance theory by establishing a direct link between attention and a global representation of all the semantic associations between a scene and its objects.

The present work extends our understanding of the relationship between attention and scene semantics in

Table 1. Fixation-Location Results From the General Linear Mixed-Effects Model

	Fixed effects					Random effects	
Predictor	β	95% CI	*SE*	*z* statistic	*p*	By participant (*SD*)	By scene (*SD*)
Intercept	0.14	[–0.08, 0.36]	0.11	1.27	.20	0.19	1.12
Concept	1.22	[0.98, 1.46]	0.12	9.90	< .001	0.11	1.26
Center proximity	1.59	[1.45, 1.73]	0.07	21.81	< .001	0.35	0.64
Concept × Center Proximity	0.36	[0.23, 0.50]	0.07	5.29	< .001	0.08	0.68

Note: CI = confidence interval.

several novel ways. First, this work uniquely focuses on grounding the study of scene semantics in a general computational model of conceptual knowledge. This approach has been highly successful in other areas of cognitive science, such as computational linguistics (Armeni et al., 2017; Brennan, 2016; Hale et al., 2015), but has so far not been applied to scene perception. The ability to generate semantic scene representations computationally that can then be used to test the influence of meaning on attentional control represents an important way forward. Second, prior work examining the spatial distribution of scene semantics across a scene has been region based rather than object based, but the literature suggests that attention is strongly biased toward object representations (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010). The approach introduced here offers a method for studying the distribution of semantic density across a scene while simultaneously taking explicit account of perceptual objects and their concepts. Third, the integration of conceptual knowledge across vision and language is an important topic of research in cognitive science, and the common use of vector-space models in both domains can provide a foundation for linking semantics across them. In this regard, it is particularly interesting that a model derived entirely from nonvisual information was able to account for the influence of scene semantics on visual attention. In future work, it will be important to determine whether the same semantic representations can serve both vision and language when they operate together.

More broadly, using semantic vector-space models to index stored scene knowledge opens up interesting avenues for future computationally grounded work on other aspects of scene semantics. For example, whereas the scenes we used here were typical real-world scenes without any added semantically inconsistent objects, a large body of previous work suggests that semantically inconsistent objects, once fixated, are given additional attentional priority (Biederman et al., 1982; Henderson et al., 1999; Võ & Henderson, 2011). From an information-theoretic perspective, semantically anomalous objects in scenes carry important information because they violate our expectations. In this sense, the stored semantic knowledge such as that captured by ConceptNet is the very kind of information an observer would need in order to identify a semantic-category outlier in the first place. The current approach could likely be generalized to account for semantic-inconsistency effects by identifying and upweighting semantic-outlier objects (i.e., objects with very low average similarity values relative to the other objects in the scene). Additionally, vector-space models of semantics could also serve as a quantitative tool for experimental design. For example, in studies manipulating semantic objects, vector-space models could be used to select which semantically inconsistent object should be included to achieve a specified amount of semantic inconsistency relative to all the other objects in the scene or the scene category.

In summary, we tested whether stored semantic knowledge guides attention in real-world scenes by combining eye-tracking data with concept maps derived from vector-based semantic representations of object concepts. Importantly, the vector-space representations were derived entirely independently of the scenes we tested and, indeed, were not based on scene representations at all. We found that the greater the semantic coherence of a scene region as represented by concept maps, the more likely that region was to be attended. These findings suggest that humans use their stored semantic representations to help attentively process complex scenes, a result with implications for theories and models in a wide range of areas including cognitive science, computer vision, linguistics, and visual neuroscience.

Transparency

Action Editor: M. Natasha Rajah
Editor: Patricia J. Bauer
Author Contributions

T. R. Hayes and J. M. Henderson conceived of and designed the study. T. R. Hayes collected and analyzed the data and drafted the manuscript. Both of the authors revised the manuscript and approved the final version for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the National Eye Institute of the National Institutes of Health (Award No. R01-EY027792). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Open Practices

Data and materials for this study have not been made publicly available, and the design and analysis plan were not preregistered.

ORCID iD

Taylor R. Hayes https://orcid.org/0000-0003-0078-3642

Acknowledgments

We thank Serena Bandtell for assistance with collecting eye-tracking data and Elizabeth H. Hall and Anthony Lagunda for assistance with collecting object-segmentation and labeling data.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797621994768

References

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

Armeni, K., Willems, R. M., & Frank, S. L. (2017). Probalistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience and Biobehavioral Reviews*, *83*, 579–588.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Eds.), *The semantic web* (pp. 722–735). Springer.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143–177.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, *7*, 299–313.

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

Cronin, D. A., Hall, E. H., Goold, J. E., Hayes, T. R., & Henderson, J. M. (2020). Eye movements in real-world scene photographs: General characteristics and effects of viewing task. *Frontiers in Psychology*, *10*, Article 2915. https://doi.org/10.3389/fpsyg.2019.02915

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

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

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Hale, J., Lutz, D., Luh, W.-M., & Brennan, J. (2015, June). Modeling fMRI time courses with linguistic structure at various grain sizes. In T. O'Donnell & M. van Schijndel (Eds.), *Proceedings of the 6th workshop on Cognitive Modeling and Computational Linguistics* (pp. 89–97). Association for Computational Linguistics. https://www.aclweb.org/anthology/W15-1110

Hayes, T. R., & Henderson, J. M. (2019). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin & Review*, *26*(5), 1683–1689.

Hayes, T. R., & Henderson, J. M. (2020). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, *82*(3), 985–994.

Hayhoe, M. M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

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

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

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

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *3*(2), Article 19. https://doi.org/10.3390/vision3020019

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

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*, 850–856.

Henderson, J. M., Weeks, P. A., & 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.

Holmqvist, K., Nyström, R. M., Dewhurst, R., Andersson, R., Jorodzka, H., & van de Weijer, J. (2015). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*, 194–203.

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

Kuo, Y.-L., Lee, J.-C., Chiang, K.-Y., Wang, R., Shen, E., Chan, C.-W., & Hsu, J. Y.-J. (2009). Community-based game design: Experiments on social games for commonsense data collection. In P. Bennett & R. Chandrasekar (Eds.), *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 15–22). Association for Computing Machinery. http://doi.acm.org/10.1145/1600150.1600154

Lenat, D. B., & Guha, R. V. (1989). *Building large knowledge-based systems: Representation and inference in the Cyc project*. Addison-Wesley Long-Man.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology*, *4*, 565–572.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547–552.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://arxiv.org/abs/1301.3781

Neider, M. B., & Zelinsky, G. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614–621.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond center bias? A new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience*, *11*, Article 491. https://doi.org/10.3389/fnhum.2017.00491

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

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 102–123.

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://www.aclweb.org/anthology/D14-1162

R Core Team. (2019). *R: A language and environment for statistical computing* (Version 3.6.0) [Computer software]. Retrieved from https://www.R-project.org/

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

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In R. Meersman & Z. Tari (Eds.), *On the move to meaningful Internet systems 2002: CoopIS, DOA, and ODBASE* (pp. 1223–1237). Springer.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In S. Singh & S. Markovitch (Chairs), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 4444–4451). AAAI Press.

SR Research. (2010). *EyeLink 1000 user manual* (Version 1.5.2). Author.

SR Research. (2017). *SR Research Experiment Builder user manual* (Version 2.1.140). Author.

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

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

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*, 766–786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Võ, M. L. H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception & Psychophysics*, *73*, 1742–1753.

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: A game for collecting common-sense facts. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. Olson (Eds.), *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 75–78). Association for Computing Machinery. https://doi.org/10.1145/1124772.1124784

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202–238.

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

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, Article 54. https://doi.org/10.3389/fpsyg.2014.00054

Yarbus, A. L. (1967). *Eye movements and vision*. Plenum.