Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short Communication

Meaning maps capture the density of local semantic features in scenes: A reply to Pedziwiatr, Kümmerer, Wallis, Bethge & Teufel (2021)

John M. Henderson [a,b,*], Taylor R. Hayes [a], Candace E. Peacock [a,b], Gwendolyn Rehrig [b]

[a] Center for Mind and Brain, University of California, Davis, USA
[b] Department of Psychology, University of California, Davis, USA

ARTICLE INFO

Keywords:
Eye movements
Scene perception
Attention

ABSTRACT

Pedziwiatr, Kümmerer, Wallis, Bethge, & Teufel (2021) contend that Meaning Maps do not represent the spatial distribution of semantic features in scenes. We argue that Peszìwiatr et al. provide neither logical nor empirical support for that claim, and we conclude that Meaning Maps do what they were designed to do: represent the spatial distribution of meaning in scenes.

Pedziwiatr et al. (2021, henceforth PKWBT) claim that Meaning Maps (Henderson & Hayes, 2017) “are insensitive to image meaning when predicting human fixations”. They reach this conclusion based on two problematic arguments. These arguments are: (1) Because Meaning Maps do not capture object-scene semantic consistency, they do not capture any aspects of semantic content; (2) Because Meaning Maps do no better than DeepGaze 2 (DG2) both generally and with regard to object-scene consistency, Meaning Maps reduce to the same type of non-semantic physical features used by DG2, a deep-learning saliency model trained on human eye movements (Kümmerer, Wallis, & Bethge, 2016). Neither of these conclusions follow from the premises, arguments, or data presented by PKWBT, and neither is correct.

Because PKWBT challenge the fundamental ability of Meaning Maps to represent meaning and their usefulness as a tool for studying meaning in scenes, we begin by touching on a few general preliminaries concerning Meaning Maps and the theory they were developed to investigate. This will then set the stage for our main points concerning PKWBT that follow.

First, the theoretical claim at stake in our work is simple: When humans visually perceive the real world, visual-spatial attention is driven in large part by our understanding and interpretation of what we are seeing, along with what we are trying to accomplish (Henderson, 2003, 2007, 2011). Research supporting this idea has a long history in visual cognition (Buswell, 1936; Yarbus, 1967) and is reinforced by a large body of evidence (Einhäuser, Rutishauser, & Koch, 2008; Tatler, Hayhoe, Land, & Ballard, 2011; Torralba, Oliva, Castelhano, & Henderson, 2006; Williams & Castelhano, 2019; Wu, Wick, & Pomplun, 2014). PKWBT acknowledge this point to some degree, allowing that there are task effects on attention in scenes while denying a general influence of semantic content. Importantly there is a growing body of behavioral and neural evidence that fundamental visual-spatial attention systems are strongly influenced by the semantic properties of meaningful visual stimuli, and indeed that semantic properties often override physical properties in the control of attention. This evidence has been observed in traditional attention paradigms (Malcolm, Rattinger, & Shomstein, 2016; Shomstein, Malcolm, & Nah, 2019), simplified object displays (Nuthmann, de Groot, Huettig, & Olivers, 2019), and in scene perception research (Võ, Boettcher, & Draschkow, 2019; Williams & Castelhano, 2019; Wu et al., 2014). At this point there is little remaining doubt that the influence of physical (image-based, non-semantic) saliency on visual-spatial attention can be overridden by many factors including meaning, and what remains to be determined is how this overriding is accomplished (Luck, Gaspelin, Folk, Remington, & Theeuwes, 2020).

Given the strong evidence for a central role of semantics in attentional control, why is there still such an emphasis on physical features in much of the attention literature, including the scene attention literature? We suspect that this focus on physical salience arises for at least two important reasons (Henderson, 2017). First, visual-spatial attention has traditionally been studied using stimuli that mostly or entirely lack semantic content, in large part because psychophysical and related methods are more easily deployed for such stimuli. Second, attentional modeling has similarly traditionally focused on physical features that exclude semantic content because modeling attention over image computable physical features has historically been more tractable than modeling over semantic features (though see Hayes and Henderson

* Corresponding author.
E-mail address: johnhenderson@ucdavis.edu (J.M. Henderson).

https://doi.org/10.1016/j.cognition.2021.104742
Received 12 January 2021; Received in revised form 13 April 2021; Accepted 15 April 2021
0010-0277/© 2021 Published by Elsevier B.V.

(2021)). Certainly, much has been learned about the fundamental nature of the attention system by taking this approach. At the same time, it is clearly not the whole story.

In visual cognition and cognitive neuroscience, a great deal of the evidence for the role of semantics on attention in complex real-world scenes has been based on manipulations of object-scene semantic relationships. For example, it is common to swap semantically consistent objects across scenes (e.g., classically swapping an octopus and a tractor in an underwater and farm scene respectively) to create semantic inconsistencies (Loftus & Mackworth, 1978). This is a powerful manipulation for investigating and establishing the causal role of semantics on attention, and it is one that we have often taken ourselves (Brockmole & Henderson, 2008; Henderson, Weeks Jr., & Hollingworth, 1999; Võ & Henderson, 2009). It is this approach that PKWBT also focus on. But a drawback of this method is that only one discrete region of a scene (the manipulated object) can be investigated in each scene, limiting the amount of data that can be collected and the conclusions that can be drawn about changes in semantic density across the entire scene.

In comparison to the literature on object-scene semantic consistency, in the literature investigating the role of physical salience on attention in scenes, physical properties are typically represented as a spatially continuous distribution of salience values across a scene in the form of a saliency map. The lack of an analogous spatially continuous representation of semantic properties has made it difficult to directly compare physical features to semantic features in scenes. Given these considerations, our goal in developing Meaning Maps was to generate a continuous representation of the spatial distribution of semantic features across a real-world scene in the same format as physical saliency maps, so that the two can be directly compared.

Because there have been no computational methods that can automatically produce semantic density maps, we capitalized on the semantic systems of human raters to tell us how meaning varies across scenes. Importantly, given that models based on physical salience do not consider global scene characteristics, in our initial work we purposefully focused on creating Meaning Maps that represented the distribution of context-free semantic density for local scene regions without taking context into account, in what we have called "context-free" Meaning Maps (Henderson, 2020; Henderson, Hayes, Peacock, & Rehrig, 2019). In other words, we purposefully excluded contextualized meaning such as object-scene semantic relationships from Meaning Maps, not because context would be impossible to include, but as a conscious decision given our specific goals at the time. Indeed, this is one of the reasons we created Meaning Maps instead of using rating paradigms that already existed (Antes, 1974; Mackworth & Morandi, 1967; t' Hart, Schmidt, Roth, & Einhäuser, 2013).

Returning to PKWBT, the conclusion that Meaning Maps represent physical features rather than semantic features requires assuming that raters ignore the instructions they are given. In our original context-free rating procedure, subjects are asked to rate each individually-presented patch based on their assessment of how informative and recognizable that patch is. There is no reason to believe that raters perversely ignore these instructions and rate physical features instead. Indeed, we have successfully generated Meaning Maps designed to capture different semantic features simply by changing the rating instructions. For example, we have generated what we call "contextualized" Meaning Maps by presenting exactly the same patches used in context-free Meaning Maps, but with each individual patch shown with its scene (Peacock, Hayes, & Henderson, 2019). We have also generated "Grasp Maps" using exactly the same patches with instructions focused on whether the region depicts an entity that can be grasped (Rehrig, Peacock, Hayes, Henderson, & Ferreira, 2020). Importantly, when the instructions are changed, subjects change their ratings to reflect the semantic features they are asked to rate, leading to different maps, even though the physical features are held constant. If raters were simply rating physical features in each patch, the ratings would not change as a function of the semantic characteristics of the rating task. Therefore, contrary to the unsupported claim in the target article, we have direct evidence that Meaning Maps reflect semantic features. Furthermore, neurocognitive work in our lab provides converging evidence supporting this conclusion. For example, cortical areas along the ventral visual stream supporting higher level scene processing show activation that is associated with the semantic values of fixated scene regions captured by Meaning Maps, but not with simple visual features (Henderson, Goold, Choi, & Hayes, 2020). Ongoing research in our lab shows that this activation is more extensive than the activation associated with DG2, strongly suggesting that the two representations do not capture the same information. We also see a similar dissociation between physical salience and semantic features represented by Meaning Maps in EEG work examining the time-course of scene perception (Kiat, Hayes, Henderson, & Luck, 2020).

Meaning Maps offer a powerful general method for investigating the spatial distribution of semantic features (or what we have called "semantic density") in complex scenes. By design, our original context-free Meaning Maps capture some aspects of scene meaning and not others. Contrary to the conclusion of PKWBT, this limitation in scope is not evidence that context-free Meaning Maps do not represent any aspect of meaning. PKWBT focus on object-scene semantic relationships, arguing that because neither context-free Meaning Maps nor DG2 account for the influence of object-scene consistency on attention, Meaning Maps and DG2 must reduce to the same type of non-semantic underlying representation. This conclusion obviously does not follow: The fact that two representational systems are equally unable to account for some phenomenon does not logically require the conclusion that they are identical to each other. Here the two types of representations fail for different reasons, as is often the case when competing models are compared.

There are many other aspects of scene meaning that the original context-free Meaning Maps also do not capture. For example, they do not capture aspects of meaning that depend on the viewer's task. They do not capture the increased meaning of water to a thirsty person. They do not capture the heightened meaning of a spider to an arachnophobe. They do not capture changes to meaning that might arise as a consequence of an unfolding event. They do not capture individual differences in meaning based on the unique histories of individual people. They do not capture a viewer's transient motivation. There is nothing about these cases that leads to the conclusion that Meaning Maps therefore do not represent scene meaning at all. And critically, it is a simple matter to extend the "classic" context-free Meaning Map approach to include additional types of semantic features, as we have done (Peacock et al., 2019; Rehrig et al., 2020). Importantly, whereas Meaning Maps can easily be extended to investigate a variety of semantic features, it is far less clear whether deep learning models like DG2 can ever in principle capture object-scene semantic features, or indeed any type of semantic feature.

PKWBT do a disservice to many fields of inquiry by incorrectly and unnecessarily dismissing a method that has already proven useful for investigating scene semantics across several domains of research, and that will likely continue to prove useful in new domains. For example, Meaning Maps have already been used to investigate scene memory (Bainbridge, Hall, & Baker, 2019; Ramey, Yonelinas, & Henderson, 2020), language production (Ferreira & Rehrig, 2019; Henderson, Hayes, Rehrig, & Ferreira, 2018; Rehrig et al., 2020), mind wandering (Krasich, Huffman, Faber, & Brockmole, 2020), active vision in immersive VR environments (Haskins, Mentch, Botch, & Robertson, 2020), and infant attentional development (Klotz, Hayes, Pomaranski, Henderson, & Oakes, 2021). Ongoing work in our lab uses Meaning Maps to investigate the nature and time-course of scene perception in the brain (Henderson et al., 2020; Kiat et al., 2020). These disparate lines of research demonstrate the utility of meaning maps for studying scene semantics. Questioning all of this work without basis unnecessarily confuses and undermines both existing and potentially new and important empirical and theoretical lines of investigation.

PKWBT also conclude that Meaning Maps do not predict attention in scenes as well as DG2 in a free viewing task. This conclusion is based on empirical results presented by PKWBT that will require replication and corroboration conducted in the course of normal science. Because it is not directly relevant to our main focus here, we set it aside for the purposes of this commentary.

In sum, there is neither empirical nor logical reason to dismiss Meaning Maps, and much reason to conclude that they do what they were designed to do: represent the spatial distribution of meaning in scenes.

CRediT authorship contribution statement

John M. Henderson: Writing - original draft. **Taylor R. Hayes:** Writing - review & editing. **Candace E. Peacock:** Writing - review & editing. **Gwendolyn Rehrig:** Writing - review & editing.

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 decision to publish or preparation of the manuscript.

References

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

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications, 10*(1), 5. https://doi.org/10.1038/s41467-018-07830-6.

Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: Effects of object-scene consistency. *Visual Cognition, 16*(2–3), 375–390. https://doi.org/10.1080/13506280701453623.

Buswell, G. T. (1936). *How people look at pictures*. Chicago: University of Chicago Press. https://doi.org/10.1037/h0053409.

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

Ferreira, F., & Rehrig, G. (2019). *Linearisation during language production: Evidence from scene meaning and saliency maps. Language, cognition and neuroscience* (pp. 1–11). https://doi.org/10.1080/23273798.2019.1566562.

Haskins, A. J., Mentch, J., Botch, T. L., & Robertson, C. E. (2020). Active vision in immersive, 360° real-world environments. *Scientific Reports, 10*(1), 14304. https://doi.org/10.1038/s41598-020-71125-4.

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: What a vector space model of semantics can tell us about attention in real-world scenes. *Psychological Science* (in press).

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

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

Henderson, J. M. (2011). Eye movements and scene perception. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 593–606). Oxford ; New York: Oxford University Press.

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

Henderson, J. M. (2020). Meaning and attention in scenes. In *, 73. Psychology of learning and motivation* (pp. 95–117). Elsevier. https://doi.org/10.1016/bs.plm.2020.08.002.

Henderson, J. M., Goold, J. E., Choi, W., & Hayes, T. R. (2020). Neural correlates of fixated low- and high-level scene properties during active scene viewing. *Journal of Cognitive Neuroscience, 32*(10), 2013–2023. https://doi.org/10.1162/jocn_a_01599.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*(October), 743–747. https://doi.org/10.1038/s41562-017-0208-0.

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), 19. https://doi.org/10.3390/vision3020019.

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports, 8*(1), 1–9. https://doi.org/10.1038/s41598-018-31894-5.

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

Kiat, J., Hayes, T., Henderson, J., & Luck, S. (2020). Assessing the time course of saliency and meaning: Representational similarity analysis of ERP responses to natural scenes. *Journal of Vision, 20*(11). https://doi.org/10.1167/jov.20.11.1629, 1629–1629.

Klotz, S., Hayes, T. R., Pomaranski, K., Henderson, J. M., & Oakes, L. (2021, April). *Experience and age guide Infants' attention to meaning in scenes* (Society for Research in Child Development).

Krasich, K., Huffman, G., Faber, M., & Brockmole, J. R. (2020). Where the eyes wander: The relationship between mind wandering and fixation allocation to visually salient and semantically informative static scene content. *Journal of Vision, 20*(9), 10. https://doi.org/10.1167/jov.20.9.10.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv*, 1610.01563 [Cs, q-Bio, Stat] http://arxiv.org/abs/1610.01563.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565–572. https://doi.org/10.1037/0096-1523.4.4.565.

Luck, S. J., Gaspelin, N., Folk, C. L., Remington, R. W., & Theeuwes, J. (2020). Progress toward resolving the attentional capture debate. *Visual Cognition*, 1–21. https://doi.org/10.1080/13506285.2020.1848949.

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

Malcolm, G. L., Rattinger, M., & Shomstein, S. (2016). Intrusive effects of semantic information on visual selective attention. *Attention, Perception, & Psychophysics, 78* (7), 2066–2078. https://doi.org/10.3758/s13414-016-1156-x.

Nuthmann, A., de Groot, F., Huettig, F., & Olivers, C. N. L. (2019). Extrafoveal attentional capture by object semantics. *PLoS One, 14*(5), Article e0217051. https://doi.org/10.1371/journal.pone.0217051.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica, 198*, 102889. https://doi.org/10.1016/j.actpsy.2019.102889.

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition, 206*, 104465. https://doi.org/10.1016/j.cognition.2020.104465.

Ramey, M. M., Yonelinas, A. P., & Henderson, J. M. (2020). Why do we retrace our visual steps? Semantic and episodic memory in gaze reinstatement. *Learning & Memory, 27* (7), 275–283. https://doi.org/10.1101/lm.051227.119.

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(9), 1659–1681. https://doi.org/10.1037/xlm0000837.

Shomstein, S., Malcolm, G. L., & Nah, J. C. (2019). Intrusive effects of task-irrelevant information on visual selective attention: Semantics and size. *Current Opinion in Psychology, 29*, 153–159. https://doi.org/10.1016/j.copsyc.2019.02.008.

t' Hart, B. M., Schmidt, H. C. E. F., Roth, C., & Einhäuser, W. (2013). Fixations on objects in natural scenes: Dissociating importance from salience. *Frontiers in Psychology, 4* (JUL), 1–9. https://doi.org/10.3389/fpsyg.2013.00455.

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

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766.

Võ, M. L. H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology, 29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009.

Võ, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 24.

Williams, C. C., & Castelhano, M. S. (2019). The changing landscape: High-level influences on eye movement guidance in scenes. *Vision, 3*(3), 33. https://doi.org/10.3390/vision3030033.

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5*(FEB), 1–13. https://doi.org/10.3389/fpsyg.2014.00054.

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