CHAPTER FOUR

Meaning and attention in scenes

John M. Henderson*
Center for Mind and Brain and Department of Psychology, University of California, Davis, CA, United States
*Corresponding author: e-mail address: johnhenderson@ucdavis.edu

Contents

Abstract

Perception of a complex visual scene requires that important regions be prioritized and attentionally selected. What is the basis for this selection? Although much research has focused on the spatial distribution of image salience as an important factor guiding attention, relatively little work has focused on the spatial distribution of semantic features (meaning) across a scene. To address this imbalance, we have recently developed a new method for measuring, representing, and evaluating the spatial distribution of meaning in scenes and its influence on attention. In this method, the spatial distribution of meaning is represented as a meaning map. Meaning maps are generated from crowd-sourced responses given by naïve subjects who rate a large number of scene patches drawn from each scene. Meaning maps are coded in the same format as traditional saliency maps, and therefore both types of maps can be directly compared and evaluated against the spatial distribution of attention derived from viewers' eye fixations. In this review I provide an overview of my lab's research comparing the influences of meaning and image salience on attentional guidance in real-world scenes. Overall, we have found that both the spatial distribution of meaning and physical

Psychology of Learning and Motivation, Volume 73
ISSN 0079-7421
https://doi.org/10.1016/bs.plm.2020.08.002

salience across a scene predict the spatial distribution of attention, but when the correlation between meaning and image salience is statistically controlled, only meaning uniquely accounts for variance in attention. I discuss the theoretical implications of these findings and point to new questions for the future.

Explaining how we perceive real-world visual environments is a fundamental goal in visual cognition and vision science. Such an explanation will require unraveling how the brain solves the difficult computational problems that arise given the complex visual scenes we encounter in our daily lives. The world contains an enormous amount of visual information, but the processing capacity of human vision and visual cognition are severely limited in their bandwidth: Only a small fraction of the information latent in the visual world can be analyzed at any given moment. Efficient visual cognition therefore requires selecting the information that is most relevant at the current moment for understanding and acting on the world.

A key mechanism for making real-world scene perception tractable is visual attention, the mechanism of preferentially selecting some regions of a scene over others for detailed analysis. In natural scene perception, attentional selection is associated with directing the eyes sequentially through a scene (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson, 2003, 2017; Henderson & Hollingworth, 1999; Land & Hayhoe, 2001; Liversedge & Findlay, 2000; Rayner, 2009; Yarbus, 1967).

The need to select and attend to individual scene regions in order to perceive and understand the contents of those regions is not immediately intuitive. Indeed, the classic change blindness phenomenon in which (for example) an object can completely disappear and reappear in a scene is so striking exactly because it upends our intuitions about what we can visually apprehend all at once (Henderson & Hollingworth, 2003; Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1997). And as "change blindness blindness" demonstrated, most people believe they are seeing and understanding the entire visual scene in front of them even though this is demonstrably false (Levin, Momen, Drivdahl, & Simons, 2000). Contrary to this intuition, typically the eyes must fixate each scene region containing relevant information for the viewer to perceive that region's content, including visual details, identities, and semantic features. It is also typically the case that the local scene regions must be attended for the contents of those regions to be encoded into short- and long-term memory. What we see and understand about the world is in a very real sense determined by where we look (Henderson, 2003).

Given the importance of visual attention for visual perception and cognition, a critical issue concerns understanding the nature of the representations and processes that guide the eyes through a visual scene in real time (Henderson, 2011). Historically the pendulum has tended to swing back and forth between emphasizing the cognitive factors that guide attention and the physical features that guide attention. For example, in a classic textbook study by Yarbus (1967), the parts of a painting that were looked at depended on the question the viewer was trying to answer. When asked about the ages of the people in the picture, the viewer spent more time attending to the faces of those people. When asked to determine the material circumstances of those people, the viewer spent more time on furniture and other artifacts. These results are classically taken to demonstrate that the viewer's task interacts with their general understanding of the image and of the world to guide their attention to relevant image regions.

Although it is widely acknowledged that a perceiver's understanding of a scene and associated stored knowledge play an important role in attentional guidance, recently the majority of research in this field has nonetheless focused on physical image properties and physical salience as the primary drivers of attention. According to what we refer to as *image guidance*, attention is directed to scene regions based on physical image properties generated in a bottom-up manner from the scene (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002a). These models generally propose that attention is controlled by contrasts in primitive image features such as luminance, color, and edge orientation (Treisman & Gelade, 1980; Wolfe, 1994; Wolfe & Horowitz, 2017). A central concept in this theory is the saliency map, which is generated by compiling the local regions of contrast or difference over the primitive features, typically in a winner take all scheme. The saliency map then serves as the basis for attentional control, with attention captured or "pulled" to the currently most visually salient scene region represented by the saliency map (Borji et al., 2013, 2014; Harel, Koch, & Perona, 2006; Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002b). On this view, because the saliency map is generated from primitive features that are semantically uninterpreted, scene regions are prioritized for attentional selection based on physical image properties alone. That is, in this view attentional guidance is at its heart based on a reaction to the physical image properties of the scene. For this reason, we refer to this general class of explanation for attentional guidance as Image Guidance Theory.

It is helpful to note that although "salience" and "saliency map" have traditionally been reserved in the scene attention literature (and the human cognitive and cognitive neuroscience literatures more generally) for image-based saliency of the sort discussed above, it has taken on different interpretations in other literatures, and especially in computer vision, where it now often refers to the output of any model that is used to predict attention regardless of the underlying assumptions and computations. This difference in definitions can lead to confusion. For this reason, it is helpful to be clear about the definition used in this chapter. Here I follow the tradition of reserving the terms "salience" and "saliency map" for concepts in the Koch and Ullman tradition based on the idea that the human visual system computes difference maps from semantically uninterpreted primitive image features that are then combined and used to guide attention (Koch & Ullman, 1985). To highlight this definition, we also sometimes use "image salience" and "physical salience" as synonyms of "salience." We can then contrast theories based on these types of ideas with theories that posit other bases for prioritizing attention over a scene.

1. Cognitive guidance of attention

We can contrast image guidance with cognitive guidance (Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009). Cognitive guidance turns the emphasis back around to the fundamental idea that in complex meaningful scenes, attention is directed to scene regions based on cognitive representations that are activated or retrieved in identification, memory, and semantic systems given the physical scene input, plus the prior and general state of the cognitive system. In this view, visual attention is primarily controlled by the viewer's current interpretation and understanding of a scene's semantic content, including the content of specific objects and local sub-regions of the scene (Buswell, 1935; Yarbus, 1967). When a viewer is engaged in an explicit task or has explicit goals, representations of the task and goals are also cognitively active, and semantically relevant regions are prioritized according to those active goal representations (Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Henderson, 2003, 2007; Henderson & Hollingworth, 1999; Tatler, Hayhoe, Land, & Ballard, 2011; Võ & Wolfe, 2013). For example, when trying to determine the material wealth of the occupants of a room, a viewer's attention will be directed to scene elements such as clothing and furniture by active cognitive representations of the scene interacting with the active goal representation (Yarbus, 1967).

Cognitive knowledge structures relevant for guiding attention can be based on memory systems that encode general concepts and semantic features (semantic knowledge) as well as representations related to specific previously experienced scenes (episodic knowledge), both of which can interact with the viewer's task and goals (Henderson et al., 2007, 2009; Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999). One way to think about the distinction between cognitive guidance and image guidance is that in image guidance, attention is "pulled" to scene regions based on physical image properties, whereas in cognitive guidance, attention is "pushed" by the cognitive system to scene regions based on an internal model of the scene (Henderson, 2007). An important aspect of cognitive guidance is that it emphasizes the central role of meaning and semantics in directing attention. In general, cognitive guidance is consistent with evidence suggesting that viewers attend to semantically informative regions of a scene (Antes, 1974; Buswell, 1935; Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Wu, Wick, & Pomplun, 2014; Yarbus, 1967), as well as regions that are task-relevant (Castelhano, Mack, & Henderson, 2009; Einhäuser, Rutishauser, & Koch, 2008; Foulsham & Underwood, 2007; Hayhoe & Ballard, 2014; Neider & Zelinsky, 2006; Rothkopf, Ballard, & Hayhoe, 2007; Tatler et al., 2011; Torralba, Oliva, Castelhano, & Henderson, 2006; Turano, Geruschat, & Baker, 2003; Yarbus, 1967).

2. Cognitive guidance, cognitive relevance theory, and the flat landscape

Henderson et al. (2009) proposed a general theoretical framework for attention in scenes based on cognitive guidance that completely removed the concept of an image-based saliency map and replaced it with knowledge-based cognitive guidance operating over an alternative visuo-spatial representation of potential attentional targets. In this view, which we have referred to as the cognitive relevance framework (Henderson et al., 2009), scene regions, typically objects (Nuthmann & Henderson, 2010), are prioritized for attention based on cognitive knowledge structures interacting with task goals as described above. Critically, in contrast to the key assumption of image guidance, in this framework potential saccade targets generated from the visual stimulus are not ranked for priority according to their physical salience, but instead are ranked by cognitive relevance based on currently activated cognitive representations interacting with current goals. Of course, the scene image in this model still plays a central role in two important ways:

It serves as input for activating cognitive knowledge structures in semantic and episodic memory, and it is used to generate a representation of (unranked) potential attentional targets. However, critically, image feature contrasts do not directly provide the basis for attentional selection and guidance as they do in image saliency models.

To make this framework more concrete, we can unpack these two ideas. First, scene locations are prioritized for attention on the basis of their potential cognitive relevance, rather than on the basis of their visual salience. Cognitive relevance is based on semantic knowledge about the type of scene in view, episodic knowledge about that particular scene if it has been experienced previously, and current scene interpretation and understanding (Henderson & Ferreira, 2004). This information will be available, both from initially generated semantic and spatial representations (i.e., gist) and from more detailed representations of local scene regions and objects that have already been attended and interpreted. These representations interact with active goals related to knowledge of the task to assign priority for attention. At the same time, the visual scene is parsed to generate a visuospatial representation that explicitly codes potential attentional targets. This representation can be thought of as a "flat landscape" (in contrast to a peaked salience map) in that the potential attentional targets are not yet ranked for attentional priority (Henderson et al., 2007). Attentional target ranking is assigned to the flat landscape of potential targets based on cognitive representations, with attention then directed to those potential targets based on their ranking of cognitive relevance.

What role does physical image salience play in this framework? The image properties associated with image salience are computed early in the visual system and contribute to the parse of the scene into perceptual objects and regions, and to the generation of the visuospatial representation of the scene and its regions over which attention can be directed. Regions are more likely to be included in the visuospatial representation if they are more different from their surround (i.e., are more salient). Critically, though, the proposal is that priority ranking for attention is not based on image salience, but instead on cognitive representations.

In summary, according to cognitive guidance, the scene image leads to activation and retrieval of higher-level cognitive representations, interpretations, and goals which are combined in a cognitive model, and this model is the basis for ranking potential attentional targets. Cognitive guidance theory therefore predicts that the spatial distribution of semantic features in a scene plays a key role in guiding attention.

3. Investigating cognitive guidance: Meaning maps

In the last several decades, a large percentage of the research on attentional guidance in scenes has been (and to a large extent, continues to be) motivated by ideas related to image guidance, with far less work on cognitive guidance and almost no work on the role of the semantic content of the scene on attentional guidance. For example, as of July 2020, Google Scholar returned 98,000 titles given the search term "saliency map," with over 6200 in the first 6 months of 2020 alone. This is a surprising state of affairs given the centrality of semantic understanding in scene perception. After all, when we look at the world, we do not simply perceive a constellation of image features or a gradation of salience. Instead, we perceive a meaningful world that includes individual objects and their semantic content (those are roses and they probably smell great), higher level concepts (this is a garden), and a web of spatial relationships (the roses look great next to the trellis). Why then has the emphasis in attention been so focused on image salience? There are likely several reasons (Henderson, 2017). One appeal of saliency models of this type is that visual salience is neurobiologically plausible in the sense that the visual system is known to compute the assumed primitive features early in visual analysis. In addition, beginning with the introduction of early formative saliency models, image saliency has been relatively easy to quantify and compute, providing quantitative predictions about attentional priority (Borji & Itti, 2013; Harel et al., 2006; Itti & Koch, 2001; Itti et al., 1998; Parkhurst et al., 2002b; Torralba et al., 2006). These models can take a complex image and generate a saliency map without human intervention. That is, saliency maps are image computable. This has made the study of image salience tractable in both behavioral research and in neuroscience (Henderson, 2007, 2017). In contrast, it has been far more difficult to generate quantitative or computational models of scene semantics, a likely reason that saliency models have been more popular (Henderson, 2007, 2017). This difference has made it difficult to experimentally compare the influence of image salience and cognitive models, because that would require representing both physical salience and cognitive content in a format that allows for comparable quantitative predictions of attentional priority across a scene. Given this difficulty, studies of cognitive-based models of attention have typically focused on one or at most a small number of specific scene regions or objects whose meaning in the context of the scene can be measured and manipulated (Brockmole & Henderson, 2008;

De Graef, Christiaens, & d'Ydewalle, 1990; Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009). However, these types of studies do not allow a direct comparison of image salience and meaning across the entire scene.

Much of the recent research in my lab has therefore been focused on addressing this challenge: If we want to compare image guidance and cognitive guidance theories of attention, how can we generate and represent the spatial distribution of semantic features likely to be encoded in a cognitive model of a scene in a format that supports direct comparison with an image saliency map? To address this question, we took inspiration from two classic scene viewing papers (Antes, 1974; Mackworth & Morandi, 1967) as well as our own previous work (Henderson et al., 2007) and introduced a new approach based on what we call *meaning maps* (Henderson & Hayes, 2017). The key idea of a meaning map is that it represents the spatial distribution of the semantic features in a scene in the same format as a saliency map represents the spatial distribution of physical salience.

Meaning maps are based on crowd-sourced responses given by large numbers of naïve subjects who rate small scene patches based on their semantic attributes. Specifically, for our standard context-free ratings, photographs of real-world environments are divided into dense arrays of objectively defined circular overlapping patches at two spatial scales. The spatial scales and numbers of patches that we use are based on simulations showing that ground truth visual properties of scenes can be recovered from them (Henderson & Hayes, 2017). Large numbers of raters each rate a randomly selected subset of individually presented patches taken from the entire set of scenes to be rated. Rating questions can vary as desired; our baseline question asks people to rate patches based on how informative and recognizable they are. Meaning maps are constructed for each scene by averaging the ratings by pixel over patches and raters and smoothing the results. Critically, the ratings show that meaning is spatially distributed non-uniformly across scenes, with some scene regions relatively high in semantic content and other regions relatively low. Meaning maps represent this spatial distribution of semantic content pixel by pixel, and so offer a foundation for directly comparing the relative roles of meaning and image salience on attentional guidance. In the same way that image saliency maps generate predictions concerning attentional priority that can be tested against eye movements or other measures of attention, so too can meaning maps. This allows contrasting predictions from scene semantics and image salience to be directly compared for the same scenes and viewers (Henderson & Hayes, 2017).

Meaning maps and saliency maps provide predictions about attentional priority: Which regions of a scene are likely and unlikely to be attended? The critical empirical question we can then ask is: How well do these predictions match observed spatial distributions of attention produced by people when they view scenes? Following common practice in the scene attention literature, we operationalize the spatial distribution of attention over a scene as an attention map derived from fixation density. In this way, attention maps reflect the spatial distribution of eye fixations across the scene, with some scene regions receiving relatively more and some regions relatively fewer fixations. Attention maps can either be unweighted, or weighted by the duration of each fixation. Notably, attention maps represent attention in the same format and on the same scale as meaning and saliency maps. This allows meaning and saliency maps to be directly assessed against attention maps. In this way, the degree to which the spatial distribution of meaning and salience predict the distribution of attention over scenes can be determined.

An important challenge when comparing meaning maps and saliency maps is that it is highly likely that image salience and semantic content are correlated in scenes (Henderson et al., 2007). An important implication of this finding is that previous results demonstrating a relationship between saliency maps and attention cannot be taken directly as evidence for a functional role of salience in guiding attention. As will become clear below, in all of the studies reviewed in this chapter, the correlation between meaning maps and saliency maps is taken into account in the analyses.

4. Review of meaning map results

In an initial study introducing meaning maps, Henderson and Hayes (2017) asked people to view a set of photographs of real-world environments (i.e., scenes) while their eye movements were recorded. Subjects were given two viewing tasks, a memorization task in which they were asked to prepare for a memory test that would be given after the viewing session, and an aesthetic judgment task in which they were asked to indicate how much they liked each scene. In the analysis, attention maps were produced from viewer fixations, and these attention maps were then compared to meaning and saliency maps. The results showed that both meaning and image salience were correlated with attention, but that when the correlation between meaning and salience was statistically controlled, only meaning accounted for unique attentional variance. In other words, meaning accounted for all the variance accounted for by salience, plus additional variance, whereas saliency account for only a subset of the variance in attention accounted for by meaning and no additional variance. This result was found in both the scene memorization and the aesthetic judgment tasks. Given the strong observed correlation between meaning and salience, and the finding that only meaning accounted for variance in attention when that correlation was controlled, we concluded that meaning was the main factor guiding attention through the scenes in this study.

In this first study, we focused on the spatial distribution of attention by measuring the density of fixations across each scene (Henderson & Hayes, 2017). However, fixations also vary in durations, and this variation reflects visual and cognitive processes (Glaholt & Reingold, 2012; Henderson, Nuthmann, & Luke, 2013; Henderson & Pierce, 2008; Henderson & Smith, 2009; Luke, Nuthmann, & Henderson, 2013; Nuthmann, Smith, Engbert, & Henderson, 2010; van Diepen, Ruelens, & d'Ydewalle, 1999). Therefore, we reanalyzed the data from this study to see if the results would differ when we explicitly included fixation durations (attentional dwell time) at each location (Henderson & Hayes, 2018). This required generating attention maps in which each fixation was weighted by its duration. In these maps, fixations longer in duration were weighted more heavily. We then asked whether the meaning maps or saliency maps best matched the duration-weighted attention maps. Using the same analysis methods as we used for the unweighted analyses in Henderson and Hayes (2017), we replicated all of the critical data patterns that we observed in the original study (Henderson & Hayes, 2018). Once again, both meaning and salience were associated with attention, but when the shared variance was partialled out, only meaning accounted for unique variance in attention. In sum, both when attention maps were based only on location and when they included fixation duration, the answer was the same: only meaning maps uniquely predicted attention.

4.1 Scene description tasks

In our initial investigation comparing meaning maps to image saliency maps, subjects were engaged in memorization and aesthetic judgment tasks (Henderson & Hayes, 2017, 2018). In those tasks, subjects gave their responses after they viewed the scenes, either right after each scene (aesthetic judgment) or at the end of the session (memorization). It may be that under these conditions, subjects are not strongly motivated to direct their attention as carefully or under as much control as they would in a task that requires greater real-time scene engagement. Perhaps in a more on-line task, image saliency would play a greater role. To test this possibility, we investigated how well meaning and image salience account for attention when people are actively engaged in tasks that required them to respond to each scene continuously in real time. For this purpose, in collaboration with Gwendolyn Rehrig and Fernanda Ferreira, we developed a scene description task (Henderson, Hayes, Rehrig, & Ferreira, 2018).

The idea behind scene description is that language production is incremental: People interleave planning and talking rather than planning an entire utterance before beginning to speak (Ferreira & Swets, 2002). This means that in a scene description task, people typically plan and produce small units of speech (for example, words and phrases) that are tied to each scene region as that region is attended. Scene description can therefore be used to examine how semantic information and image salience are related to attention when attention to specific scene regions is functional and necessary for the task in real time. We used this basic scene description paradigm in two experiments. First, in an action description experiment, subjects described what someone might do in each scene. Second, in a general scene description experiment, subjects simply described each scene however they liked. In both experiments, the scene was presented for 30 s, and subjects began their description when the scene first appeared and continued talking while it was in view. Their eye movements were also recorded during this entire time. The eyetracking data were then analyzed in the same way as they were in our first two studies. The results again showed that both meaning and salience were associated with the spatial distribution of attention, but critically, when the correlation between meaning and salience was statistically controlled, only meaning accounted for unique variance in attention. This result was seen in both the action description and general scene description experiments.

4.2 Free viewing and contextualized maps

In the literature on the influence of image salience on attention in scenes, the preferred task has been free viewing, with no specific viewing task given to viewers, and eyetracking data from free viewing has typically been used to benchmark saliency models (Itti et al., 1998; Parkhurst et al., 2002a, 2002b). In contrast, our experiments comparing meaning and image salience reviewed so far were based on directed viewing tasks with explicit viewing instructions.

Perhaps these tasks are biased toward semantic features, and image saliency would dominate in free viewing. To test this hypothesis, Candace Peacock and colleagues conducted a free viewing experiment that was otherwise similar to our previous experiments (Peacock, Hayes, & Henderson, 2019b). If free viewing is meaning-neutral, then perhaps an image saliency advantage would appear under those instructions. In the experiment, subjects were asked to view each scene "naturally, as they would in their daily lives," and were not required to provide any response either during or after viewing.

The results were once again clear cut: Meaning was a better predictor of the spatial distribution of attention than image salience, with meaning accounting for substantial additional unique variance over salience but salience accounting for no unique variance over meaning. Consistent with Henderson and Hayes (2018), this conclusion did not change depending on whether fixations were unweighted or were weighted by fixation duration (Peacock et al., 2019b).

4.3 Is meaning mandatory? Scene viewing tasks with irrelevant meaning

So far, across the five tasks in which we compared the influence of meaning and image salience on attention in scenes (memorization, aesthetic judgment, scene description, action description, and free viewing), meaning was potentially relevant to accomplishing the task. Perhaps all of these tasks more or less bias attention toward meaning, but if we could explicitly make saliency relevant and meaning irrelevant, salience would better predict attention. To test this hypothesis, we extended our study of attention and examined the role of meaning and image salience in three experiments in which meaning was completely irrelevant. In one study, meaning was irrelevant and saliency was relevant for performing the task, and in the second study, neither meaning nor saliency were relevant.

In the first study, as part of her PhD research, Candace Peacock conducted two experiments designed to focus viewers on image properties and to induce them to ignore scene semantics (Peacock, Hayes, & Henderson, 2019a). In one experiment, subjects were asked to rate each scene for its overall brightness. Note that in this brightness rating task, attending to brightness was critical to accomplishing the task, whereas attending to meaningful scene regions was completely irrelevant. Brightness is a content-free visual feature that is central to saliency computations. In a second experiment, subjects were asked to count the number of bright areas in each scene. Here again, attending to the relative brightness of areas in a scene was critical to performing the task, but attending to meaningful scene regions was completely irrelevant.

The reasoning behind these experiments was this: If, in the previous experiments, attention was guided by meaning because the tasks used in those experiments emphasized the semantic content of the scenes, then the relationship between meaning and attention should no longer hold in the new tasks that focused on brightness, a physical feature of the scene images. On the other hand, if the guidance of attention by scene semantics is a fundamental property of the attention system, then the strong relationship between the spatial distribution of meaning and the spatial distribution of attention should continue to hold even when meaning is irrelevant to the viewing task.

The striking finding was that in both the brightness rating and brightness search experiments, the results were very similar to those of the prior experiments: When the correlation between meaning and salience was controlled, only meaning uniquely accounted for significant variance in attention. These results demonstrated that the relationship between meaning and attention is not restricted to viewing tasks that bias the attentional system toward meaning. It appears that meaning is used to guide attention in scenes even when meaning is irrelevant to the task, supporting theories in which scene semantics play a central and perhaps mandatory role in setting attentional priority in scenes.

Perhaps any scene viewing task that requires engaging with a scene, including evaluating a scene-dependent feature or image property of the scene itself such as brightness, mandatorily leads to control of attention by scene semantics. Could it be that a task in which a scene is present but all aspects of that scene are irrelevant to the task would show a stronger effect of image salience and a lesser effect of meaning on attention? To test this idea, Taylor Hayes conducted a third experiment using a scene-independent visual search paradigm in which scene semantics and image salience were both unrelated to the search (Hayes & Henderson, 2019b). In this task, subjects were asked to search for letter targets that were randomly superimposed on top of scenes. The target letters were sufficiently difficult to find that subjects had to look for them carefully with eye movements. Importantly, in the 40 critical scenes that were analyzed, the letter targets were absent so that subjects would have to search throughout the trial. The absence of targets in the critical scenes also meant that we did not have to worry about any potential effects of the target letters themselves on attention.

The data were analyzed in the same way as the previous experiments, and the main result was similar. Once again, the spatial distribution of attention remained more strongly correlated with meaning than with image-based saliency, and when the correlation between meaning and salience was controlled, only meaning uniquely accounted for significant variance in attention. This result was observed even though the spatial distribution of meaning across a scene was completely irrelevant to the letter search task, which could be accomplished by completely ignoring the scene.

In sum, the results from these three experiments support the view that scene semantics play a central role in setting attentional priority in scenes. Furthermore, it appears that meaning is used to guide attention in scenes even when meaning is irrelevant to the task, consistent with the idea that the extraction and use of scene semantics in setting attentional priority is mandatory when an interpretable real-world scene is present.

4.4 Verbal encoding

We often use language to encode visual information in working memory. Scene meaning is likely easier to encode verbally than image salience. Could it be that the advantage of scene meaning over image salience in accounting for attention is due to verbal encoding of the scene? To test this hypothesis, Gwendolyn Rehrig and colleagues conducted two experiments in which people were asked to do a secondary articulatory suppression task at the same time that they viewed scenes (Rehrig, Hayes, Henderson, et al., 2020). In the first experiment, articulatory suppression was manipulated between-subjects, and the verbal suppression task was to repeat a different sequence of three digits aloud while viewing each scene. In the second experiment, articulatory suppression was manipulated within-subject in a block design, and the suppression task was to repeat a different sequence of the names of three simple shapes aloud while viewing each scene. The viewing task was to prepare for a later scene memory test. As usual, meaning maps and saliency maps were generated for each of the scenes.

The logic of these experiments was this: If verbal encoding mediates the relationship between meaning and attention, then meaning should explain greater (especially unique) variance in attention than image salience in the control no-suppression conditions but not in the suppression conditions. On the other hand, if verbal encoding does not underlie the meaning advantage over image salience, then meaning should explain greater (and especially unique) variance in attention over salience whether or not subjects are engaged in a verbal suppression task.

The results were clear. Once again, in both experiments, meaning explained more of the variance in visual attention than image salience did, and only meaning explained unique variance in attention, both with and without the suppression tasks. The results offer no support for the hypothesis that the meaning advantage over image salience in accounting for attention during scene viewing is due to verbal encoding of the scenes.

4.5 Contextualized meaning maps

In the experiments discussed so far, local scene meaning was based on human ratings of individual scene patches presented to raters independently of full scenes. These experiments therefore focused on the role of what we have called scene-intrinsic context-free meaning (Hayes & Henderson, 2019a, 2019b). By scene-intrinsic, we mean that the meaning derives from the contents of the scene alone and not from the contents in interaction with particular tasks or goals. By context-free, we mean that the meaning derives from each local region without regard to the overall scene meaning. We constructed initial meaning maps in this way because we were interested in investigating the role of local scene meaning on attention rather than the relationship between local semantic content and the rest of the scene. We also did not want to bias comparisons with image salience by including contextual information that saliency map models do not know about.

However, the meaning of an object or local scene element is often influenced by the context in which that object appears (Henderson et al., 1999; Loftus & Mackworth, 1978; Spotorno, Tatler, & Faure, 2013; Võ & Henderson, 2009). It could be that meaning in the context of the scene (which we refer to as contextualized meaning) is more related than context-free meaning to how attention is distributed in a scene. To examine this hypothesis, Candace Peacock and colleagues generated new contextualized meaning maps from meaning ratings for scene patches that were presented along with their full scenes (Peacock et al., 2019b). Contextualized meaning maps were generated using the same patches and the identical method as the context-free meaning maps, except that the raters were asked to rate how informative or recognizable each patch was in the context of the larger scene. To aid this rating decision, each patch was circled in green in the context scene. We then compared attention maps from free viewing to the original context-free and the new contextualized meaning maps. The important result was that both types of meaning maps produced very similar results; neither predicted attention better than the other, and both predicted substantial unique variance over image salience. This convergence in results suggests that the context-free meaning maps are generally a good representation of the spatial distribution of local meaning across a scene. These results also suggest that the context-free meaning maps do not lose much critical semantic information despite the fact that sometimes only parts of large objects and scene regions are shown in the rated patches. (In the contextualized ratings, the larger objects are visible in the context scene image.) Interestingly, the results suggest that a good deal of the variance in the distribution of attention across a scene can be captured by local semantic features without consideration of how those features relate to the rest of the scene.

4.6 Early vs late meaning effects

Perhaps image salience is more likely to guide attention early in viewing, when a scene first becomes visible, but is subsequently overridden or suppressed over time as meaning comes on-line to guide attention (Anderson, Donk, & Meeter, 2016; Anderson, Ort, Kruijne, Meeter, & Donk, 2015; Henderson & Hollingworth, 1999; Parkhurst et al., 2002a, 2002b). This hypothesis predicts that very early in scene viewing, before suppression can be applied, attentional guidance by salience should be stronger, and that later, when sufficient time is available for suppression to operate, it should be weaker or absent. This possibility is important theoretically because it would suggest that image-based saliency maps are in fact generated for scenes, but that they are then inhibited as cognitive representations become active and cognitive guidance becomes dominant. To test this hypothesis, in each of the studies reviewed here, sub-analyses were performed in which the earliest fixations were examined to determine whether salience had a dominant or at least larger effect compared to meaning for the first few fixations in scenes. In every case, the results were very similar to the overall pattern: meaning and salience were both associated with the spatial distribution of attention, but when the correlation between meaning and salience was statistically controlled, only meaning accounted for unique variance in attention even in the earliest fixations, including the very first viewer-generated fixation. These results are not consistent with the hypothesis that image salience plays a more important role early in scene viewing.

4.7 Center bias

Viewer center bias refers to the common finding that human subjects tend to spend more time looking near the center than the periphery of a scene image

(Tatler, 2007). *Image center bias* refers to the possibility that more information is to be found at the center of an image than the periphery. For example, a photographer may place the most visually salient or meaningful information near the center of a photograph. The majority of influential image saliency models therefore include significant *model center bias* to account for viewer center bias (Bruce, Wloka, Frosst, Rahman, & Tsotsos, 2015). Because most image saliency models either add model center bias or are built to generate model center bias to account for viewer center bias, model predictions are typically generated from a combination of the model's built-in core assumptions and the added model center bias. It is therefore instructive to ask how well saliency models are able to predict attention based on their core assumptions separated from their center biases, and how well those core assumptions do compared to meaning maps.

To answer these questions, Taylor Hayes compared how well three influential and widely cited image saliency models (Itti and Koch saliency model with Gaussian blur, Itti et al., 1998; Koch & Ullman, 1985; Harel et al., 2006; graph-based visual saliency (GBVS) model, Harel et al., 2006; and attention by information maximization (AIM) saliency model, Bruce & Tsotsos, 2009) predicted attention compared to their model center biases (Hayes & Henderson, 2019a). These image saliency models were chosen because they represent three approaches for using bottom-up low-level image features to generate saliency maps. They also incorporate different types and different amounts of model center bias. For example, the two standard saliency models (Itti and Koch with blur, and GBVS) include substantial model center bias, whereas AIM includes substantially less model center bias. These models were also compared to context-free meaning maps.

The basic approach was to separate the center bias of each model from its core model predictions. This was accomplished by running each model over a large independent set of scene images (the MIT-1003 benchmark data set) that were not used in the attention experiments, and then to extract the general center bias that each model produced across this entire set of scenes. Each model's core assumptions (model predictions without its center bias) and model center bias were then assessed separately against attention data from scene viewing. This analysis used the attention data from five eye movement experiments over three viewing tasks: scene memorization, aesthetic judgment, and visual search. Each of these tasks produces different amounts and patterns of viewer center bias, providing a varied data set to compare to the saliency models.

We found that for viewing tasks that produced significant viewer center bias (memorization and aesthetic judgment), the image saliency models actually performed significantly worse (on average accounted for 23% less variance) than their center biases alone, whereas meaning maps performed significantly better (on average accounted for 10% more variance) than center bias alone in all three tasks. This is a striking finding because it suggests a large part of the predictive work in saliency models is done by their center biases and not by their core assumptions. In sum, these results suggested that when viewer center bias is present, adding low-level image feature saliency actually accounts for attention less well than a model simply based on center bias.

5. Conclusion

In this chapter I have summarized the idea that visual attention may be driven more by the semantic content of a visual scene than directly by its image features. I have also reviewed the meaning map approach to investigating this issue, and shown how we can use it to compare the roles of semantic features and image features (image salience) on attention during scene viewing. The reviewed results strongly support the hypothesis that meaning plays a fundamental and mandatory role in attentional guidance in real-world scenes.

An advantage of saliency maps compared to meaning maps is that they are image computable: They can be derived automatically (without human intervention) by computational models. On the other hand, meaning maps are not image computable: They require human raters. For this reason, saliency models might be more satisfying. However, in our view, the interesting psychological claim of the image salience hypothesis is independent of its computability. This claim is that human attention in real-world scenes is driven by contrasts in basic semantically uninterpreted image features. The alternative hypothesis we have proposed is that image salience effects are actually meaning effects due to the fact that image features and semantic features in scenes are correlated (Henderson & Hayes, 2017). And as reviewed above, this is what we have found over many experiments, with very little if any unique variance accounted for by image salience once the variance accounted for by semantic features has been controlled. This comparison is independent of computability, and instead is based on the fundamental theoretical assumptions of the two competing psychological theories.

That is, we are concerned with psychological principles, not modeling. Though there are certainly advantages to instantiating theoretical assumptions as a working model, there is no logical requirement that testing psychological theories of attention requires image computability.

It is important to highlight that meaning maps do not provide a theory of scene semantics. They are simply a method for tapping into and representing human judgments concerning how the semantic features in a scene are distributed across space. That is, meaning maps provide a representation of the spatial distribution of meaning across a scene, but do not offer direct insight into the nature of scene semantics or how they are represented in the mind and brain. That said, the meaning map approach may be useful as a tool for getting a handle on the nature of scene semantics. For example, by changing the nature of the rating question, we might be able to create semantic feature maps that represent more specific targeted aspects of scene semantics. These maps could then be compared to the distribution of attention to determine how well they match the attended areas of a scene. The contextualized meaning maps discussed in this chapter provide one example of this approach in which we asked whether meaning maps based on scene context differ from those that are context-free. As another example, we have recently investigated "grasp maps" that represent whether an object can be manipulated, a particular semantic feature of scene space. We compared grasp maps to general meaning maps and image saliency maps and found that both meaning maps and grasp maps were predictive of attention (Rehrig, Peacock, Hayes, Henderson, & Ferreira, 2020).

Similarly, we might consider other scene-intrinsic semantic features, as well goal-related semantic features that are based in part on their relationship to the viewer's goals. In this way we might be able to unravel how different types of meaning are related to each other and to performance over different perceptual and cognitive tasks. The meaning map approach provides a tool for pursuing these questions. The meaning maps we have investigated so far based on context-free ratings may not be the type of meaning most associated with attention, and we may therefore be underestimating the relationship between semantic features and attention.

Funding

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.

Acknowledgement

We thank the members of the UC Davis visual cognition group for their feedback on this work.

Conflicts of interest

The authors declare no conflicts of interest.

References

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Psychonomic Bulletin & Review*, *23*(6), 1794–1801.

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 9.

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

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89.

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

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69. https://doi.org/10.1109/TIP.2012.2210727.

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.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 5–5.

Bruce, N. D., Wloka, C., Frosst, N., Rahman, S., & Tsotsos, J. K. (2015). On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, *116*, 95–112.

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

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6.1–15.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317–329. https://doi.org/10.1007/BF00868064.

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., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57–84. https://doi.org/10.1006/jmla.2001.2797.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception*, *36*(8), 1123–1138. https://doi.org/10.1068/p5659.

Glaholt, M. G., & Reingold, E. M. (2012). Direct control of fixation times in scene viewing: Evidence from analysis of the distribution of first fixation duration. *Visual Cognition*, *20*(6), 605–626.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, *19*, 1–8. https://doi.org/10.1.1.70.2254.

Hayes, T. R., & Henderson, J. M. (2019a). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, *82*, 985–994. https://doi.org/10.3758/s13414-019-01849-7.

Hayes, T. R., & Henderson, J. M. (2019b). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin & Review*, *26*(5), 1683–1689. https://doi.org/10.3758/s13423-019-01642-5.

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

Hayhoe, M. M., & Ballard, D. (2014). Modeling task control of eye movements minireview. *Current Biology*, *24*(13), R622–R628. https://doi.org/10.1016/j.cub.2014.05.020.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 49–63. https://doi.org/10.1167/3.1.6.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11), 498–504. 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.), *Vol. 2015. The oxford handbook of eye movements* (pp. 593–606). Oxford; New York: Oxford University Press. Issue January.

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., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Elsevier Ltd. https://doi.org/10.1016/B978-008044980-7/50027-6

Henderson, J. M., & Ferreira, F. (2004). *Scene perception for psycholinguists*. https://doi.org/10.4324/9780203488430.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747. https://doi.org/10.1038/s41562-017-0208-0.

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*, 10.

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., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271. https://doi.org/10.1146/annurev.psych.50.1.243.

Henderson, J. M., & Hollingworth, A. (2003). Global transsaccadic change blindness during scene perception. *Psychological Science*, *14*(5), 493–497. https://doi.org/10.1111/1467-9280.02459.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856.

Henderson, J. M., Nuthmann, A., & Luke, S. G. (2013). Eye movement control during scene viewing: Immediate effects of scene luminance on fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 318–322. https://doi.org/10.1037/a0031224.

Henderson, J. M., & Pierce, G. L. (2008). Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin and Review*, *15*(3), 566–573. https://doi.org/10.3758/PBR.15.3.566.

Henderson, J. M., & Smith, T. J. (2009). How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition*, *17*(6–7), 1055–1082. https://doi.org/10.1080/13506280802685552.

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.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227. https://doi.org/10.1007/978-94-009-3833-5_5.

Land, M. F., & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25–26), 3559–3565. https://doi.org/10.1016/S0042-6989(01)00102-X.

Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change blindness blindness: The metacognitive error of overestimating change-detection ability. *Visual Cognition*, *7*(1–3), 397–412. https://doi.org/10.1080/135062800394865.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*(1), 6–14.

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.

Luke, S. G., Nuthmann, A., & Henderson, J. M. (2013). Eye movement control in scene viewing and reading: Evidence from the stimulus onset delay paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 10–15. https://doi.org/10.1037/a0030392.

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.

Neider, M. B., & Zelinsky, G. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621.

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

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *117*(2), 382–405. https://doi.org/10.1037/a0018924.

Parkhurst, D., Law, K., & Niebur, E. (2002a). Modelling the role of salience in the allocation of visual selective attention. *Vision Research*, *42*(1), 107–123.

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

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019a). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*(1), 20–34. https://doi.org/10.3758/s13414-018-1607-7.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). 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.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Rehrig, G., Hayes, T. R., Henderson, J. M., et al. (2020). When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention. *Memory and Cognition*. https://doi.org/10.3758/s13421-020-01050-4.

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*. Advance online publication. https://doi.org/10.1037/xlm0000837.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*, 368–373.

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

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267.

Spotorno, S., Tatler, B. W., & Faure, S. (2013). Semantic consistency versus perceptual salience in visual scenes: Findings from change detection. *Acta Psychologica*, *142*(2), 168–176.

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

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), 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.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136. https://doi.org/10.1016/0010-0285(80)90005-5.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, *43*(3), 333–346. https://doi.org/10.1016/S0042-6989(02)00498-4.

van Diepen, P., Ruelens, L., & d'Ydewalle, G. (1999). Brief foveal masking during scene perception. *Acta Psychologica*, *101*(1), 91–103.

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), 24.1–15. https://doi.org/10.1167/9.3.24.

Võ, M. L. H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, *126*(2), 198–212. https://doi.org/10.1016/j.cognition.2012.09.017.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. https://doi.org/10.3758/BF03200774.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 1–8. 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*(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.