

Regarding Scenes

John M. Henderson

University of Edinburgh, Edinburgh, Scotland

ABSTRACT—*When we view the visual world, our eyes flit from one location to another about three times each second. These frequent changes in gaze direction result from very fast saccadic eye movements. Useful visual information is acquired only during fixations, periods of relative gaze stability. Gaze control is defined as the process of directing fixation through a scene in real time in the service of ongoing perceptual, cognitive, and behavioral activity. This article discusses current approaches and new empirical findings that are allowing investigators to unravel how human gaze control operates during active real-world scene perception.*

KEYWORDS—*scene perception; real-world scene; visual saliency; eye movements; gaze control; visual context*

It has been known for at least 30 years that the gist of a scene can be apprehended very rapidly, well within the duration of a single fixation or period of relative gaze stability (Potter, 1976). It has been known for even longer that viewers tend to move their eyes through scenes when they look at them (Buswell, 1935; Yarbus, 1967; see Fig. 1). Given fast gist understanding, why do we bother to move our eyes? Recent studies of change detection, object identification, and scene memory show that close or direct fixation is necessary to perceive local visual details, to unambiguously identify objects, and to encode object and scene information into short- and long-term memory (Henderson & Hollingworth, 1999; Hollingworth & Henderson, 2002). What we see and understand about the visual world is tightly tied to where our eyes are pointed.

Why is fixation critical to these perceptual and cognitive processes? First, high-resolution visual information is acquired from only a very limited region of the scene surrounding the fixation point, with visual quality falling off precipitously and continuously from central vision into a low-resolution visual surround. The high resolving power of central vision is partly a consequence of the optical and anatomical structure of the eye and retina. Also, “cortical magnification” preferentially maps

central vision onto the visual cortex, ensuring that more computational power is devoted to fixated regions. Therefore, to acquire high-quality visual input from a scene region, fixation must be directed to it. Second, there is a very tight link between attention and fixation. Although visual-spatial attention can be dissociated from where the eyes are fixated in laboratory demonstrations, attention is typically directed to the fixated location and the location to be fixated next. Attention is time-locked to eye-movement dynamics as eye-movement control circuits hold fixation and then release the eyes to the next fixation site. This time-locked relationship between shifts in attention and gaze appears to be mandatory, in part due to the tight neural integration of systems that control covert attention and those that control eye movements.

Given the importance of fixation for perceptual and cognitive processing during scene perception, a critical issue concerns the representations and processes that govern where and for how long the eyes are directed to a particular scene region. This issue has become the focus of intense investigation in the past few years, with recent research emphasizing two general classes of factors that may drive gaze: bottom-up image properties, and cognitive knowledge structures used in a top-down manner.

WHERE DO WE LOOK?

In a sense, we can think of fixation as either being “pulled” to a particular scene location by the visual properties at that location, or “pushed” to a particular location by cognitive factors related to what we know and what we are trying to accomplish. Intuitively, a bright or colorful area of a scene might attract the eyes in a bottom-up manner. At the same time, a viewer might want to look at scene regions that are relevant given current tasks and goals, whether or not those regions are visually prominent. A good deal of research has been devoted to studying whether fixation is pulled by the stimulus or pushed by cognitive processes. An initial emphasis on stimulus control was motivated in part by research in attention suggesting that image differences (e.g., a red patch among green patches) perceptually “pop out.” Furthermore, because image properties are easier to model than knowledge structures, initial attempts to generate computational models and quantitative predictions of human eye movements have tended to focus on bottom-up input.

Address correspondence to John M. Henderson, Department of Psychology, 7 George Square, University of Edinburgh, Edinburgh, EH8 9JZ United Kingdom; e-mail: john.m.henderson@ed.ac.uk.

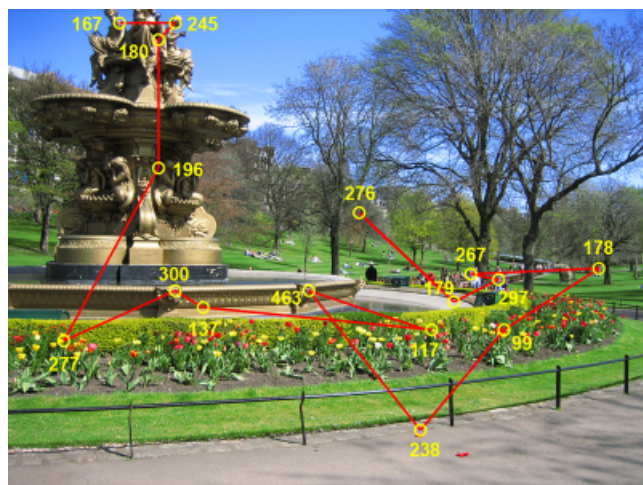


Fig. 1. An example scan pattern over a photograph of a real-world scene, Princes Street Gardens, Edinburgh. Circles represent fixations, lines represent saccades, and the numbers above the circles represent the durations of fixations in milliseconds.

The importance of the “pulling” influence of image properties on the selection of fixation sites has been investigated in two ways. First, the image properties of fixation locations have been analyzed to determine if they differ systematically from nonfixated regions. A correlation between fixation locations and image statistics is typically observed. For example, fixated regions tend to contain more edges than nonfixated regions. Second, computational models of visual saliency have been implemented based on known functional properties of the visual cortex, and these models have been used to predict fixation locations (Itti & Koch, 2001). In this approach, the image gives rise to a representation (a saliency map) that explicitly denotes scene areas that are different from surrounding areas. Image properties considered include intensity, contrast, orientation, color, and motion. The maps generated for each image property over multiple spatial scales are then combined to create a single saliency map. The intuition behind this approach is that a scene region that differs from its surrounding area (e.g., a red patch surrounded by green) is potentially informative and so is worthy of fixation. This intuition also aligns with the pop-out effect. The salient regions in the map are predictions about which scene regions should be fixated, and correlations between model predictions and human fixations have been observed.

Both the scene-statistics and saliency-map approaches rely on correlations rather than experimental manipulation. These types of studies therefore do not allow us to unambiguously conclude that differences in image properties directly cause the eyes to be directed to particular locations. A third (causal) factor, such as the meaning of the fixated region, could correlate with both image statistics and fixation probability, because meaningful objects are likely to differ from scene background in image properties. When we explicitly tested this possibility by asking an independent group of participants to rate how meaningful fixated and nonfixated patches were, we found that fixated re-

gions differed in both their image statistics and their semantic content compared to regions that were not fixated (Henderson, Brockmole, Castelano, & Mack, 2007). In future work, it will be necessary to directly manipulate image statistics, holding semantic content constant, to determine whether visual saliency plays a causal role in driving gaze through a scene. But based on what is already known today, there is good reason to doubt that the “pull” of image properties is the main factor driving eye movements.

Human gaze control is intelligent in that it draws not only on currently available visual input but also on cognitive knowledge structures, including short-term and episodic memory for the current scene; stored long-term visual, spatial, and semantic information about other similar scenes; and the goals and plans of the viewer. This fact is easily illustrated by considering where you would look if you wanted to know the current time. Presumably you wouldn’t look at the brightest or the most colorful thing in your visual field, but rather to a location likely to provide the time (perhaps your wrist, or a clock on the wall). Fixation sites are far less strongly tied to visual saliency when meaningful scenes are viewed during active tasks (Land & Hayhoe, 2001). Even the initial saccades in a scene tend to move the eyes toward the likely location of a search target, whether or not the target is present, because a rapid understanding of the general meaning of a scene and its spatial layout provides important constraints on where a particular object is likely to be found (Castelano & Henderson, in press).

The fact that gaze control draws on stored knowledge structures implies that these structures must somehow be integrated with a representation of the specific image that is currently in view. For example, if you are going to look at a spot on the wall to learn the time, you need to have activated relevant knowledge about where the clock is to be found, and you need to have a visual representation specifying where the wall is and providing a potential visual target to direct your eyes to. An important current issue in gaze control concerns how these types of information are combined. One approach uses knowledge structures to modify the bottom-up saliency map (Rao, Zelinsky, Hayhoe, & Ballard, 2002). For example, if you know what the clock looks like (e.g., a black circular object with a white face), the current scene input might be filtered to produce a saliency map that highlights regions with just those visual properties. Then the eyes could be driven by the saliency map. However, not all knowledge structures can easily be converted to visual properties that can be used to generate a saliency map. For instance, it would be difficult to directly create a saliency map based on knowledge about where a particular type of object might be found, because location is not itself a visual property over which saliency can be computed.

Another approach for combining knowledge structures and image properties is to independently compute separate saliency and spatial-context maps and then combine the two. For example, in the Contextual Guidance model, we proposed that a

context map highlighting scene regions likely to contain a specific class of target object (e.g., a coffee cup) is combined with an independent image-based saliency map (Torralba, Oliva, Castelano, & Henderson, 2006). We demonstrated that our model predicted fixation locations during object search significantly better than did a model based on visual saliency alone. We also observed that when the two were directly contrasted, the context representation alone did a much better job of predicting fixation locations than the saliency representation alone did.

Whether cognitive knowledge directly influences image-based saliency or initially generates a separate context map is an important current topic of research. The answer may depend in part on the nature of the knowledge under consideration. For example, in the case of search that can capitalize on object features, direct modulation of the saliency map might be employed. On the other hand, when knowledge instead provides spatial constraints, then a separate contextual location map might be computed (Torralba et al., 2006). These possibilities are not mutually exclusive.

HOW LONG DO WE LOOK?

In addition to fixation location, another important characteristic of gaze behavior is the observed variability in the durations of individual fixations (Fig. 1). The influence of visual and cognitive factors on fixation duration has been a central focus in the study of reading (Reichle, Pollatsek, Fisher, & Rayner, 1998) but has generally been of less concern in the literature on gaze control in scenes. Conclusions about the distribution of attention over a scene can potentially differ markedly when fixation position is weighted by fixation duration, because the distribution of processing time across a scene is a function of both the spatial distribution of fixations and the durations of those fixations.

Average fixation duration during scene viewing is about 330 milliseconds, but there is substantial variability around this mean both within an individual and across individuals. In reading, much of this variability is controlled by visual and cognitive factors associated with the currently fixated word. In scene perception, it is not yet clear whether fixation durations are controlled by the fixated region. It has been shown that individual fixation durations can be affected by global scene properties such as luminance. Individual fixation durations are also influenced by the nature of the viewing task, with longer durations during scene memorization than when searching through a scene. However, these effects could be due to changes in global fixation duration parameters. For example, eye movements might simply be slowed down overall when the going becomes more difficult, as when we slow our general walking gait in the face of rougher terrain. Alternatively, fixation durations could reflect moment-to-moment changes in visual and cognitive difficulty, reacting in real time to the transitory nature of the visual world.

To investigate this question, we have recently conducted a series of experiments examining whether the durations of individual fixations are controlled directly and immediately by the current stimulus (Henderson & Pierce, 2006). In these experiments, we tracked participants' eye movements as they viewed photographs of real-world scenes, and we turned off the scenes while their eyes were in saccadic movement from one location to another (Rayner & Pollatsek, 1981). Thus the scene was missing at the beginning of the next critical fixation. After a predetermined delay, we turned the scene back on. The duration of the delay was varied, and the influence of the delay on the duration of the critical fixation was measured. We found strong evidence that fixation durations are systematically influenced by currently available stimulus information: When the scene disappeared, the gaze control system often held the fixation until the scene became visible again. This finding provides a compelling demonstration that scene processing during a fixation can produce an immediate and direct influence on the duration of that fixation. It is clear that moment-to-moment attention dynamics are reflected by both fixation location and fixation duration, and a complete model of active scene perception will have to account for both.

CURRENT DIRECTIONS

Humans use knowledge about the world to guide gaze intelligently through a scene. The drive to fixate is so strong that viewers will look to a scene region that had contained an object even when they are aware that the object is no longer present (Altmann, 2004; Richardson & Spivey, 2000). Cognitive systems interact with each other and with the scene image to determine where the eyes fixate. Research on gaze control that combines computational modeling with eye tracking is providing new insights into the neural and cognitive processes that support this behavior.

Fixation also provides a pointer or index (Ballard, Hayhoe, Pook, & Rao, 1997) that anchors cognitive processes such as language understanding to entities in the world (Henderson & Ferreira, 2004; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). For example, viewers will typically look to a scene area that contains an object when that object is mentioned by a speaker. Methodologically, the drive to look at objects as they are mentioned provides an important tool for studying on-line language processing. Theoretically, the study of world-situated language is leading investigators to wonder how linguistic and visual representations are integrated in the service of ongoing cognitive and behavioral activity (Henderson & Ferreira, 2004), as when two people talk about and cooperatively interact with their visual environment. Research in this area is just beginning.

Much of what we know about active scene processing comes from the study of static scene depictions like photographs. Ultimately, depictions are stand-ins for the real environment, and what we really want to know is how active scene perception operates in the world itself. Film, video, and virtual-reality

environments offer an important middle ground for studying dynamic scene perception while still providing critical experimental control. Ultimately, what we see, understand, and remember from the visual world is tightly tied to where we look, and so a comprehensive understanding of human perception and cognition will require an understanding of gaze control in the entire range of situations that humans encounter.

Recommended Reading

- Findlay, J.M., & Gilchrist, I.D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford, England: Oxford University Press.
- Henderson, J.M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7, 498–504.
- Henderson, J.M., & Ferreira, F. (2004). (See References)
- Land, M.F., & Hayhoe, M. (2001). (See References)
-

REFERENCES

- Altmann, G.T.M. (2004). Language-mediated eye movements in the absence of a visual world: The “blank screen paradigm.” *Cognition*, 93, 79–87.
- Ballard, D.H., Hayhoe, M.M., Pook, P.K., & Rao, R.P.N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20, 723–767.
- Buswell, G.T. (1935). *How People Look at Pictures*. Chicago: University of Chicago Press.
- Castelhano, M.S., & Henderson, J.M. (in press). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*.
- Henderson, J.M., Brockmole, J.R., Castelhana, M.S., & Mack, M. (2007). Image salience versus cognitive control of eye movements in real-world scenes: Evidence from visual search. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537–562). Oxford, England: Elsevier.
- Henderson, J.M., & Ferreira, F. (Eds.). (2004). *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.
- Henderson, J.M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, 5, 438–443.
- Henderson, J.M., & Pierce, G. (2006). Direct control of fixation durations during active scene perception. *Visual Cognition*, 15, 108–112.
- Hollingworth, A., & Henderson, J.M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 113–136.
- Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2, 194–203.
- Land, M.F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, 41, 3559–3565.
- Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509–522.
- Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., & Ballard, D.H. (2002). Eye movements in iconic visual search. *Vision Research*, 42, 1447–1463.
- Rayner, K., & Pollatsek, A. (1981). Eye movement control during reading: Evidence for direct control. *Quarterly Journal of Experimental Psychology*, 33A, 351–373.
- Reichle, E., Pollatsek, A., Fisher, D.L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105, 125–157.
- Richardson, D., & Spivey, M. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, 76, 269–295.
- Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.
- Torralba, A., Oliva, A., Castelhana, 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.
- Yarbus, A.L. (1967). *Eye Movements and Vision*. New York: Plenum Press.