CellPress

Opinion

Gaze Control as Prediction

John M. Henderson[1,*]

The recent study of overt attention during complex scene viewing has emphasized explaining gaze behavior in terms of image properties and image salience independently of the viewer's intentions and understanding of the scene. In this Opinion article, I outline an alternative approach proposing that gaze control in natural scenes can be characterized as the result of knowledge-driven prediction. This view provides a theoretical context for integrating and unifying many of the disparate phenomena observed in active scene viewing, offers the potential for integrating the behavioral study of gaze with the neurobiological study of eye movements, and provides a theoretical framework for bridging gaze control and other related areas of perception and cognition at both computational and neurobiological levels of analysis.

Trends

Gaze is directed to task- and goal-relevant scene regions.

Gaze control is based on predictions concerning where specific goal- and task-relevant objects are likely to be found.

Predictions for gaze control are based on knowledge gained from past experience with scenes.

Predictions are likely to draw on memory representations of specific scene instances and general scene categories.

Let's Take a Look

Some 85 years ago, Guy Thomas Buswell established that viewers tend to look at the regions of scenes that are likely to contain the information that is most meaningful and relevant [1] (Figure 1, Key Figure). Similarly, in classic textbook demonstrations Alfred Yarbus showed that where we look in a complex scene is strongly influenced by our current goals and viewing task [2]. In the years since those groundbreaking studies, the role of knowledge in the control of gaze and attention in complex, meaningful scenes has been repeatedly shown. Despite this overwhelming evidence, the literature over the past couple of decades has focused almost exclusively on trying to explain where people look in scenes in terms of image properties alone, ignoring the viewer's understanding of the scene. This tendency to focus on the image rather than the viewer's understanding of the meaning of the scene is likely to be due in part to a version of the principle of the drunkard's search, in which the inebriated driver looks for lost car keys under a streetlamp because that is where the light is. In the case of attention in scenes, image properties fall under the light of the streetlamp; it is far easier to build models that account for where we attend based on image properties than it is to build models based on scene meaning and viewer goals. A model based on image properties requires methods to measure and quantify those properties, a condition that is already within grasping distance given advances in computational vision and visual neuroscience. By contrast, a model based on a full understanding of the scene, its meaning, and its relationship to the viewer's current goals and task requires a relatively complete model of human cognition, and that is still a few years off. From this perspective it makes sense that investigators have tended to focus on the tractable and put off the less tractable for another day.

Clearly, however, the image- or saliency-based approach is fundamentally limited when it comes to accounting for how human viewers direct their attention through real scenes [3–7]. The question then becomes how to make theoretical progress in a way that takes viewer knowledge into account. In recent years a new framework for understanding how the brain thinks and perceives has emerged. This framework conceptualizes the brain as a 'prediction machine' [8], a system that uses past experience to generate expectations or predictions concerning what and where items and events are likely to be encountered next [9–11]. When these predictions are supported by incoming evidence from the environment, all is well and nothing new is learned. When a prediction is partially or completely wrong, however, the knowledge that was used to

[1]Center for Mind and Brain and Department of Psychology, University of California, Davis, 267 Cousteau Place, Davis, CA, USA

*Correspondence: johnhenderson@ucdavis.edu (J.M. Henderson).

generate that prediction has to be updated so future predictions can be more accurate. The upshot is that errors of prediction lead to knowledge gain. For some proposals, only information associated with prediction error is coded from the environment, a process called predictive coding [12–14].

In this Opinion article, I outline the proposal that **gaze control** (see Glossary) in natural scenes can be understood as the consequence of spatial prediction. Specifically, the proposal is that where a viewer looks and attends in a complex scene is the result of a prediction about where the most meaningful and task-relevant information is to be found in that scene. Adopting this view provides an overarching theoretical context for unifying many of the disparate phenomena observed in active scene viewing. For example, it offers potential for integrating the behavioral study of gaze control with the neurobiological study of eye movements more generally [15]. It also provides a theoretical framework that can assist in bridging between gaze control and other related areas of perception, cognition, and motor control. It provides a potential framework for understanding how **overt** and **covert attention** are related. Furthermore, taking this perspective provides potential for integrating the study of language–vision interaction, where linguistic input can determine where to look [16,17] and where we look can influence what we say [18,19].

Here the emphasis is specifically on considering gaze control in terms of spatial prediction, since most research on overt attention in scenes has focused on understanding where people look [3,20–22]. However, a more encompassing theoretical framework would extend this approach to other aspects of gaze control as well, such as explaining the amount of time each scene region is **fixated** as a consequence of predictions about location, identity, and meaning (Box 1).

Glossary

Covert attention: selective processing of a location or object at the expense of other locations or objects via internal processes without a corresponding movement of the eyes.
Fixation: stable eye position directed toward a specific scene location or object used for visual information acquisition.
Fovea: the highest-resolution region of the retina.
Gaze control: the process of directing fixation through a scene in real time in the service of ongoing perceptual, cognitive, and behavioral activity.
Overt attention: selective processing of a location or object at the expense of other locations or objects due to a movement of the eyes.
Saccade: a fast, ballistic eye movement that reorients fixation from one location to another within a scene.

The Prediction Approach to Gaze Control: Examples

Four brief examples serve to illustrate the prediction approach to gaze control and the nature of the phenomena that support it. These examples are meant to be illustrative rather than exhaustive.

Object Search in Scenes

People can find common objects very quickly in complex real-world scenes. In experiments demonstrating this ability, target objects are placed at either an expected location or an unexpected location in each scene [23–28]. Viewers are asked to find those objects as quickly as possible. A typical experiment presents the name (or picture) of the target object for the current trial followed by the scene, which may or may not contain the object. The key finding is that viewers typically find objects effortlessly, often within one or two eye movements [23,24,29–31]. Indeed, viewers are so good at this that it is often not possible to study learning or repetition effects for real object search because search performance is already at ceiling [32]. Furthermore, viewers can rapidly find objects in scenes from a brief scene glimpse based on extraction of the 'gist' of the scene and can quickly find objects based on that gist even when the scene is no longer visible [33–35]. Viewers can also quickly find an object they are searching for when that object is not at all visually salient as long as the location of the searched-for object is constrained by the scene's meaning and structure [5,30]. For example, a coffee cup half-hidden by a box of Wheaties will still be fixated very quickly if it is in its expected location on the kitchen table [4]. How can we account for the strong influence of scene context on object search? The proposal here is that viewers use learned knowledge about where a given object is likely to be found from past experience with a given scene category to predict (given a new instance of that scene category) the location of the target, and this prediction is used to direct gaze.

Scene-Based Contextual Cueing

Because object search in scenes is so efficient, it can be difficult to study the learning processes associated with establishing the contextual relationships that underlie spatial prediction [32]. To