Eye Movements in Natural Behavior

Authors: Mary Hayhoe & Dana Ballard
Departments of Brain & Cognitive Science and Computer Science
University of Rochester, Rochester, NY 14627, USA
mary@cvs.rochester.edu
dana@cs.rochester.edu
Abstract

Although it has been recognized since Yarbus’ classic experiments over 50 years ago, that saccadic eye movements reflect cognitive processes, it is only recently that a number of separate lines of inquiry have merged to give an increasingly coherent understanding of the elaborate and intricate role of eye movements in cognitive function. There are three principal advances. The first is the demonstration of the role of eye movements in executing everyday visually guided behaviors. The major findings of these studies have been the ubiquitous role of the task in such movements, and the importance of learning where and when to fixate. The second advance has been the recognition of the role of internal reward in guiding eye and body movements. This has been revealed especially in neurophysiological studies. The third important advance has been the theoretical developments in the fields of reinforcement learning and game theory, together with developments in graphic simulation. Together, these developments have allowed the simulation realistic models of eye movements over extended time scales. All of these advances are critical for understanding how behavioral programs control fixations and the selection of visual information.
Over fifty years ago a Russian scientist Yarbus was able to capture the movements of the eye by attaching a mirror system to the eyeball. Although he was not first to measure eye movements, his work most clearly called attention to their intrinsically cognitive nature. Nowadays most vision scientists are familiar with his traces of a subject examining Ripken’s “The Unexpected Visitor” and the very different saccadic patterns elicited by the different questions asked of the subject [1]. The significance of this finding was that it revealed in a particularly compelling way that “seeing” is not a unitary process, and is inextricably linked to the observer’s cognitive goals. For example, the instruction to remember the position of the objects and people in the room might easily be taken as the job of vision. The fact that other instructions produced strikingly different patterns means that vision is much more complex. Although the role of eye movements in cognitive tasks has been studied extensively since that time [2-4], it is only recently that a number of separate lines of inquiry have coalesced to give an increasingly coherent understanding of the elaborate and intricate role of eye movements in cognitive function. The intent of this review is to describe these advances.

There are three principal advances that have had a direct bearing on the issue. The first is the demonstration of the role of eye movements in executing everyday visually guided behaviors [5-9]. These measurements have been driven by the development of portable eye trackers that can be worn by subjects engaged in behaviors that involve substantial body movements [10-12]. Figure 1 shows one such device. The major findings of these studies have been the ubiquitous role of the task in such movements, and the importance of learning where and when to fixate. The second advance has been the
recognition of the role of internal reward in guiding eye and body movements [13-17]. This has been revealed especially in neurophysiological studies. The third important advance has been the theoretical developments in the field of reinforcement learning, together with tremendous developments in graphic simulation [18-20]. Together, these developments have allowed the simulation of reward-based systems that incorporate realistic models of eye movements over extended time scales. The sum total of all of these advances is to shift the focus of experimental understanding of the role of eye movements. While the early focus has been on where the eyes fixate in an image, we now are more interested in why the eyes choose a location in a scene, and when they choose it.

**Eye Tracking in Natural Behavior**

Yarbus’ subjects had to suffer a mirror attached to a suction cup on the sclera, that had an unpleasant bulk. Subsequent systems, such as eye coils and the Dual Purkinje Image tracker were more comfortable, but still required the head to be held fixed. However the last ten years has seen the advent and rapid refinement of portable eye trackers that allow eye tracking during free viewing. Michael Land built one of the first of these but it was hampered by having to have the eye position calculated off line [5-7]. Modern eye trackers have fast software to do this and a recent version by Pelz, shown in Figure 1, is completely portable, using power from a backpack-mounted battery pack [12]. The new eye trackers allow the study of eye movements over extended tasks in natural settings, where a much wider variety of natural coordinated behaviors is possible.
**The Importance of Task.** Although Yarbus’ study revealed the importance of the instructions in determining where subjects look, the particular fixations do not reveal much more than that the observer attended to these locations. While a given cognitive event might reliably lead to a particular fixation, the fixation itself does not uniquely specify the cognitive event (Box 2). This problem is particularly acute in situations where subjects passively view pictures where the experimenter often has little control of, and no access to, what the observer is doing. There are indeed some regularities in fixation patterns that can be explained by image properties [21-23]. However, specific instructions to search or to remember objects lead to more specialized patterns [4,24-26,62]. Furthermore, if the task venue is set up so as to require sufficient visual acuity, then eye movements are almost invariably locked to attention [27, 28]. Thus, recent experiments where the task structure is evident have been more easily interpreted, because the task provides an external referent for the internal computations [2, 29,30]. The most novel finding of such studies is that the task dependent nature of vision is much more important than the native visual saliency of the scene (Box 1). In a variety natural tasks, the eyes are positioned at a point that is not the most visually salient, but is the best for the spatio-temporal demands of the job that needs to be done. This line of investigation has been used in extended visuo-motor tasks such as driving, walking, sports, and making tea or sandwiches [5-7, 9,31,32]. The central result of all these investigation is that fixations are tightly linked in time to the evolution of the task, and very few irrelevant areas are fixated. (See Box 1).

BOX 1
Not only is the sequence of fixations tightly linked to the task, but in addition, many of the fixations appear to have the purpose of obtaining quite specific information. For example, in cricket, players exhibit very precise fixation patterns, fixating the bounce point of the ball just ahead of its impact, since the location and time of the bounce provides batters with the information they needed to estimate the desired contact point with the bat [32]. Box 2 provides further evidence that highly task-specific information is extracted in different fixations. This specificity is indicated not only by the ongoing actions and the point in the task, but also by the durations of the fixations, which vary over a wide range [8,9]. It appears that a large component of this variation depends on the particular information required for that point in the task, fixation being terminated when the particular information is acquired [4, 11,37]. This underscores the overriding control of visual operations by the internal agenda rather than the properties of the stimulus, and the range of different kinds of visual information that can be extracted from the same visual stimulus.

Implicit in much of the research on natural tasks is the finding that eye movement patterns must be learned [6,10,38]. For example, in tea making and sandwich making, observers must have learnt what objects in the scene are relevant, since almost no fixations fall on irrelevant objects. In driving, Shinoda et al [39] showed that approximately 45 per cent of fixations fell in the neighborhood of intersections. As a consequence of this, subjects were more likely to notice Stop signs located at intersections as opposed to signs in the middle of a block. Thus it is likely that subjects have learnt that traffic signs are more likely around intersections. At a more detailed level, subjects must
learn the optimal location for the information they need. For example, when pouring tea, fixation is located at the tip of the teapot spout [7]. Presumably, flow from the spout is best controlled by fixating this location. Another important way in which eye movements reveal the influence of learning that they are often pro-active; that is, saccades are often made to a location in a scene in advance of an expected event. For example, in Land & MacLeod’s investigation of cricket, batsmen anticipated the bounce point of the ball, and more skilled batsmen arrived at the bounce point about 100 msec earlier than less skilled players [32]. These saccades were always preceded by a fixation on the ball as it left the bowler’s hand, showing that batsmen use current sensory data in combination with learnt models of the ball’s motion to predict the location of the bounce. Thus, eye movement patterns appear to be shaped by the dynamic properties of the world.

BOX 2

**Context-Dependent Neural Activity**

The eye tracking studies show that the locus and sequence of gaze positions, and the associated visual computations, are orchestrated by the ongoing cognitive goals. The new challenge is to understand how this can be achieved by the neural machinery. There are two ways in which recent neurophysiological evidence helps us understand the cognitive control of eye movements. The first is the growing evidence that neurons behave in a task specific manner, even in early cortical visual areas such as V1. The higher levels of the cortex, such as the prefrontal cortex, are of course involved in setting up this process, but
the elements of many operations are handled early in visual cortex. For example Gilbert and colleagues have shown that the perceptual judgment required of the animal determines the way in which stimulus context modulates the classical receptive field of V1 neurons [41]. Another compelling experiment that reveals this task-dependent processing in primary visual cortex is that by the Roelfsema group, who showed that monkeys solve a visual line tracing task by activating specific cells in striate cortex [42]. Interestingly, the time at which the cells responded was delayed by 100 msec when a color judgment was required in addition to line tracing, suggesting a cost to an added computation. Thus primary visual cortex can dynamically modify the way it processes visual information depending to the immediate behavioral requirements. These experiments demonstrate that even primary visual cortex can be dynamically reconfigured to selectively extract the specific information required for the momentary task, in a way suggested by the psychophysical data.

Higher cortical areas also respond in a task specific manner, shaped by experience. One of the most telling examples of this task specificity is revealed in the activity of neurons in dorso-lateral prefrontal cortex. These cells appear to code the conjunction of specific visual qualities with the learned motor response. For example, some neurons show selective responses for an image of a particular object, but only when that image instructs an eye movement in a particular direction [43,44]. Thus, the demands of the task seem to be an intrinsic component of the brain’s representational structure of visual information.
Eye Movements are driven by prospects of reward.

The second development in neurophysiology that helps us understand the cognitive control of eye movements is that of showing the influence of reward on eye movements. We know the brain must learn without an explicit programmer, but how is this done? Much research supports a reward-based learning mechanism. The brain has some way of generating alternatives and these can be scored with a secondary reward signal believed to be dopamine. The most impressive evidence for this comes from experiments by Schultz that show that dopaminergic neurons in the substantia nigra pars compacta – a part of the basal ganglia system - behave in ways predicted by mathematical models of reinforcement [17,45]. This reward system is integral to the generation of saccadic eye movements. Cortical saccade-related areas (frontal eye fields, dorso-lateral pre-frontal, and lateral intra-parietal) all converge on the caudate nucleus in the basal ganglia, and the cortical- basal ganglia- superior colliculus circuit appears to regulate the control of fixation and the timing of planned movements. This is achieved by regulation of tonic inhibition exerted by the substantia nigra pars reticulata on the superior colliculus, the mid-brain site of saccade generation. Such regulation is a critical component of task control of fixations. Hikosaka and colleagues have demonstrated that caudate cell responses reflect both the target of an upcoming saccade and the reward expected after making the movement [15,46]. Thus the neural substrate for learning where to look in the context of a task is present in the basal ganglia.

Sensitivity to reward is manifest throughout the saccadic eye movement circuitry. In the lateral intra-parietal area (LIP), the neurons involved in saccadic targeting respond
in a graded manner to both the amount of expected reward, and the probability of a reward, in the period prior to execution of the response [13,14,47,48]. Sensitivity to both these variables is critical for linking fixation patterns to task demands and is predicted by game theory and decision theory. Cells in the Supplementary Eye Fields, also signal the animal’s expectation of reward and monitor the outcome of saccades [16]. Sensitivity to stimulus probability is also revealed in build-up neurons in the intermediate layers of the Superior Colliculus, the mid-brain saccade generator. Basso & Wurtz [49] showed that firing in these neurons prior to the saccade is proportional to probability that the upcoming target will appear in that location. This forms a nice parallel with psychophysical observations showing that saccade reaction time is similarly influenced by stimulus probability [50]. Hand movements are also sensitive to reward structure. Trommershauser and colleagues [51] have shown that subjects making fast hand movements are able to learn a complicated spatially distributed target reward system and behave in a nearly optimal manner to maximize reward. All of these results are in stark contrast to the more traditional approach to understanding saccades that treats the saccade circuitry as a linear system, with fixed input-output properties. While this approach works well for describing the dynamic properties of the movements, the target selection mechanisms appear to be intimately related to behavioral outcomes.

**Simulation of Virtual Humans with Gaze Control**

In this brief review we have tried to sketch the enormous progress in understanding eye movements by understanding their surrounding task context together with the
additional rapid progress made in understanding the neural control of such movements as fundamentally reward-driven. These developments present the challenge of understanding such systems, which can be very complex. Fortunately there has been a third avenue of development and that has been in virtual reality (VR) graphics environments. Such environments now run in real time on standard computing platforms. The value of VR is that it allows the creation of human models that implement complex visuo-motor loops for the control of temporally-extended behaviors. Visual input can be captured from the rendered virtual scene, analyzed and the results used to generate motor commands that direct the graphical representation of the virtual agent's body [19]. Virtual humans have the advantages of experimental reproducibility, hardware requirements, flexibility, and ease of programming. What these simulations reveal is the different kinds of computational problems that are involved in understanding the brain’s allocation of resources. During the course of normal behavior, humans engage in a wide variety of tasks, each of which requires certain perceptual and motor resources, one of which is gaze. Thus there must be mechanisms that allocate resources to tasks. These mechanisms have been extensively studied under the headings of “attention” and “working memory.” Virtual human simulations provide a complementary approach wherein a complete system model is produced that includes precise specifications of the ongoing allocation of resources to tasks as revealed through the control of gaze.

One of the principal findings of these simulations is that while eye fixations subserve behaviors, the information they provide needs to be multifaceted, in ways that refine our notions of attention and working memory. Gaze is still directed to a point that is
the best choice at some instant and then a task-directed computation is done at that point, but the issues surrounding the direction of gaze are fundamentally different than those surrounding the particular focal computation at the fixation point. As a consequence it is very helpful to think of the visual computations that need to be done as hierarchically organized, a viewpoint that has not been as obvious from the attention/working-memory dichotomy. Table 1 shows the basic elements of a representative hierarchy highlighting the different roles of vision at each level.

Table 1. Different uses of Gaze at Different Levels of Abstraction

<table>
<thead>
<tr>
<th>Abstraction Level</th>
<th>Problem Being Addressed</th>
<th>Role of Vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Behavior</td>
<td>A visual routine needs to get state information</td>
<td>Provide State Estimation</td>
</tr>
<tr>
<td>2. Arbitration</td>
<td>Different behaviors may have competing demands for the eyes.</td>
<td>Move gaze to the best location</td>
</tr>
<tr>
<td>3. Context</td>
<td>Current set of behaviors is inadequate</td>
<td>Test for off-agenda exigencies and provide a updated set of behaviors</td>
</tr>
</tbody>
</table>

The issues that arise for vision are very different at each the levels in this hierarchy.

1. The issue at the level of individual behaviors is that of understanding how gaze can be used to compute state information necessary for meeting behavioral goals. Almost invariably, the visual computation needed in a task context is vastly simpler than that required by general-purpose vision and, as a consequence, can be done very quickly [52-55]. Psychophysical evidence for such context-specific visual computations within a fixation was described above (Box 2), and context-specific neural activity in visual cortex [41,42] provides a neural basis for such state computations.
2. At the arbitration level, the principal issue for eye movement control is that the center of gaze is not easily shared and instead generally must be allocated sequentially to different locations. Since the set of active behaviors must share perceptual and physical resources, there must be some mechanism to arbitrate when they make conflicting demands. Task arbitration, particularly dual-task performance, [56,57] has been extensively studied but not from the perspective of gaze allocation. One arbitration model [20] shows how gaze allocations may be selected to minimize the risk of losing the reward associated with a given behavior. As shown in the psychophysical studies of the effects of task context, if the experiment can be structured to make the eye movements overt, then they can be very revealing indicators of the results of this arbitration process.

3. The context level maintains an appropriate set of active behaviors from a much larger library of possible behaviors, given the current goals and environmental conditions. At this level, the focus is to maintain an appropriate set of behaviors to deal with internally generated goals. One of these goals is that the set of running behaviors be responsive to rapid environmental changes. Thus the issue for vision at this level is that of understanding the interplay between agenda-driven and environmentally driven visual processing demands. This issue has not received much attention in the cognitive literature but has been extensively studied in robotic models [58-60].
The various hierarchical levels are illustrated in the “Walter” humanoid simulation conducted by Sprague (See Box 3). In this simulation the direction of gaze is entirely controlled by rewards that are learned en route. As shown in the Box, the issues surrounding eye movements are very different depending on the abstraction venue. The main message is that the hierarchy, rather than adding complexity, simplifies the analysis because the computations at each level are much more independent of each other. Thus different models of any one aspect of the design can be tested in an environment that captures much of the complexity of real human eye movements.

Summary

There is a long history of attempts to use eye movements to infer cognitive processes. Saccades are quintessentially voluntary movements [2], and a variety of psychophysical and imaging studies support the idea that the shifts in attention made by the observer are usually reflected in the fixations [61]. The most recent research extends these ideas in several ways. Portable eye tracking technology allows us to study the effects of task requirements on the underlying cognitive process over much more extended periods. Developments in virtual environments allow us to pinpoint the information used moment by moment in a task and revealed it to be exquisitely specific. Neural recordings in monkeys engaged in specific tasks support such information specificity. The mental programs that drive gaze selection are increasingly seen to be reward-based. This direction is spurred
by extensive neural recordings that show vast areas of the brain’s gaze computation system exhibit sensitivity to reward signals, as well as developments in reinforcement learning theory that make very specific predictions, some of which have been confirmed. Lastly, reward-based gaze control has been given impetus from graphic simulations. These allow the modeling of situated gaze with vastly increased fidelity and confirm the viability of the understanding of why we look at places in the world in terms of potential rewards

Acknowledgements.

This work was supported by NIH grants EY05729 and RR09283.

References.


Figure 1. Portable eye-tracker. Portable, head mounted eye-tracker developed by Pelz and colleagues [15], based on commercially available systems that use an infra-red video camera to image pupil and the corneal reflection. A scene camera mounted on the frame of the glasses records the scene from the observer’s viewpoint. Eye position is then superimposed on the video record.
Box 1. Fixations Cluster At Locations Critical For Task Performance

When making tea or sandwiches, very few fixations fall on objects that are irrelevant to the task. The figure below shows all the fixations made in the course of making a peanut butter and jelly sandwich. The image shows the scene at the beginning of the task. Fixations on and around the plate were made while spreading the bread, putting the jars and lids near the plate etc. Note that almost all the fixations fall on task-related objects, despite the presence of other salient objects in the scene. When viewing a video replay of the subject’s view of the scene, with eye position superimposed, it is clear that fixations are tightly linked in time to the actions, such a grasping and moving objects, moving on to the next object when the needs of the current action have been met [7,9,33,34]. Similar step-by-step control of hand actions by fixation at a specific locus in the scene has been demonstrated by Johansson et al [35] in a task where subjects picked up a bar, moved it past an obstacle, and used it to contact a switch (see Figure 2). Fixations clustered at critical loci for each segment of the movement, moving on to the next locus as the action was completed. Thus the basic structure of the task allows one to link the visual operations fairly closely in time with the occurrence of eye and hand movements. Ballard et al [30] called this a “just-in-time” strategy, where observers acquire the specific information they need just at the point it is required in the task.

Figure 1. Fixation patterns. Fixations made by an observer while making a peanut butter and jelly sandwich, indicated by yellow circles. Images were taken from a camera mounted on the head, and integrated over different head positions by a method described in Rothkopf et al [36]. Note that almost all fixations fall on task relevant objects, and that the two observers fixate similar locations. Circle diameters reflect the duration of the fixation.
Figure 2. Link between fixation and the hand. Fixation locations and hand path while a subject picks up and moves a bar past an obstacle, and then contacts a switch. The solid line indicates the fixations, and the dashed line shows the fingertip position. The subject holds the bar by the end on the right, and must maneuver the tip past the triangular shaped obstacle. Numbers on the fingertip path indicate fingertip position during the fixation with the corresponding number. Fixations are made at critical points such as the point of contact for the fingers, then the end of the bar after pickup, followed by fixation near the tip of the obstacle while the bar is moved around it, and then near the switch once the bar has cleared the obstacle. Adapted from Johansson et al [35].
Box 2. What do we see when we look?

Fixation patterns in natural tasks suggests that very specific visual computations are performed at the fixation point as needed for task performance [5,8,9]. For example, when a subject first fixates on the jelly jar in the scene in Box 1, it is presumably for the purpose of recognition. Later, when they fixate the jar and grasp it, they are presumably computing geometric information to control the hand and arm. An experiment by Triesch et al [40] supports this supposition. In their experiment, subjects sorted virtual bricks of two different heights onto two “conveyor belts” (horizontal strips on the right hand side of the virtual work-space) according to different rules that vary the points at which the brick height is relevant in the task. In one condition, subjects simply picked up the bricks in front-to-back order and placed them on a belt. In a second condition, subjects picked up the tall bricks first and put them on a belt, and then, picked up the small bricks and put them on the same belt. In the third condition, the tall bricks were put on the front belt, and the short bricks on the back belt. In the first case, size is irrelevant. In the second, size is relevant for pickup only. In the third, it is relevant for both pickup and placement. On some trials, the brick changed size while it was being moved to the belt. Subjects rarely noticed the change when size was irrelevant, suggesting that subjects did not visually represent the brick size in these trials. Interestingly, when size was relevant only for pickup, subjects were less likely to notice changes than when it was relevant for placement as well, suggesting that subjects did not retain size in working memory when it was no longer needed. On some trials, subjects tracked the brick while they moved it across the workspace, so that they were fixating the brick at the point when it changed. In many of these cases, subjects were still unaware of the change. This suggests that subjects may not
represent particular stimulus features such as size when they are performing other computations such as guiding the arm.

**Figure 1: Moving Bricks Task.** View of the virtual work-space as a subject (a) picks up, (b) carries, and (c) places a brick on a conveyor belt. The pink dots show the fingers. The cross-hair shows fixation.
Box 3: Different Gaze Functions Illustrated in a Virtual Reality Simulation

1. Visual Routines for Updating State Information

2. Use of Gaze for Competing Tasks

3. Choosing an Appropriate set of Behaviors for Different Contexts

Three different hierarchical functions of gaze, shown in Table 1 are shown in a graphic simulation that uses a human-like figure ‘Walter.’ The model is based on the use of reward-guided visual behaviors that use vision exclusively for perception. (Top) Walter is shown using a visual routine for determining his location on the sidewalk. The top left hand inset shows the result of a boundary detector looking for sidewalk edges. (Middle) A delayed exposure shows Walter allocating gaze to the service of three competing ongoing
behaviors that need gaze to update internal models. Gaze is given to the behavior with the most to lose as measured by reward expectation. Initially, the proximity of an obstacle attracts three successive fixations (red gaze vectors), subsequently the sidewalk following behavior is given a fixation (blue) followed by three successive litter pickup fixations (green). (Bottom) Depending on the context, Walter activates different sets of behaviors from a larger repertoire. The fixations from the different behaviors are indicated by different colors. For example while in the ‘On Crosswalk’ context, the behaviors are twofold: 1) following the crosswalk (gray fixations) and 2) checking for the other side (pink fixations). Adapted from (20).