

Visual memory and motor planning in a natural task

Mary M. Hayhoe

Center for Visual Science, University of Rochester,
Rochester, NY, USA



Anurag Shrivastava

Center for Visual Science, University of Rochester,
Rochester, NY, USA



Ryan Mruczek

Center for Neural Science, Rochester Institute of Technology,
Rochester, NY, USA



Jeff B. Pelz

Center for Neural Science, Rochester Institute of Technology,
Rochester, NY, USA



This paper investigates the temporal dependencies of natural vision by measuring eye and hand movements while subjects made a sandwich. The phenomenon of change blindness suggests these temporal dependencies might be limited. Our observations are largely consistent with this, suggesting that much natural vision can be accomplished with “just-in-time” representations. However, we also observe several aspects of performance that point to the need for some representation of the spatial structure of the scene that is built up over different fixations. Patterns of eye-hand coordination and fixation sequences suggest the need for planning and coordinating movements over a period of a few seconds. This planning must be in a coordinate frame that is independent of eye position, and thus requires a representation of the spatial structure in a scene that is built up over different fixations.

Keywords: natural tasks, saccades, eye-hand coordination

Introduction

One of the fundamental issues in visual perception is how visual mechanisms operate over time scales longer than a single fixation. Visual operations are normally embedded in the context of extended behavioral sequences. However, we have limited understanding of how visual processes operate in the service of natural, ongoing, behavior. A central aspect of vision in its natural context is how we make the transition from the computations within a fixation to those that operate between fixations. To what extent does the current computation depend on information acquired in previous fixations, or are visual operations within a fixation essentially independent? This question has traditionally been addressed in the context of integration of information across saccadic eye movements: whether there is such an integrated representation of a visual scene, and what the contents of that representation might be (Irwin, 1991; Rayner & Pollatsek, 1983). The conclusion from a large body of previous work is that representation of information acquired in prior fixations is very limited. Evidence for limited memory from prior fixations is provided by the finding that observers are extremely insensitive to changes in the visual scene during an eye movement, film cut, or similar masking stimulus (Hendersen, 1992; Hochberg, 1986; Irwin, 1991; Irwin, Zacks, & Brown, 1990; Pollatsek & Rayner, 1992; O'Regan, 1992; Rensink, O'Regan, & Clark, 1997;

Simons, 1996). Many of these, and more recent studies, have been reviewed by Simons and Levin (1997) and Simons (2000), and this insensitivity to changes has been described as “change blindness.” Since detection of a change requires a comparison of the information in different fixations, change blindness has been interpreted as evidence that only a small part of the information in the scene is retained across fixations. Irwin suggests that it is limited by the capacity of working memory, that is, to a small number of individual items whose identity is remembered better than their location (Irwin, 1996). Thus memory from prior fixations is primarily semantic in nature, suggesting a large degree of independence of the visual computations within individual fixations.

It is not clear, however, how much we can generalize these findings to natural vision. What information in a scene do observers actually need, and how much of this information persists past a given fixation? Although some studies have examined change blindness in the real world (Simons & Levin, 1997), most paradigms examine a single visual or motor operation over repeated trials. Visual function in this context may be fundamentally different from active participation in a real scene. Observers almost certainly fine-tune their behavior to the experimental demands. For example observers are very sensitive to the probabilistic structure of the trials, and match the distribution of attention to expected events (Mack & Rock, 1996). In natural behavior, the observer performs a sequence of different computations, whose

initiation and timing is controlled by the observer, not by the experimenter. This active initiation of behaviors is likely to be important. For example, viewing a picture of a scene is very different from acting within that scene, simply because the observer needs different information. Some evidence for the importance of observer actions is given by Wallis and Büthoff (2000), who showed that drivers and passengers in a virtual environment have different sensitivity to changes in the scene. Other evidence also suggests the importance of the immediate task in determining what is detected (Folk, Remington, & Johnston, 1992; Hayhoe, Bensinger, & Ballard, 1998).

Another difference that is likely to be important is the nature of the stimulus array. Investigations of change blindness typically involve viewing either two-dimensional pictorial representations of scenes or simple arrays of letters or geometric figures. These displays differ from normal scenes in their spatial structure. One difference is spatial scale. The visual angle subtended by an image of a room in a typical experimental display, for example, is very different from being in a real room, and it is not clear how such infidelities in spatial scale might affect observers' representations of the spatial structure of the scene. Depth information introduces an additional level of spatial complexity in normal vision and poses a greater challenge for the visuo-motor apparatus.

Both active participation and spatial structure are likely to be important for understanding the visual representations that are used to locate targets for the eyes and hands, and to coordinate the movements of eyes, head, hands, and body. These behaviors are important requirements of any situation, and Chun and Nakayama (2000) pointed out the potential importance of implicit memory structures for guiding attention and eye movements around a scene. They argue that such guidance requires continuity of visual representations across different fixation positions. In contrast to the findings of the change blindness experiments, there is evidence to suggest that subjects do in fact build an implicit memory representation of the spatial structure of the environment. Chun and Jiang (1998) showed that visual search is facilitated (by 60-80 msec) by prior exposure to visual contexts associated with the target. They suggest that this reflects sensitivity to the redundant structure in a scene, which remains invariant across multiple gaze points. It seems likely that observers are sensitive to this invariance. Other evidence for an influence of prior views is "priming of popout." This is the reduction of both search latencies and saccade latencies to locations or features that have been recently presented (Maljkovic & Nakayama, 1994; McPeck, Skavenski, & Nakayama, 2000). Such mechanisms do not require conscious intervention, and exhibit greater memory capacity, longer durability, and greater discriminability than explicit short-term visual memory. Chun & Nakayama proposed that both contextual cueing

and priming of pop-out might be mechanisms that guide attention and eye movements in scenes.

The goal of the present investigation was to examine the fixation patterns in natural behavior, in order to gain insights about the way that natural behavior might depend on information in prior fixations. We recorded eye and hand movements while making a sandwich. This task was modeled on a similar one used by Land, Mennie, and Rusted (1999), who recorded fixation patterns while observers made a cup of tea. Like tea-making, the sandwich making task allows the observer considerable flexibility while still providing an explicit set of behavioral goals for the observer. Some explicit manifestation of the observer's goals is required for understanding the behavior. The focus of the current investigation was to understand the temporal dependencies of natural behavior. Thus, to what extent is visual information that was acquired in prior fixations needed for performing the task. In particular, to what extent is such information needed for guiding eye and hand movements? Our observations confirm earlier studies demonstrating the transient, task-specific nature of the information extracted within a fixation (Ballard, Hayhoe, & Pelz, 1995; Hayhoe, Bensinger, & Ballard, 1998; Land et al, 1999). Thus much visual processing is accomplished within a fixation. Because of this, change blindness may not be much of a limitation in normal performance, because much of the information in a scene is not in current use. However, we also observe several aspects of performance that point to the need for some representation of the spatial structure of the scene that is built up over different fixations. Patterns of eye-hand coordination and fixation sequences suggest the need for planning and coordinating movements over a period of a few seconds. This planning must be in a coordinate frame that is independent of eye position and thus requires a representation of the spatial structure in a scene that is built up over different fixations.

Methods

Subjects wore an eye-tracker mounted on the head, and were seated at a table with the items required for making a sandwich. They were thus free to make natural movements. No instructions were given except to make a peanut butter and jelly sandwich and to pour a glass of soda. Observations were made on 11 subjects. Seven of the subjects made the sandwich with the layout demonstrated in Figure 1 (top).

The necessary items were laid out on the table in front of the observer, with a few background items irrelevant to the task. Four subjects made the sandwich with a more cluttered layout, as shown in Figure 1 (bottom), where a number of arbitrarily chosen irrelevant items (other food items, tools, silverware) were



Figure 1. The two scene layouts used in the experiment.

interspersed with the items required for the task. Before the experiment, the layout was occluded by a cardboard sheet showing the calibration points. Following calibration, this was withdrawn, and the subjects immediately began the task. The research followed the World Medical Association Declaration of Helsinki and was approved by the University of Rochester Research Subjects Review Board. Informed consent was obtained from the subjects.

Monitoring Eye Position

Monocular (left) eye position was monitored with either an Applied Science Laboratories Model 501 or an ISCAN eyetracker. The ISCAN was used with the uncluttered scene, and the ASL was used in the cluttered scene. Both are headband mounted, video-based, IR reflection eyetrackers. The eye position signal was sampled at 60 Hz and had a real time delay of 50 msec. The accuracy of the eye-in-head signal is approximately 1°

over a central 40° field. Both pupil and first Purkinje image centroids are recorded, and horizontal and vertical eye-in-head position is calculated based on the vector difference between the two centroids. This technique reduces artifacts due to any movement of the headband with respect to the head. (Errors in reported eye position caused by movement of the headband with respect to the head were less than 0.1° , measured over a sequence of movements at a peak velocity of $60^\circ/\text{sec.}$) Both trackers provide a video record of eye position. The ISCAN headband held a miniature “scene-camera” to the left of the subject’s head, aimed at the scene. The ASL’s scene camera was mounted so as to be coincident with the observer’s line of sight. The tracker creates a cursor, indicating eye-in-head position, that is merged with the video from the scene-camera, providing a video record of the scene from the subject’s perspective on the scene-monitor, with the cursor indicating the intersection of the subject’s gaze with the working plane. Because the scene-camera moves with the head, the eye-in-head signal indicates the gaze point with respect to the world. Head movements appear on the record as full field image motion. (Because the ISCAN scene camera was not coaxial with the line of sight, calibration of the video signal was strictly correct for only a single plane. Calibration was close to the plane of the table, so the parallax error was significant when subjects lifted objects out of that plane toward the body.)

The eye tracker was calibrated for each subject before each trial. The subject was seated at the work surface, with all items within reach. At this distance, the plate close to the observer subtended about 20° of visual angle, and the peanut butter and jelly subtended about 7° . All the items were within about a 90° region. Calibration was performed using a nine-point grid, over a region of about 50° by 40° . (This region moves with the subject’s head.) Following data collection, which took about two minutes per subject, the video records were analyzed on a frame-by-frame basis, recording the time of initiation and termination of each eye and hand movement, the location of the fixations, the nature of the hand actions, and periods of track loss. These detailed records formed the basis of the summary statistics described below. For 4 of the subjects, the image of the eye provided by the tracker was superimposed on the record from the scene camera either as a transparent overlay, or in the top corner of the scene image. The eye image is shown in Figure 2. Two crosshairs indicate the tracker’s calculation of center of the pupil and corneal reflection. If either of these signals is lost, the corresponding crosshair disappears. This provides a mechanism for checking the scene video for transient track losses and blinks. The movement of the eye can also be seen in the eye image, providing an additional source of information for identifying fixations and measuring their duration.

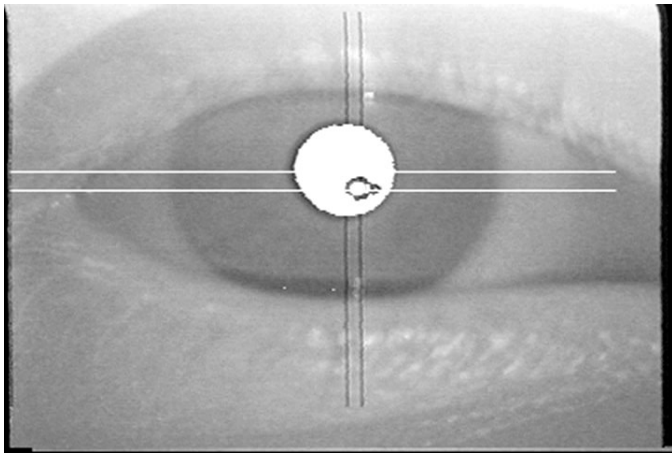


Figure 2. The eye image with cross-hairs indicating pupil center and corneal reflection.



Movie 1. A segment of the experimental task.

Results

Task Specific Fixations

In agreement with Land et al's (1999) observations on tea-making, we found that fixation patterns were highly directed, and used to acquire specific information just as it was needed for the momentary task. A description of a small segment of the task is shown in Figure 3 and Movie 1. At the beginning of the segment, the subject is fixating the completed sandwich on the plate, guiding the knife to cut the sandwich with the right hand, and the left hand steadies the bread. Gaze is then transferred to the edge of the plate to guide placement of the knife with the right hand. The left hand simultaneously begins to move toward the lid of the jelly jar on the table.

While the right hand completes placement of the knife, the eye fixates the jelly jar briefly, then fixates the lid to guide pickup with the left hand. The eye then returns to the jelly jar to guide the lid towards the jar. Just before the left hand, holding the lid, makes contact with

the jar, the right hand also moves toward the jar to coordinate with the left hand in screwing it on the jar, and so on. Thus the first fixation is for guiding knife putdown. This requires both directional and distance information for controlling the arm and computing the contact forces. The fixation on the lid on the table is required to guide pickup. This involves computing information to control the grasp, including the position, orientation, and size of the lid, and perhaps recalling from memory information about surface friction and weight to plan the forces. The intervening fixation on the jar may provide information for the future movement to place the lid on the jar. The final fixation on the rim of the jar initially guides the direction and posture of the left hand to contact the jar, then the right hand movement and posture to the jar, and then the lid placement and the screwing action. Thus fixations are tightly locked to the task, and their role is well-defined. Fixations on task-relevant objects were typically close in time to their use in the task. For example, in this subject, except for fixations immediately on viewing the scene, the soda was not fixated until the subject was about to pour a glass.

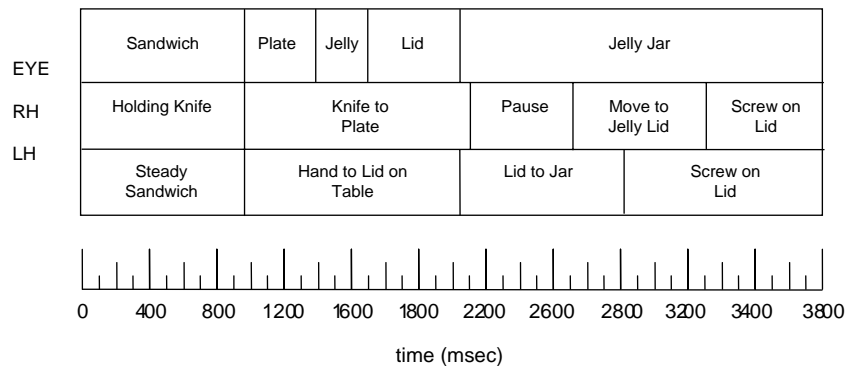


Figure 3. Sequence of eye and hand movements shown in Movie 1.

Fixations appear to play a specific role, depending on momentary task context. The locations of the fixations on the objects were different for different actions, for example, subjects fixate the middle of the jar for grasping with the hand in a vertical posture, and the rim for putting on the lid, with the hand in a horizontal posture. This suggests that the visual information being extracted controls the pre-shaping in one case, and the orientation of the lid in the other. To the extent that information is obtained at the moment it is needed, visual computations depend only on the information available within that fixation.

In the first scene, there were a number of objects surrounding the work area, as shown in Figure 1 top, a monitor, camera, tools etc. Subjects rarely fixated these background items. This occurred on only 0.02 of the fixations (one or two per subject). In the second, cluttered, scene, shown in Figure 1 bottom, a variety of irrelevant objects were present, interspersed with the items needed for the task. There was an approximately equal number of irrelevant as relevant items. In this case, an average of 0.2 ± 0.04 of the fixations were made on irrelevant items.

Fixation Durations

The duration of each fixation was calculated from the video transcriptions. The frequency distributions are shown in Figures 4 and 5. Data for the three subjects in Figure 4 were recorded with the image of the eye provided by the tracker, superimposed on the record from the scene camera. This allowed careful monitoring of the fixation durations measurements, since a transient track loss sometimes results in a deviation of the cursor position, and thus appears like the termination of a fixation. Movement of the eye during the track loss could be observed directly in the eye image. The distributions are quite similar for the different subjects. The data for the seven subjects in Figure 5 did not have the eye image available on the video record. It is therefore possible that these data are partially contaminated by transient track losses. This should not be a major factor, however, as segments of the tape where the cursor disappeared, indicating a track loss, were eliminated from the analysis. The most distinctive feature of these distributions is their wide spread. Fixations range from under 100 msec to over 1500 msec. There is some variation between subjects, but most of the distributions have a mode between 100 and 200 msec, which is less than for reading or picture viewing (Henderson & Hollingworth, 1999). The very long fixations are usually associated with some prolonged action of the hands that required continuous guidance, such as spreading, scooping out peanut butter, pouring, or undoing the tie on the bread bag. Land et al (1999) observed a similarly wide spread of fixation durations in their tea-making task. For the long fixations, it is important to note that the noise in the tracker made it

impossible to identify small saccades within a radius of about 1.5° around fixation. If these were present, the number of long fixations would be overestimated. Although it is impossible to know what the role of individual fixations is from such observations, it appears that, to a first approximation, the fixation durations are determined by the momentary task demands. Gaze often departs just at the point a hand movement is complete, or there is no longer need for visual guidance. An example of this is given in Figure 3, and in the accompanying video, where the eye departed from controlling knife placement when the knife was close to the plate, and the remainder of the movement could be controlled using somatosensory information. The eye then arrived to guide lid pickup just as the left hand approached the lid. Similar time-locking of fixations to critical stages of the actions was observed by Johansson et al (2001). This is, of course, an incomplete description of determinants of fixation duration. In a number of instances vision may not be providing critical information for the ongoing action. For example, screwing the cap on the soda bottle can be completed under proprioceptive control and it is not obvious what role is being played by fixation during such periods.

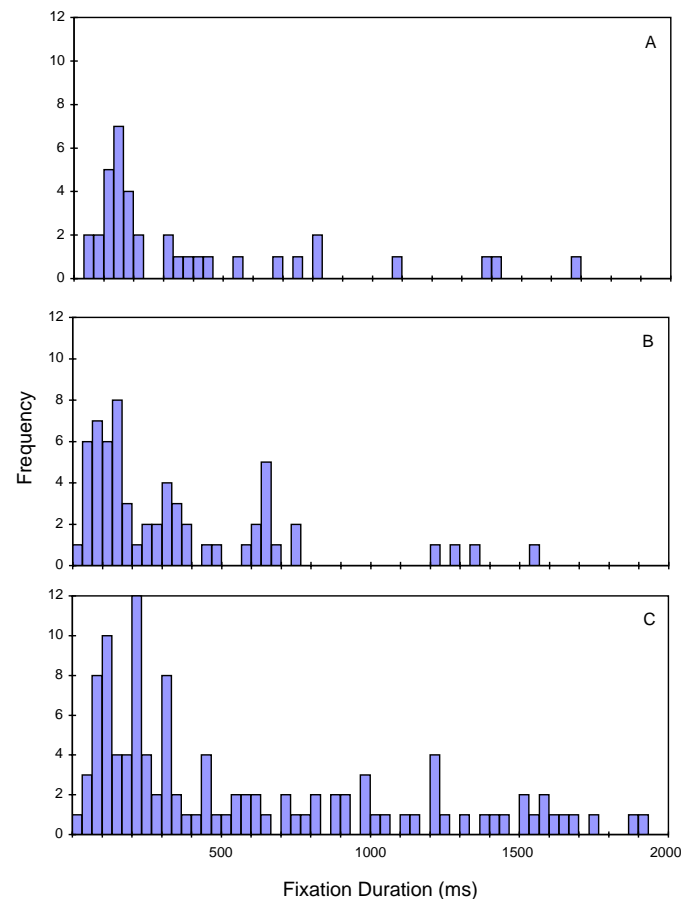


Figure 4. Distribution of fixation durations for 3 subjects using the eye image.

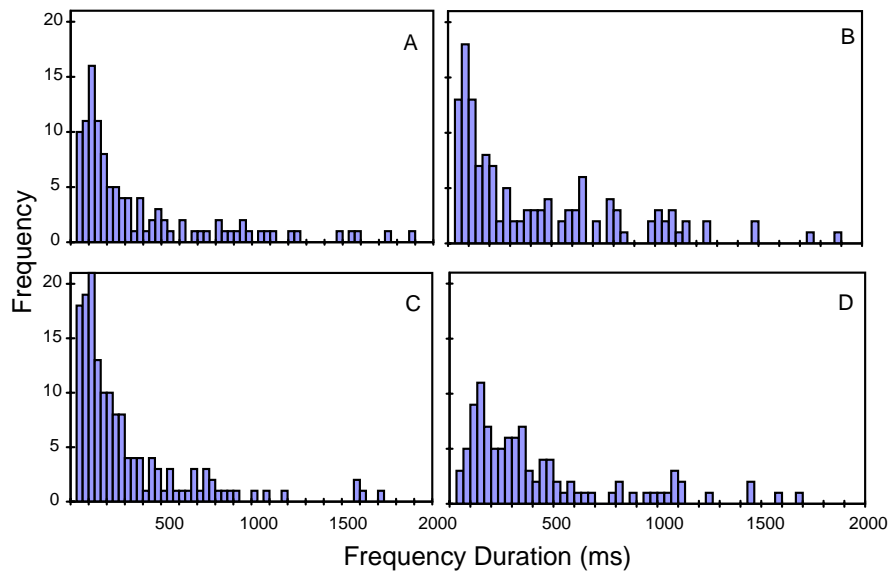


Figure 5. Distribution of fixation durations for another 4 subjects without the eye image.

An interesting feature of the distributions is the frequency of very short fixations. All subjects show a number of fixations of two to four video frames (66 – 133 msec). The measurement of the short fixations for these subjects is thus highly reliable, within the temporal resolution limits of the video record (30 Hz). The short fixations do not appear to play a single specialized role. We examined all the fixations of 100 msec or less and attempted to categorize them according to the context. The frequency of the various categories is shown in Table 1. Very short fixations have previously been observed between the primary and corrective saccade to a remembered target location (Becker & Fuchs, 1969). In our experiment only about 0.1 of the short fixations could be potentially classified as preceding a corrective saccade. We classified corrective saccades as those where the eye landed on an object (for example the bottle), and then moved to an adjacent location on the object (the neck), followed by an action involving the object (pouring). The occurrence of the action suggests that the second fixation locus is the intended target, though this is not known with any certainty. Of the other short fixations, 0.24 occurred while guiding some kind of reaching movement, either for picking up or placement; 0.07 were on a task-relevant object located more or less on the path of the saccade, between the pre-saccadic position and the location of the next item to be manipulated. Notably these objects were ones needed at some other point in the task. For example, a brief fixation might occur on the knife, positioned between plate and bread, as the subject moved from plate to bread to open the bread bag. About 0.12 of the fixations occurred following some kind of occlusion of the point of interest by a hand or following a blink. The remaining 0.37 of the fixations could not be obviously classified. Thus the

fixations occur on a variety of occasions and are not limited to the interval before a corrective saccade.

Table 1. Categories of Fixations Less than 100 msec.

Frequency	Type of fixation
0.24	guide reach
0.1	corrective
0.07	in path
0.12	occlusion
0.37	other

These short fixations are of interest because the time to program a saccade is reliably found to be in the 200-250 msec range. Consequently these brief fixations must be part of pre-programmed sequences of saccades (Becker & Jurgens, 1979). This planning must be done in a spatial, not retinal coordinate frame, and the partially programmed second saccade updated for the first movement. Thus pre-programmed sequences of saccades point to the existence of a representation in spatial coordinates, independent of eye position.

Initial Fixations

The strict capacity limits on the information that can be retained across fixation positions raise the question of whether there is some representation of scenes that is built up over time. O'Regan & Levy-Schoen (1983) and Irwin (Irwin, 1991; Irwin et al, 1990) suggest there is some sparse, post-categorical description of the objects and their locations in a scene accumulated over different eye positions. This seems plausible, since in most ordinary environments observers have prolonged exposure to the scene, and multiple opportunities to accumulate

information, despite the capacity limits of visual short-term memory. However, it is not known what subjects do in natural viewing. We were therefore interested to observe what subjects look at when they first view a novel scene. Do they in fact make a series of exploratory eye movements as we might expect if they are building a representation of scene layout for later use? We therefore examined the fixations made by subjects after the scene was initially exposed by removing the calibration display, and before the first reaching movement, which indicated that they had begun the task. We found that on the initial exposure, subjects scan the scene and make a series of fixations on the objects, before the first reaching movement is initiated. Each of the 11 subjects made between 3 and 21 fixations on this initial exposure. The mean number of fixations was 8.9 +/- 1.5, as shown in Table 2.



Movie 2. Initial fixations of one subject.

Table 2. Number and Duration of Pre-Task Fixations, and Frequency of Fixations on Irrelevant Objects.

Pre-task fixations	8.9 +/- 1.5
Fixation duration	197 msec +/- 26
Pre-task	
Irrelevant objects	0.17 +/- 0.04
During task	
Irrelevant objects	0.48 +/- 0.07
Pre-task	

An example of one subject's fixations on first view is given in Figure 6. This subject makes a series of short fixations on the bread, the peanut butter, in between the peanut butter and the jelly, two fixations on the bread, then on the jelly, between soda and jelly, and then to the bread bag to guide the first reaching movement. A second subject's initial fixations are shown in Movie 2. In the case of subjects who used the cluttered scene, these initial fixations were distributed fairly equally between relevant and irrelevant objects (0.48 +/- 0.07, on irrelevant objects). During task performance, however, the proportion of fixations on irrelevant objects went down to 0.16 +/- 0.04. This suggests that subjects are doing something different in the initial fixations. The initial fixations were typically quite short (mean 197 msec +/- 26). Thus the information being acquired in these fixations does not take extensive visual analysis.

Eye-Hand Latencies

Little is known about the targeting of reaches in natural behavior. A straightforward way in which the target of a reach might be selected is for the subject to visually search the peripheral retina for the desired object, and then to program both the reach and the accompanying saccade on the basis of this information. Experiments on the relative timing of eye and hand movements to a target reveal eye-hand latencies close to zero, consistent with this speculation (Abrams et al, 1990). However, in a typical experiment the target is usually presented at the onset of the trial, and there is little opportunity to locate the target ahead of time, unlike the natural world, where objects are continuously available. When the target is continuously present observers have the opportunity to plan for the arm movement. Such planning is an essential component of motor behavior, and allows speedier movements as well as coordination with other movements, such as the other hand or the body. We measured the latency between eye and hand movements for all the reaches that subjects make. The initiation of both eye and hand movements was taken from the video record, using the first frame on which a translation could be detected. The frequency distributions of eye-hand latencies for seven subjects are

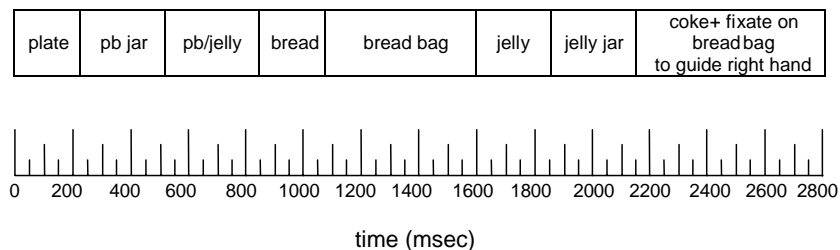


Figure 6. Sequence of fixations made by a subject on first viewing the scene.

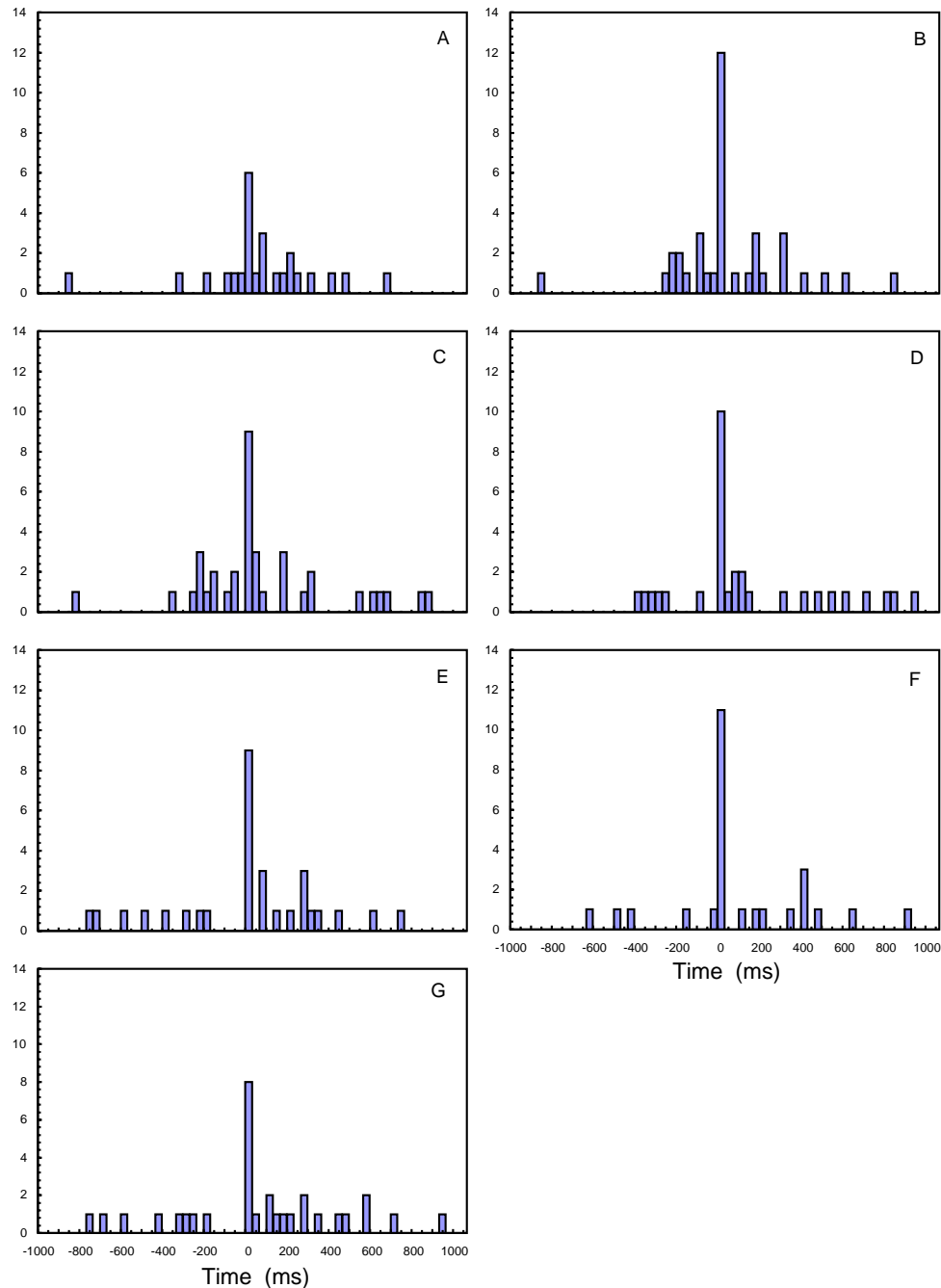


Figure 7. Eye-hand latency distributions for 7 subjects. The hand leads for negative values.

shown in Figure 7. (For a small number of the reaches the hand was not visible in the video at the beginning of the movement and these were omitted.)

Most (0.87 ± 0.03) of the reaching movements were accompanied by a fixation on the target. When the hand movement was not accompanied by a fixation, it was almost always for the purpose of placing an object on the table. Only a very few of the pickup actions were not accompanied by fixation at some stage of the movement. This suggests that foveal information was less critical for

the control of putdown actions. These reaches must have been controlled using either peripheral vision, visual memory, or perhaps somatosensory information about the height of the table. Even when the reach was accompanied by a fixation, there was substantial flexibility in the stage of the reach when the fixation occurred. On a number of occasions, 0.19 ± 0.02 , a substantial fraction of the movement was accomplished without fixation on the target. Although the predominant strategy is for eye and hand to depart close together in time, all subjects

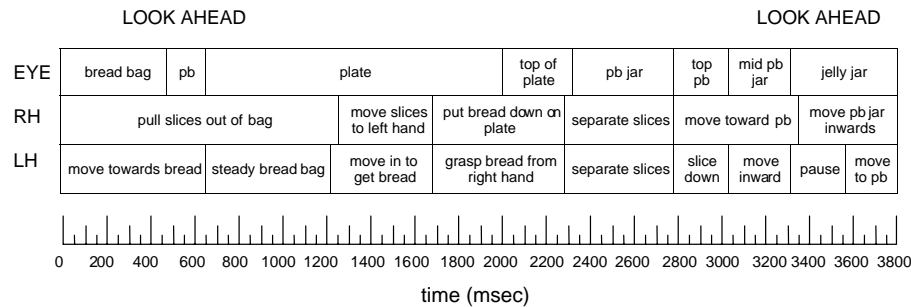


Figure 8. Sequence of eye and hand movements during a segment of the task, showing look-ahead fixations on the peanut butter jar, and later on the jelly jar.

show a number of movements where the reach was initiated well ahead of, or later than, the eye movement, as shown in Figure 7. Presumably, these reaches could be completed without further visual input, or with peripheral guidance. Similarly, the eye frequently fixates the object for as much as a second before the initiation of the reach. These large lags and leads result from the interweaving of visual control of the two hands, with some movements starting while the eye is supervising the other hand's action. An example of this can be seen in Figure 3, where the movement of the left hand towards the lid begins at the same time that the eye and right hand move to put down the knife, about 800 msec after the start of the record. The eye does not move to the lid until about 600 msec later (at 1400 msec), after the right hand movement is complete. These long relative latencies suggest that the next eye or hand movement may be planned as much as a second ahead of time. For example, if fixation of an object is required for final guidance of the reach, the fixation must be planned to some extent when the reach is initiated, so as to be there when needed. Since several fixations intervene between the eye and hand movement to the object, this planning must occur in a representation that is independent of eye position. This can be seen in Figure 3. While the left hand moves to the jelly lid, a fixation is made on the plate, then on the jelly, before the saccade to the jelly lid is initiated. While these arguments are indirect, they make a plausible case for visual representations that span fixations and are maintained over a period of a second or more, to coordinate visually guided movements.

Fixations Prior to Reaching

Reaching behavior provides another clue that a representation of the locations of objects is preserved across fixations. In 11 subjects examined, we found that 0.3 (+/- 0.06) of the reaches that subjects made to pick up objects were preceded by a fixation on that object in the recent past (less than 8 sec). (This is prior to the fixation on the object during the actual reach.) An example of this is given in Movie 3. The subject fixates the jelly while picking up the peanut butter jar lid, fixates it again 1300

msec later while screwing on the lid, and then fixates it for the third time 3,660 msec after the first look, this time maintaining fixation until the reach to the jelly is initiated. These fixations on objects that were to be picked up shortly afterwards may indicate that the subject is planning a reach, and is looking to the object to acquire its spatial location for guiding the next movement. Another example of lookahead fixations is shown in Figure 8. It seems likely that the spatial memory information facilitates the targeting of the saccade and perhaps initiates programming the reach. Similar "look-ahead" fixations were observed by Pelz et al (2001) in a hand-washing context. As subjects approached the wash basin they fixated the tap, soap, and paper towels in sequence, before returning to fixate the tap to guide contact with the hand.



Movie 3. An example of a reach preceded by a fixation.

Discussion

In general, these observations suggest that the visual operations within a given fixation are highly specific to the immediate task. The dependence of fixation location on the immediate task was also observed by Land in the tea-making task (Land, Mennie & Rusted, 1999). Land et

al described performance as a sequence of “object related acts”. Thus the sequence: pick up an object, move it to a new location, put the object down, would constitute an object-oriented action, where fixation would be required for picking up, then for targeting the location for placement and guiding the placement. To pick up an object, observers typically fixate the point on the object where the hand makes contact. Similar step-by-step control of hand actions by fixation at a specific locus in the scene has been demonstrated under more controlled circumstances by Johansson et al (2001) in a task where subjects picked up a bar, moved it past an obstacle, and used it to contact a switch. Fixations clustered at critical loci for each segment of the movement, moving on to the next locus as the action was completed. Other natural behaviors, such as driving, playing cricket, and table tennis also reveal stereotyped fixation patterns for acquisition of information critical to the momentary task needs. In driving, Land has shown that drivers reliably fixate the tangent point of the curve to control steering around the curve (Land & Lee, 1994). In cricket, players exhibit very precise fixation patterns, fixating the bounce point of the ball just ahead of its impact (Land & McLeod, 2000). A similar pattern is seen in table tennis (Land & Furneaux, 1997). In a task where observers copy a pattern of colored blocks, Ballard et al (1995) showed that block color and location are acquired in separate fixations on the pattern, just before block pickup and placement, respectively.

The specificity of the information acquired in different fixations is indicated not only by the ongoing hand actions and the point in the task, but also by the durations of the fixations, which vary over a wide range. 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. In addition to the current observations, other evidence suggests that ongoing task is a primary factor in fixation duration. In the block-copying task, fixations for acquiring block location took about 75 msec longer than those for acquiring color (Hayhoe et al, 1998). In addition, different distributions of fixation duration are observed for reading than for viewing pictorial representations of scenes (Henderson & Hollingworth, 1999; Viviani, 1991). Pelz et al (2000) observed different distributions for three phases of a model-building task. There were three phases of the task: reading the instructions, searching for the pieces, and putting the pieces together, each with a characteristic distribution of fixation durations. Epelboim et al (1995) also observed shorter fixation durations for tapping than simply for looking at a sequence of lights on a table. The argument that the time required for acquisition of the currently needed information can, of course, be made in only the most general terms. In any particular instance, the duration of a fixation will depend on a variety of other factors, such as the time to program the next

saccade, the degree of pre-planning of the next saccade, and the time taken by the hands to complete other aspects of the task such as a manipulation or a reach.

Thus different visual goals require different computations. While such task dependence is to some degree inevitable, the extent to which fixation durations vary moment by moment during task performance 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. The intrinsic salience of scene objects does not appear to be a major factor in attracting fixations in normal vision, and models that depend entirely on salience, such as that of Itti & Koch (2000) cannot be generally applicable. The specificity of the information extracted within a fixation suggests a large degree of independence of the visual computations within individual fixations, to the extent that the particular information extracted does not depend on information from prior fixations. This is consistent with the body of work indicating limited memory across fixation positions. For at least some proportion of the task, observers appear to access the information explicitly at the point of fixation, at the time when it is needed, as opposed to relying on information from prior fixations. This behavior is consistent with O'Regan's suggestion that the scene serves as a kind of external memory that can be quickly accessed when needed (O'Regan, 1992; Ballard et al, 1995).

Integrated Representations for Motor Planning

However, some aspects of natural behavior cannot be accounted for this way. Land & Furneaux (1997) noted the need for some kind of visual buffer both in driving, where the current information controls the steering action about 800 msec later, and in piano playing, where the fixations lead the note played by about a second. In the tea making task, Land et al (1999) also noted a number of instances where objects were found more easily when they had been fixated a few seconds previously. The current observations provide further evidence that memory across fixations is needed as a basis for motor planning and coordination. First, observers consistently scan the scene with a small number of brief fixations before beginning the task. It seems plausible that this provides information about the identity and location of objects in the scene. The existence of a coarse scene representation has been postulated by O'Regan, Irwin, and coworkers (O'Regan & Levy-Schoen, 1983; O'Regan, 1992; Irwin 1991; Irwin Zacks & Brown, 1990). Ullman (1984) also suggested the need for such a representation, which he suggested was extracted using some kind of general-purpose routine. To this point, however, there has been no evidence that observers in fact construct such a

representation in normal viewing. The scanning behavior observed here hints at such a general-purpose representation. However, it would be necessary to observe scanning as a common occurrence when observers view novel scenes, for this argument to have much force. Other evidence shows that information about the spatial organization of scenes is preserved across fixations. For example, De Graef and Verfaillie show encoding of spatial relationships of “bystander” objects that are not the target of a saccade (de Graef et al, 2001; Verfaillie et al, 2001). Melcher & Kowler (2001) showed memory for both the identity and location of about 8 objects in multiple scenes following inspection periods of a few seconds.

It seems likely that one function of an integrated representation is for targeting (and planning) eye and hand movements. In normal viewing, the target is frequently present in the peripheral retina, and can be located on the basis of stimulus features, so it is not obvious that spatial memory from a prior fixation would be useful in target selection. However, other evidence supports the idea that prior fixations facilitate target selection. For example, Epelboim et al (1995) found that the time taken to tap a specified sequence of colored lights arrayed on a table rapidly decreased as the task was repeated. Zelinsky et al (1997) found faster search times and fewer saccades for target objects when subjects were given a pre-view of the spatial array prior to a search task. McPeck & Nakayama (1999), showed that saccades to colored targets have shorter latency if a target of the same color has been presented on the previous trial. A similar result has been found in Frontal Eye Field neurons of monkeys by Bichot & Schall (1999). Chun & Jiang (1998, 1999), also showed that visual search is facilitated by prior exposure to the spatial context.

The behavior of observers in this study is consistent with the suggestion that a spatial memory representation is used in targeting eye and hand movements. The fixation distributions revealed an unexpectedly large number of very short fixations in the range 70-130 msec. This is much shorter than the time normally required to program a new saccade. Saccades evoked by a sudden onset typically occur with a latency of 200-250 msec. Thus these very short fixations show that observers must pre-program two or more of the saccades. Zingale & Kowler (1987) have demonstrated that saccades can be pre-programmed by showing that the latency to initiate a sequence of saccades increased with the number of saccades in the sequence. Very brief fixations have also been observed in circumstances where two targets are in competition, such as a double-step task (Becker & Jurgens, 1979). Theeuwes and colleagues (Theeuwes et al, 1998, 1999; Irwin et al, 2000) observed short fixations to a distractor stimulus that was suddenly presented when subjects were preparing to saccade to a search target. They interpreted these brief fixations as the consequence of a concurrently programmed second saccade to the target, which terminated the fixation on the distractor. McPeck

et al (2000) also demonstrated concurrent programming of saccades in a similar situation where two targets were in competition. Saccades to the wrong stimulus were often followed by a second saccade to the correct stimulus with a very brief inter-saccadic interval. The frequency of very short fixations we observe in the sandwich task indicate that pre-programming, or concurrent programming of more than one saccade, is a common occurrence in ordinary movements, and not restricted to particular experimental situations. The significance of pre-planning is that programming of the second (and subsequent) saccade in a sequence must initially occur in a reference frame that is independent of the eye, and the second saccade is using information acquired prior to the immediately preceding fixation. This implies the existence of some form of spatial memory representation that is precise enough to support saccadic targeting. McPeck & Keller (2002) observed that neurons in the superior colliculus show activity related to preparation of the second saccade even while the first saccade is still in progress. Thus neural activity for more than one saccade can be maintained concurrently, even at levels close to the motor output, and the neural activity for the second saccade must be able to take into account the eye displacement by the first saccade. Freedman et al (1996) have demonstrated that cells in the superior colliculus code gaze position in space (or with respect to the body), not retinal error. Thus the intrinsic organization of the saccadic system appears to be in spatial coordinates.

Under special conditions, the latency of saccades evoked by a flashed stimulus can be in the 70-130 msec range, comparable to the fixation durations observed here. These are called express saccades (Fischer & Ramsberger, 1984). These are seen when the fixation point is turned off before the stimulus appears, and usually involves substantial practice. Usually the stimulus is in one of two positions left or right of fixation. These saccades are commonly thought to be a special kind of saccade that by-passes high level cortical control mechanisms (Fischer & Breitmeyer, 1987). However, the high frequency of very brief fixations in natural contexts suggests that planning is a fundamental aspect of saccade programming, and that the short latencies observed in the express saccade paradigm are simply a result of motor planning. This is consistent with the suggestion of Kowler (1991), and is supported by recordings of the Superior Colliculus, that indicate some metrical preparation indicated by increased activity in build-up neurons (Munoz & Wurtz, 1995).

Reaching movements also suggest the existence of a spatial representation independent of eye position. On a number of occasions a reaching movement was initiated up to a second ahead of the eye movement to the target. As indicated in Figure 3, several fixations could intervene between the initiation of the movement and the grasp. This means that the programming and control of the reach was accomplished with the eye in two or more

positions with respect to the scene, and the reach must be guided either by a spatial memory representation, or a visual representation that is independent of eye position. This is interesting because neurophysiological evidence from cells in the intraparietal sulcus suggests that reaching movements are programmed in an eye-centered coordinate frame (Batista et al, 1999). The above evidence suggests, instead, that this coordinate frame must be exocentric, rather than eye-centered.

In general, the relationship between the eye and hand movements is much more flexible than expected on the basis of previous experimental work. The wide range of eye-hand latencies is very different from the usual single trial experiments where the eye-hand latency is usually close to zero (Abrams et al, 1990). This difference is presumably a consequence of the opportunity for motor planning afforded by the continuous presence of the scene, as well as the need to interleave control of the two hands. Despite the wide range of relative latencies, it is interesting to note the predominance of latencies close to zero, suggesting a preference for simultaneously initiated, synergistic movements. Land et al (1999) also measured eye-hand latencies in their tea-making task. They also observed a number of long lead times for the hand, although their distribution was strongly biased toward positive values, where the eye leads the hand.

The frequent looks to objects, a few seconds prior to reaching for them, are suggestive of movement planning in a spatial coordinate frame. These fixations on objects that were to be picked up shortly afterwards may indicate that the subject is looking to the object to acquire its spatial location for guiding the next movement. As described above, similar "look-ahead" fixations were observed by Pelz et al (2001) in a hand-washing context. The high frequency of these look-ahead fixations in the present task, as well as in hand-washing and tea-making Land et al (1999), suggest that this is a ubiquitous aspect of natural behavior. Pelz et al interpreted these look-ahead fixations in terms of their perceptual role, suggesting that they provide continuity of perceptual experience. It also seems likely that fixating the location of a future target facilitates the programming of the saccade, and perhaps initiates programming the reach (McPeck & Nakayama, 1999; Chun & Nakayama, 2000; Zelinsky et al, 1997). It is known that accurate saccades can be made on the basis of memory for stimulus location when the original stimulus is no longer present (eg Miller, 1980; Gnadt et al, 1991; Hayhoe et al, 1992; Colby, 1998). However, in normal viewing, the target is continuously present in the peripheral retina, and can be located on the basis of stimulus features, so it is not obvious that spatial memory would be useful in target selection. Its usefulness becomes more apparent when the need for motor planning is taken into account. In the case of reaching movements, the slower velocity of the arm relative to eye movements makes early initiation of the arm movements particularly useful.

Conclusions

In conclusion, examination of eye and hand movements in natural behavior suggests that much of what the visual system has to do is computed at the moment it is needed for the particular task, and does not appear to be heavily dependent on information acquired in prior gaze positions, in agreement with prior work (Ballard et al, 1995). Thus the limitations of short term memory, and the related susceptibility to change blindness may not be much of a limitation for normal visual function. However, there must be some scene representation that corresponds to perceptual awareness. O'Regan (1992) and Irwin (1991) have postulated that there is some integrated representation of the scene, but suggest that the representation of spatial information is imprecise and that the representation is semantic in nature. The evidence presented here, however, supports the suggestion of Chun & Nakayama (2000) that the spatial information cannot be imprecise, but must be able to support high precision movements.

Acknowledgments

This research was supported by National Institutes of Health Grants EY-05729 and RR-09283 Thanks to Chris Chizk for assistance with the experiments. Commercial relationships: None.

References

- Abrams, R., Meyer, D., & Kornblum, S. (1990). Eye-hand coordination: Oculomotor control in rapid aimed limb movements. *Journal Experimental Psychology: Human Perception and Performance*, 15, 248-267. [PubMed]
- Ballard, D, Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Cognitive Neuroscience*, 7, 66-80.
- Batista, A., Buneo, C., Snyder, L., & Andersen, R. (1999). Reach plans in eye-centered coordinates. *Science*, 285, 257-260. [PubMed]
- Becker, W., & Fuchs, A. F. (1969). Further properties of the human saccadic system: Eye movements and correction saccades with and without visual fixation points. *Vision Research*, 9, 1248-1258. [PubMed]
- Becker, W., & Jurgens, R. (1979). An analysis of the saccadic system by means of double-step stimuli. *Vision Research*, 19, 967-983. [PubMed]
- Bichot, N.P. & Schall, J.D. (1999) Effects of similarity and history on neural mechanisms of visual selection. *Nature Neuroscience*, 2, 549-554. [PubMed]

- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36, 28-71. [PubMed]
- Chun, M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, 10, 360-365.
- Chun, M., & Nakayama, K. (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition*, 7, 65-82.
- Colby, C. L. (1998) Action-oriented spatial reference frames in cortex. *Neuron*, 20: 15-24. [PubMed]
- De Graef, P., Verfaillie, K., & Lamote, C. (2001). Transsaccadic coding of object position: Effects of saccadic status and allocentric reference frame. *Psychologica Belgica*, 41, 29-54.
- Epelboim, J., Steinman, R., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C., & Collewijn, H. (1995) The function of visual search and memory in sequential looking tasks. *Vision Research*, 35, 3401-3422. [PubMed]
- Fischer, B., & Breitmeyer, B. (1987). Mechanisms of visual attention revealed by saccadic eye movements. *Neuropsychologia*, 25, 78-83. [PubMed]
- Fischer, B., & Ramsperger, E. (1984). Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, 57, 191-195. [PubMed]
- Folk, C., Remington, R., & Johnston, J. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal Experimental Psychology: Human Perception and Performance*, 18, 1030-1044. [PubMed]
- Freedman, E., Stanford, T., & Sparks, D. (1996) Combined eye-head gaze shifts produced by electrical stimulation of the superior colliculus in rhesus monkeys. *J. Neurophysiol.* 76, 927-952. [PubMed].
- Gibson, B., & Jiang, Y. (1998). Surprise! An unexpected color singleton does not capture attention in visual search. *Psychological Science*, 9, 176-182.
- Gnadt, J., Bracewell, R., & Andersen, R. (1991) Sensorimotor transformation during eye movements to remembered visual targets. *Vision Research*, 31, 693-715. [PubMed]
- Hayhoe, M. M. (2000). Vision using routines: a functional account of vision. *Visual Cognition*, 7, 43-64.
- Hayhoe, M., Bensinger, D., & Ballard (1998). Task constraints in visual working memory. *Vision Research*, 38, 125-137. [PubMed]
- Hayhoe, M., Lachter, J., & Moeller, P. (1992). Spatial memory and integration across saccadic eye movements. In K. Rayner (Ed), *Eye movements and visual cognition: Scene perception and reading* (pp. 130-145). New York: Springer-Verlag.
- Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp.261-283).Berlin: Springer.
- Henderson, J., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, 10, 438-443.
- Hochberg, J. (1986). Representation of motion and space in video and cinematic displays. In K. Boff, L. Kauffman, & J. Thomas (Eds), *Handbook of perception and human performance* (Vol. 1, pp. 22.21-22.64). New York: Wiley.
- Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology*, 23, 420-456. [PubMed]
- Irwin, D. (1992). Memory for position and identity across eye movements. *Journal Experimental Psychology: Learning, Memory, & Cognition*. 18, 307-317.
- Irwin, D. (1996) Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, 5, 94-100.
- Irwin, D. & Gordon, R. (1998) Eye movements, attention, and trans-saccadic memory. *Visual Cognition*, 5, 127-155.
- Irwin, D. E., Zacks, J. L., & Brown, J. S. (1990). Visual memory and the perception of a stable visual environment. *Perception and Psychophysics*, 47, 35-46. [PubMed]
- Irwin, D., Colcombe, A., Kramer, A., & Hahn, S., (2000) Attentional and oculomotor capture by onset, luminance, and color singletons. *Vision Research*, 40, 1443-1458. [PubMed]
- Itti, L. & Koch, C. (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489-1506. [PubMed]
- Jiang, Y., Olson, I. R., & Chun, M. M. (2000) Organization of visual short-term memory. *Journal Experimental Psychology: Learning, Memory, and Cognition*. 26, 683-702. [PubMed]
- Johansson, R., Westling, G., Backstrom, A., Flanagan, J. R. (2001) Eye-hand coordination in object manipulation. *J. Neuroscience*, 21, 6917-6932. [PubMed]

- Kowler, E. (1991) The role of visual and cognitive processes in the control of eye movement. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes* (Reviews of Oculomotor Research, Vol. 4, pp. 1-70). Amsterdam: Elsevier.
- Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, 369, 742-744. [PubMed]
- Land, M., & Furneaux, S. (1997) The knowledge base of the oculomotor system. *Philosophical Transactions, Royal Society of London, Series B*, 352, 1231-1239. [PubMed]
- Land, M. F. & McLeod, P. (2000) From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience*, 3, 1340-1345. [PubMed]
- Land, M., Mennie, N., & Rusted, J. (1999). Eye movements and the roles of vision in activities of daily living: making a cup of tea. *Perception*, 28, 1311-1328. [PubMed]
- Land, M. (1996) The time it takes to process visual information while steering a vehicle [Abstract]. *Investigative Ophthalmology & Visual Science*, 37, S525.
- Levin, D., & Simons, D. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, 4, 501-506.
- Mack, A., & Rock, I. (1996). *Inattention blindness*. Cambridge, MA: MIT Press.
- McConkie, G., & Currie, C. (1996). Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 563-581. [PubMed]
- McPeck, R., & Keller, E. (2001a). Short-term priming, concurrent processing, and saccade curvature during a target selection task in the monkey. *Vision Research*, 41, 785-800. [PubMed]
- McPeck, R.M., & Keller, E.L. (2002) Superior colliculus activity related to concurrent processing of saccade goals in a visual search task. *Journal of Neurophysiology*, 87, 1805-1815. [PubMed]
- McPeck, R., Maljkovic, V., & Nakayama, K. (1999). Saccades require focal attention and are facilitated by a short-term memory system. *Vision Research*, 39, 1555-1565. [PubMed]
- McPeck, R., Skavenski, A., & Nakayama, K. (2000). Concurrent processing of saccades in visual search. *Vision Research*, 40, 2499-2516. [PubMed]
- Maljkovic, V., & Nakayama, K. (1994) Priming of pop-out: I Role of features *Memory & Cognition*, 22, 657-672. [PubMed]
- Melcher, D., & Kowler, E. (2001) Visual scene memory and the guidance of saccadic eye movements. *Vision Research*, 41, 3597-3611. [PubMed]
- Miller, J. (1980) The information used by the perceptual and oculomotor systems regarding the amplitude of saccadic and pursuit eye movements. *Vision Research*, 20, 59-68. [PubMed]
- Munoz, D. & Wurtz, R. (1995) Saccade-related activity in monkey superior colliculus: II. Spread of activity during saccades. *Journal of Neurophysiology*, 73, 2334-2348. [PubMed]
- O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal Psychology*, 46, 461-488. [PubMed]
- O'Regan, J. K., & Levy-Schoen, A. (1983). Integrating visual information from successive fixations: Does trans-saccadic fusion exist? *Vision Research*, 23, 765-769. [PubMed]
- O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of "mudsplashes." *Nature*, 398, 34. [PubMed]
- O'Regan, J. K., Deubel, H., Clark, J., & Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7, 191-211.
- Pelz, J. B., Canosa, R., Babcock, J., Kucharczyk, D., Silver, A., & Konno, D. (2000). Portable Eyetracking: A Study of Natural Eye Movements, *Proceedings SPIE Volume 3959, Human vision and electronic imaging V* (pp. 566-582). Bellingham, WA: SPIE. [Abstract]
- Pelz, J. B., & Canosa, R., (2001). Oculomotor Behavior and perceptual strategies in complex tasks. *Vision Research*, 41, 3587-3596. [PubMed]
- Pollatsek, A., & Rayner, K. (1992). In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 166-191). New York: Springer-Verlag.
- Pylyshyn, Z. (1989). The role of location indices in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, 32, 65-97. [PubMed]
- Rayner, K., & Pollatsek, A. (1983) Is visual information integrated across saccades? *Perception & Psychophysics*, 34, 39-48. [PubMed]
- 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. *Journal Psychological Science*, 8, 368-373.
- Simons, D. (1996). In sight, out of mind: When object representations fail. *Psychological Science*, 7, 301-305.
- Simons, D. J. (2000). Change blindness and visual memory [Special issue]. *Visual Cognition*, 7, Psychology Press, Hove, UK.
- Simons, D., & Levin, D. (1997). Change blindness. *Trends in Cognitive Science*, 1, 261-267.

- Simons, D., & Levin, D. (1998). Failure to detect changes to people in real-world interactions. *Psychonomic Bulletin & Review*, *5*, 644-649.
- Theeuwes, J., Kramer, A., Hahn, S., & Irwin, D. (1998). Our eye do not always go where we want them to go: capture of the eyes by new objects. *Psychological Science*, *9*, 379-385.
- Theeuwes, J., Kramer, A., Hahn, S., Irwin, D., & Zelinsky, G. (1999). Influence of attentional capture on oculomotor control. *J. Experimental Psychology: Human Perception & Performance*, *25*, 1595-1608. [[PubMed](#)]
- Verfaillie, K., De Graef, P., Germeys, F., Gysen, V., & Van Eccelpoel, C. (2001). Selective transsaccadic coding of object and event-diagnostic information. *Psychologica Belgica*, *41*, 89-114.
- Wallis, G., & Büthoff, H. (2000). What's scene and not seen: Influences of movement and task upon what we see. *Visual Cognition*, *7*, 175-190.
- Ullman, S. (1984). Visual routines. *Cognition*, *18*, 97-157. [[PubMed](#)]
- Viviani, P. (1991). In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes* (Reviews of oculomotor Research, Vol. 4, pp. 1-70). Amsterdam: Elsevier.
- Zelinsky, G, Rao, R. Hayhoe, M. & Ballard, D. (1997) Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, *8*, 448-453.
- Zingale, C. M., & Kowler, E. (1987). Planning sequences of saccades. *Vision Research*, *27*, 1327-1341. [[PubMed](#)]