

# Vision Using Routines: A Functional Account of Vision

Mary Hayhoe

*Center for Visual Science, University of Rochester, Rochester, New York, USA*

This paper presents the case for a functional account of vision. A variety of studies have consistently revealed “change blindness” or insensitivity to changes in the visual scene during an eye movement. These studies indicate that only a small part of the information in the scene is represented in the brain from moment to moment. It is still unclear, however, exactly what is included in visual representations. This paper reviews experiments using an extended visuo-motor task, showing that display changes affect performance differently depending on the observer’s place in the task. These effects are revealed by increases in fixation duration following a change. Different task-dependent increases suggest that the visual system represents only the information that is necessary for the immediate visual task. This allows a principled exploration of the stimulus properties that are included in the internal visual representation. The task specificity also has a more general implication that vision should be conceptualized as an active process executing special purpose “routines” that compute only the currently necessary information. Evidence for this view and its implications for visual representations are discussed. Comparison of the change blindness phenomenon and fixation durations shows that conscious report does not reveal the extent of the representations computed by the routines.

## INTRODUCTION

Our conscious perception of the visual world suggests that the task of the visual system is to create some global representation of the space and objects around us. However, a variety of studies reveal that observers are quite insensitive to a change in the visual scene when the change is made during a saccadic eye movement, or in the presence of some kind of masking stimulus. Many of these studies have been recently reviewed by Simons and Levin (1997) and the

---

Please address all correspondence to M. Hayhoe, Center for Visual Science, University of Rochester, Rochester, NY14627, USA. Email: mary@cvs.rochester.edu

I am indebted to my collaborators in the work discussed here: Dana Ballard, Jeff Pelz, and David Bensinger. Supported by NIH Grant EY05729 and R24 RR06853.

phenomenon has been described as “change blindness”. The implication of these studies is that failure to notice a change in the visual display means that particular piece of information is not part of the internal visual representation. This raises the question of how well the internal representation of the visual world reflects the information in the retinal image. What are the implications of this for perception? If visual representations are limited, as implied by the change blindness results, how do we go about determining what information is actually included in the internal representation, and ‘processed’ by the visual system? The visual display can be manipulated in many ways, such as changing objects, stimulus properties, or regions. What manipulations should be chosen and how can we evaluate the magnitude or importance of a change for the visual representation? The suggestion in this paper is that the content of the representation varies from moment to moment in concert with the requirements of the ongoing visual tasks. Thus image changes should not be noticed if they are made to task-irrelevant information. The task specificity of visual representations also has a more general implication, that vision should be conceptualized as an active process that extracts only the information from the visual stimulus that is currently needed. This functional approach differs from the more traditional approach to vision, whereby the brain is thought to reconstruct a general-purpose representation of the information in the scene. This paper explores the evidence for a functional approach and its implications for understanding visual representations.

## THE ROLE OF VISION IN EVERYDAY LIFE

To appreciate the importance of task context, consider an ordinary circumstance like making a snack: For example, a peanut butter and jelly sandwich and a glass of cola. How is vision used to achieve this? Following up on a study by Land, Mennie, and Rusted (1998), who recorded fixation patterns while making a cup of tea, we monitored people’s eye movements while they made a sandwich and poured a cola.<sup>1</sup> Over the 2-minute period that it took to make the sandwich, gaze was almost exclusively directed on the objects involved in the task, as noted by Land in his study. In the entire sequence of approximately 250 fixations, only one or two were to irrelevant parts of the scene. A moment-by-moment record of the sequence of fixations and the actions of the two hands was made for the entire period. A description of a segment of the behaviour part way through the task is shown in Fig. 1. The subject fixates the bread for about 500msec to guide placement of the bread on the plate. Gaze is then transferred to the peanut butter jar for a period of 1400msec, first to guide the left hand to

---

<sup>1</sup>Subjects were given no specific instructions as to how to do this. Eye movements were monitored by a head-mounted ISCAN infra-red video based eye tracker, and the data were in the form of a video record from the subject’s viewpoint with the direction of gaze superimposed on the tape.

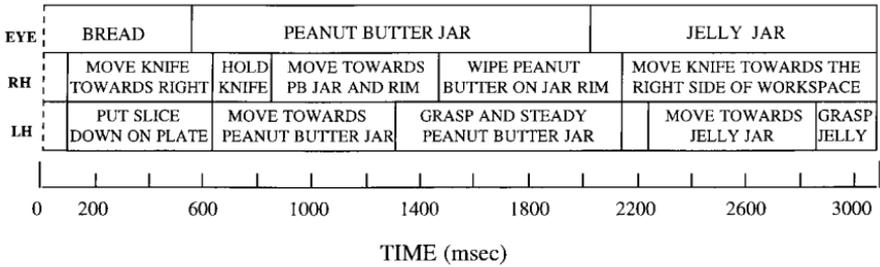


FIG. 1. A segment of behaviour in making a peanut butter and jelly sandwich. The actions of the eye and hands are shown as a function of time.

grasp the jar and then to guide the right hand to wipe off the excess peanut butter on the rim. While this is in progress, gaze is transferred to the jelly jar in order to guide the subsequent movement of the left hand (initiated 200msec after the gaze change) to grasp the jar, and so on. It can be seen that fixations are very tightly locked to the task. Thus vision is being used for locating objects (peanut butter jar, jelly jar) or specific locations of objects, for guiding the reaching movements and the preshaping of the hand for the grasp, for monitoring the spreading of the peanut butter, and for guiding the position of an object in one hand to a particular location relative to the object in the other hand (the knife to the rim). The role of vision from moment to moment is determined almost exclusively by the current stage in accomplishing the task. There appears to be little room for other functions. Performance is also highly serialized. Subjects rarely looked ahead to acquire information required for upcoming actions, such as a fixation on the jelly jar while spreading the peanut butter, returning to the peanut butter and then back to the jelly for pickup.

Another compelling feature of the behaviour was the similarity between subjects. The sequence of fixations and reaches were almost identical for four subjects despite the unspecific nature of the instructions. The locations of the fixations on the objects were also very reproducible between subjects, for example, subjects fixate the mouth of the bottle when pouring and then transfer gaze to the level of cola in the glass about halfway through. Thus, many details of the fixations, and by inference the ongoing visual computations, are governed by the task goals, together with the physical constraints of the world. The goal of vision can be seen as the active extraction of specific, task relevant information, and the particular information being extracted is indicated by the particular location of fixation and the immediate behavioural context. Perhaps more importantly, the complete task appears to be composed of the sequential application of a small number of special purpose behavioural primitives executed in different contexts; for example, locating the next object needed, fixating that object, guiding the hand for pickup, monitoring a state (such a level of the cola), and so on. Each of these primitives requires the extraction of very

specific information from the visual display. Since these operations are the ones we usually think of as visual perception, how much non-specific information is represented in addition?

## WHAT ARE VISUAL ROUTINES?

If only limited information in the image is used for vision at a particular point in time, there must be some way of selecting it in a way that is appropriate to the current circumstances. One way of doing this is by special purpose routines. The idea of visual routines was first introduced by Ullman (1984), to describe the perception of spatial relationships, such as the apparently effortless perceptual process of judging whether a point is inside a closed curve. The essential property of a routine was that it instantiated a *procedure*<sup>2</sup> as opposed to requiring a specialized detector of some kind. For example, seeing if two points are on the same contour can be done using an operation such as tracing the contour. A procedure such as this can cope with a large variety of spatial relations and do so in a computationally efficient manner—two of Ullman's requirements for a routine. Ullman proposed a set of basic operations used in assembling routines, such as a shift of the processing focus, as might occur when fixation or attention is directed to a location. Another basic operation was to “mark” an environmental location, as might occur when the information at that location is required in a subsequent operation. Thus, many quite elementary (low level) visual operations may require specialized computations, since it would be impossible to do all possible computations in anticipation of their need—they must be done on demand. In the peanut butter and jelly sandwich example, the primitives such as “locate the next object”, “monitor a variable or state”, or the extraction of visual properties for pre-shaping grasp, would constitute commonly used routines. For example in filling the cup, only the level of cola in the cup has to be monitored. Given the context of the white cup and dark brown cola, this becomes a very simple computational task. This is a very different way of conceptualizing vision that emphasizes its functional characteristics rather than its structural properties. In this framework, understanding the nature and composition of the routines becomes one of the central issues in understanding vision. The present paper argues that it is necessary to examine vision in the context of a well-defined ongoing task in order to understand the routines, and consequently the composition of visual representations.

---

<sup>2</sup> A similar procedural approach was proposed by Just and Carpenter (1976), who examined visual tasks such as mental rotation. Ullman's emphasis is on somewhat lower level perceptual processes, but the ideas are closely related.

## VISION AS A DYNAMIC PROCESS

The specificity and time-varying nature of visual computations revealed in the example of making a sandwich shows the need to observe extended behavioural sequences in order to understand visual representations. To think about vision in terms of functional demands, we need to distinguish between processes that operate at different time scales. It has been pointed out by Newell (1990) that any complex system, such as the brain, must be organized hierarchically in space and time. We are familiar with the conceptualization of vision as a spatial hierarchy, but are less used to distinguishing different temporal levels in the hierarchy. Table 1 shows a modified version of Newell's conceptualization of the brain's temporal hierarchy that underlies behaviour at different time scales, and at correspondingly different levels of abstraction. Basic cognitive functioning, for example, reasoned decision making, takes place at the time scale of tens of seconds. Sensorimotor tasks, such as dialing a telephone number, span several seconds, requiring working memory and occur over several fixation positions. The visual processing that occurs within a given fixation, such as locating a search target and initiating an eye movement, takes a few hundred milliseconds. The basic neural operations underlying this, such as activating a visual area, occur on a shorter time scale of tens of milliseconds.

Although there is some ambiguity about how to categorize, in general, the different time scales, we can see that in the visual system there is an important difference in processes that operate within a fixation, and those that operate between fixations. Processes operating within a fixation, such as simple object recognition, are the ones most usually studied. Those that operate across fixations are central to the issue of change blindness, since detection of a change requires the comparison of perceptual events across different fixation positions or between visual stimuli separated in time by some masking event.

Understanding change blindness and the events that occur on the time scale of fixations presents a challenge. As is clear from the peanut butter and jelly sandwich example, vision ordinarily operates in the presence of ongoing goal-directed movements. This means that many visual behaviours span several fixations. In the spatial domain observers must maintain constancy of visual direction across different eye and head positions as a basis for co-ordinated actions.

TABLE 1

Newell's temporal hierarchy, showing different functional levels and time scales

<i>Time</i>	<i>Process</i>	<i>Description</i>
10sec	Cognition	Reasoned decision making
1sec	Working memory	Sensory-motor tasks, e.g. driving
300msec	Visual routines	Extract context-dependent information, e.g. visual search
80msec	Neural operations	Elemental sensory input, e.g. spatial filters

In the temporal domain, visual information acquired in one gaze position must be related in some way to the previous ones. Thus, to understand visual behaviour we must understand the transition between events operating on the time scale of a few hundred milliseconds to those on the scale of a few seconds. These time scales correspond to the operation of visual routines and to the consequent ongoing behaviour.

## EVIDENCE FOR TASK SPECIFICITY

The central issue for a functional account of vision is: What is the evidence that the representations that guide ongoing behaviour are task specific? To make the case for task-specific representations I first consider the computational and physiological evidence that is consistent with this idea, and then some of our own recent experimental work.

### Computational argument

The crucial advantage of task-specific routines is computational efficiency. Segmenting images and representing even simple properties of objects and scenes reliably is computationally very complex. In the case of human vision the retinal image is changing every several hundred milliseconds as the observer changes gaze. Thus, the relevant information must be extracted very quickly. However, the continuous redirection of gaze to the region of interest can be used to advantage. Direction of gaze can be used to specify the location of the needed information in the visual scene and the time when it is needed. This makes complex internal representations unnecessary because the observer/autonomous system can acquire the necessary information from the scene during performance of the task, rather than using the information inherent in an internal model. Consequently this approach to computational problems is known as Active Vision (Bajcsy, 1988; Ballard, 1991; Brooks, 1986) and seems well suited for modelling human vision. The argument for this important computational role of fixation is developed more extensively in Ballard, Hayhoe, Pook and Rao, (1998; see also Just & Carpenter, 1976; Ullman, 1984).

It might be argued that such minimal representations are inadequate to support the complexity of visual perception. Simple representations may merely reflect the requirements of simple behaviours. Thus, it is important to establish if the normal range of visual behaviours is possible from limited representations. One demonstration that this is possible is in driving, where McCallum (1995) demonstrated that a very limited set of sensory information was sufficient to allow a simulated vehicle to learn to drive successfully on a freeway, passing slower cars in front and avoiding faster cars behind. The vehicle's perceptual world was limited to eight perceptual variables, each with two or three

possible values (e.g. gaze object = car, shoulder, road; gaze direction = forward, backward, and so on). Indeed, limited representations are probably advantageous in this case as it limits the complexity of the learning problem (Ballard et al., 1998).

### Physiological argument

A visual system that uses special purpose routines is intrinsically top down, in the sense that the way incoming stimuli are handled depends on the current state of the visual system as determined by the ongoing behaviour. Is this consistent with what we know about the activity of the visual system? It is becoming increasingly evident that the operation of the visual cortex depends critically on the ongoing behaviour. Since Mountcastle's seminal experiments demonstrating the importance of behavioural context on visual responses of single units in area 7 in the posterior parietal cortex, a large number of studies have revealed the ubiquitous effects of attention and behavioural context on the responses of cells in extrastriate visual areas such as MT, MST, V4, V2, and even as early as primary visual cortex (reviewed in Gilbert, 1998). Recent findings of Gilbert and colleagues are particularly dramatic. The spatial tuning of V1 cells in behaving monkeys depended on the perceptual judgement required by the animal, as well as to the spatial distribution of attention (Crist, Ito, Westheimer, & Gilbert, 1997). These task-driven effects interact with the effects of spatial context outside the classical receptive field, indicating the need to consider the operation of visual cortex as a network, rather than an invariant, spatially localized response to specific features. Additionally, in a series of studies, Anderson and his colleagues have demonstrated that the activity of cells in lateral intraparietal area (LIP) and 7a encode the intention to change gaze or point to a target (Snyder, Babista, & Andersen, 1997). An elegant study by Gottlieb, Kusunok, and Goldberg (1998) showed that cells in LIP had little or no response to stimuli brought into their receptive fields by a saccade unless they were salient for the animal. Saliency could be established either by making a stimulus task relevant, or by its sudden onset in the scene, an event that would normally attract the animal's attention. Interestingly, the firing of the cell was not strictly linked to presence of the stimulus in the receptive field, but began before the stimulus was brought into the receptive field by the eye movement, and continued even when a second eye movement to fixate this stimulus took it out of the receptive field. Thus, the firing is tied to the spatial location of the relevant stimulus and reflects something more than a simple enhancement of ongoing activity.

Evidence from brain imaging studies also clearly point to the task specificity of cortical activity. PET studies by Corbetta, Miezin, Dohmeyer, Schulman, and Petersen (1991) conducted while subjects performed visual search tasks, show different regions of activity in visual cortex during a search for a colour,

compared with search based on another property such as motion or shape. This directly reveals that the same visual stimuli can be associated with very different activity patterns depending on the behavioural goal, even when a simple visual judgement such as colour is required. Interestingly, this differential activation of extrastriate cortex was not present in a divided attention task where the target item could be either colour, motion, or shape. In this case it appeared that the selection was done in the dorsolateral prefrontal cortex, a region not activated in the selective attention condition. Thus, the entire set of regions involved in the task, as well as activation in the extrastriate regions, depended on task structure. Recent fMRI experiments also show task specific activation of hMT+ (the human analogue of MT in primates), where activity was graded depending on the task (Beauchamp, Cox, & DeYoe, 1997). Thus, hMT+ was activated most when speed information from a peripheral spatial location was required, less when the colour of stimuli in this location was required, and least when a luminance discrimination at the fixation point was required. Each of these tasks requires different involvement of hMT+

## Psychophysics

We have recently accumulated a body of evidence for the specificity of visual representations (Ballard, Hayhoe, & Pelz, 1995; Ballard, Hayhoe, Li, & Whitehead, 1992). Much of this has been done in the context of a block-copying task. Observers copied a pattern of coloured blocks, as shown in Fig. 2. The model pattern is at the top left, the workspace for building the copy directly below it, and blocks for use in making the copy are in the resource area on the right. Subjects move the blocks with the mouse. Similar experiments with real blocks have also been performed. In this case the observers moved the blocks with their hand, and were freely moving. In this task subjects use frequent eye movements to the model pattern to acquire information just when it is needed. Subjects frequently looked twice at a particular block in the model in the process of copying it, in preference to using visual memory, even for a very short interval. The eye and hand traces demonstrating such a sequence are shown in Fig. 2. This suggests that observers prefer to acquire information just as it is needed, rather than holding an item in memory. A plausible interpretation of this is that colour is acquired in the first fixation as a basis for the search and pickup of a block of that colour. The other fixation, following pickup, is presumably for finding the relative location of that block in the model. That is, colour and location of a single block may be acquired in separate fixations, rather than being bound together as object properties during a single fixation. This points to extremely reduced visual representations and minimal visual information maintained from prior fixations, consistent with the results of the change blindness experiments.

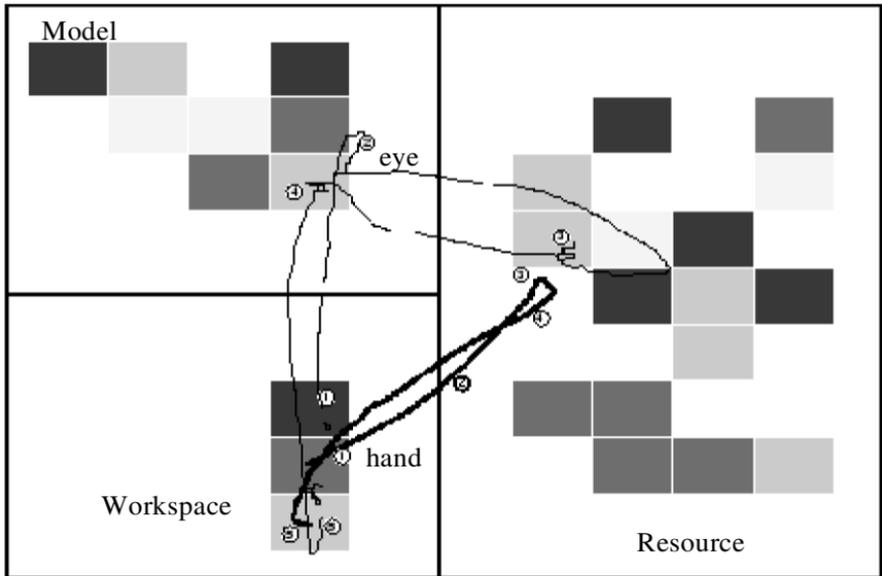


FIG. 2. Display for copying the pattern of blocks in the model area (top left), using blocks picked up in the resource area on the right. The copy is made in the workspace (bottom left) using the mouse to move the blocks. The narrow trace shows a typical eye movement sequence while copying a block, and the thick trace shows the associated cursor movements.

Such inferences from the particular fixation pattern chosen by observers are not in themselves very rigorous. For example, a fixation on the model pattern may be simply a verification of the information while waiting for the slower hand to move to the putdown location. In one test of this hypothesis we restricted the time that the model pattern was visible during the task (Bensinger, 1997). The pattern became invisible during the period after the observer had picked up the blocks and was placing them in the copy, making second looks before putdown impossible. What observers did in this case was to look longer at the model pattern when it was visible *before* pickup, presumably to compensate for the lack of availability during block placement. The total time looking at the model pattern was almost identical for the two paradigms. This suggests that the model fixations serve a purpose in task performance, and are not simply an epiphenomenon of some kind. Other evidence that the fixations are purposive is that making a model fixation delays initiation of the hand movement until the eye becomes available for guiding it. This issue is explored further in Ballard et al. (1995), where we show that fixations and memory can be traded off, depending on the precise conditions of the experiment.

A stronger validation of the idea that the representations are transient and geared to the immediate task needs is given by another recent experiment, where we made saccade-contingent changes during task performance (Hayhoe,

Bensinger, & Ballard, 1998). Subjects performed the block-copying task illustrated in Fig. 2, and in addition, changes were made in the display at two particular points in the task, as shown in Fig. 3. In the first condition, shown in Fig. 3(a), a display change was triggered by a saccade from the workspace area to the model pattern. One of the uncopied blocks was randomly chosen and its colour changed. On some trials this corresponded to the next block to be copied (that is, the target of the saccade). The arrow in the figure shows the saccade, and the zigzag shows when the change is made. In another condition shown in Fig. 3(b), the change was made after the subject had picked up a block and was returning to look at the model immediately before putdown. The logic was that the *same* change occurring at different points in the task should have *different* consequences, if the ongoing representations are indeed acquired for the immediate task needs. In the first condition, we hypothesized that the observer was looking to the model pattern to acquire the colour information for the next block to copy. If the observer has not stored this information in memory from previous views and needs to access it at this point in the task, then a colour change made before arrival should have no consequences. If the change occurs following pickup of a block, however, we might expect some interference in task performance because the block in that location no longer matches the block in hand.

Figure 4(a) shows that the two changes indeed have different effects. The figure shows the duration of the fixation in the model area subsequent to a colour change to the block currently being copied, compared with the control condition when no changes were made. When the change was made after pickup the duration of the subsequent fixation substantially increased over the control (103msec), and had a greater effect than the same change made before the initial fixation on the block (43msec). Thus the same stimulus change has

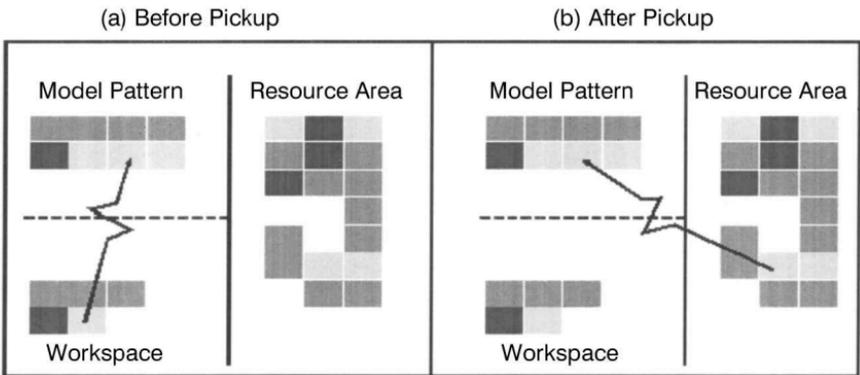


FIG. 3. Schematic of the paradigm. The jagged line indicates that a colour change is made to a model block during a saccade from the workspace to the model before picking up a new block (a) or after pickup (b).

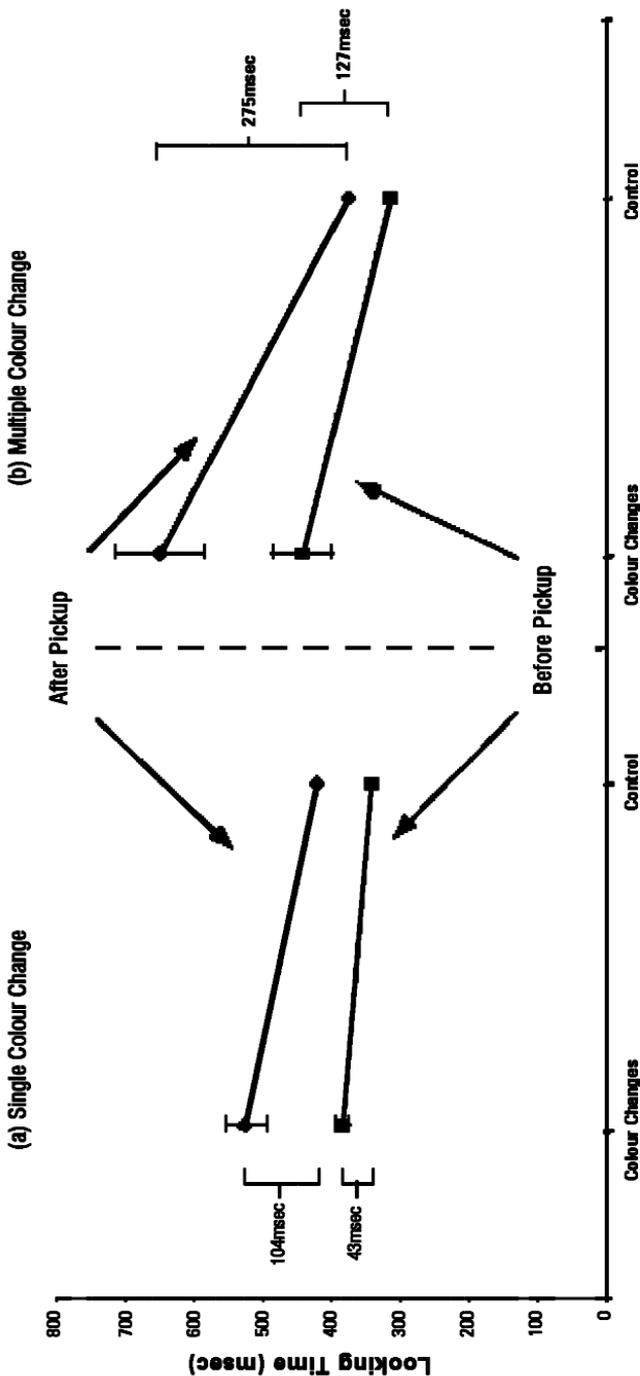


FIG. 4. (a) Duration of the fixations in the model area following a colour change of the fixated block. (When more than one fixation occurred during a visit to the model, the fixation times were summed.) On the left is the fixation time when the eye landed on the changed block. On the right is the control condition where no changes were made. The bottom line shows the model fixation at the beginning of the copying cycle, before the next block is picked up. The upper line shows the model fixation following pickup. (b) As for (a), except that all the uncopied blocks are changed.

different effects depending on point in the task, and the magnitude of the change implies that some representation of the block in hand is maintained between the resource and model fixations. The 43msec increase before the first fixation also suggests some carryover of information from the peripheral activation preceding the model fixation, though it is statistically less reliable. Figure 4(b) shows the same two conditions, but now *all* of the uncopied blocks were changed, instead of just one. In this case fixations following the change were lengthened by 218msec after pickup, and 123msec before pickup. Again, the effect of the change depends on the point in the task. It is also consistently observed that the second fixation on the model, following the pickup, is about 50msec longer in duration than the first fixation. This also suggests that the information acquired in the second fixation is different from that in the first. Acquisition of the block's location in the model appears to take longer than acquisition of colour information—another indication of task specificity. The specialized nature of the computations in a given fixation is also indicated by the results of O'Regan et al. (this issue), and by Wallis and Bulthoff (this issue). Thus, the demands of even simple sensorimotor behaviours require transient, context-dependent computations, and a purely bottom-up approach to vision seems unworkable.

However, there is another important aspect of this data. Changing several blocks, instead of just one, leads to substantially longer fixations. Even when the change precedes the first fixation on the block for colour acquisition, fixations are prolonged by over 100msec—a substantial part of a normal fixation duration.<sup>3</sup> This means that changes in neighbouring, presumably irrelevant blocks affect the representation. Thus, some global property of the model representation is interfered with when the change is made, in both conditions. It isn't clear whether this global property is strictly task relevant or not. One possibility is that the other blocks aren't entirely irrelevant, and a more global representation of the model is necessary for the saccadic process, or for keeping one's place in the task. This is consistent with the suggestion of Chun and Nakayama (this issue) who suggest that implicit memory information acquired in previous views is used to direct attention to task-relevant parts of a scene. Whatever the nature of the representation, however, the sensitivity of fixation duration to the changes differs from the change blindness results in that it was not mirrored in the subjects' verbal reports. Since the goal of the experiment was to probe the copying task, subjects were not informed of the display changes (which occurred on only 10 % of the trials). When questioned at the end of the experiment it appeared that changing a single block was only rarely noticed, out of 250 occasions when a change occurred. This lack of awareness

---

<sup>3</sup>The number of fixations made during a model inspection increased substantially as well (Hayhoe et al. 1998).

frequently coexisted with a new saccade from the changed block to one of the same colour the subject was holding. Even when many blocks were changed, subjects grossly underestimated the frequency of the changes, and the number of blocks changed. (This issue is discussed further in Hayhoe et al., 1998.) This means that fixation duration is a more sensitive indicator than perceptual report in revealing the representations that persist across saccades. The reported detection of such changes may only partially reveal the effect of the display manipulations on visual representations and consequently underestimate what is represented.

The saccade-contingent experiment delivers a mixed message. On the one hand the represented information depends on place in the task, indicating specialized representations. On the other hand it is not entirely consistent with expectations based on the change blindness results, since changes that are not noticed still lead to an effect on fixation duration. This means we cannot immediately conclude that blindness to a change means the absence of a representation. However, one possibility for resolving this inconsistency is to return to the temporal hierarchy described in Table 1. Note that events on the time scale of a second probably correspond to conscious awareness, whereas the visual routines themselves may not be accessible to awareness (Ballard et al., 1998). Perceptual insensitivity to display changes makes sense if conscious experience corresponds to changes in brain state at a time scale of the task, using task relevant variables such as "Next block", "Pickup", and "Putdown". This is illustrated in Table 2. This time scale and these variables describe working memory. The routines that govern the fixations and operations within a fixation presumably run at a shorter time scale, with different primitives. The longer fixations observed in the experiment reflect interference with the functioning of the visual routines. In this experiment, for example (and indeed in most natural performance), observers are not conscious of the eye movements themselves, but primarily of events described in task terms, such as seeing a red block, picking it up and putting it in the correct place in the copy. This is the appropriate time scale for goal-directed behaviour. If awareness corresponded to the act of moving the eyes themselves (as one can do if instructed) this would mean representing the eye movement explicitly as a variable in short-term memory, rather than as an autonomous process. This would compete with events pertinent to the task, since working memory is a capacity limited system.

TABLE 2  
Description of blocks task at different time scales

<i>Time</i>	<i>Process</i>	<i>Description</i>
10sec	High-level control	Copy pattern
1sec	Task-relevant behaviour	Copy next block behaviour
300msec	Visual routines	Visual search "get colour"
80msec	Basic operations	Colour, spatial filters

## HOW BEHAVIOURS ARE COMPOSED, AND THE SCHEDULING PROBLEM

The computational advantage of task specific representations comes at a cost. A crucial problem for any system like this is: How are more complex behaviours composed using the individual routines? That is, how is the scheduling of the individual routines handled? To examine this issue, I will discuss an example in the domain of driving. Figure 5 shows a schematization of an automated vehicle that uses complex images (natural or synthetic) as input, that are then analysed in real time to produce behaviours such as stopping at stop lights and stop signs, and following a constant distance behind another car (Salgian & Ballard, 1998, in press). This schematization has the kind of hierarchical structure envisaged by Newell. Complex behaviour such as driving is assumed to be composed of a set of sub-tasks or simple behaviours. Behaviours are composed of sub-sets of routines. The traffic light behaviour consists of looking for a stop light, stopping if it is red and waiting for a green light, then continuing. The visual routine part of the behaviour would be the detection process itself. In Salgian's model, this involves examining a restricted region of the field and looking for red in that region (Salgian & Ballard, 1998). Stop sign detection involves examining another restricted region (on the right), looking for red blobs, and if one is found, examining the spatial frequency content for a match to the word STOP. Car following (keeping a constant distance behind a lead car) has also been successfully implemented using a looming cue in a restricted region of the velocity space (Salgian & Ballard, in press). The routines are highly context specific, and reduce computational load by taking advantage of this contextual information, for example, by looking for stop signs only on the right. The routines themselves are composed of sub-sets of the basic operations. A biological interpretation of the basic operations would be the neural organization underlying extraction of features such as colour and spatial frequency (see Table 1). In earlier work, Rao and Ballard (1995) have shown that unsegmented image properties encoded by filters such as those in primary visual cortex can be effectively used as a basis for object localization and identification routines.

But how are the different behaviours scheduled? How does the vehicle avoid going by a stop sign when it is in the process of looking for stop lights? One possibility is simply to alternate behaviours at a rate appropriate for the context. This has been effective in Salgian's implementation, but it is probably too simplistic. Recent observations by Land, of eye movements while driving, suggest that at least some of the scheduling may be handled this way (Land & Furneaux, 1997). They recorded drivers' eye movements while simultaneously following a curve and avoiding a cyclist. Land and Lee (1994) have shown that drivers reliably fixate the tangent point of the curve. The driver fixates the tangent point and the cyclist in succession at intervals of about half a second. Land

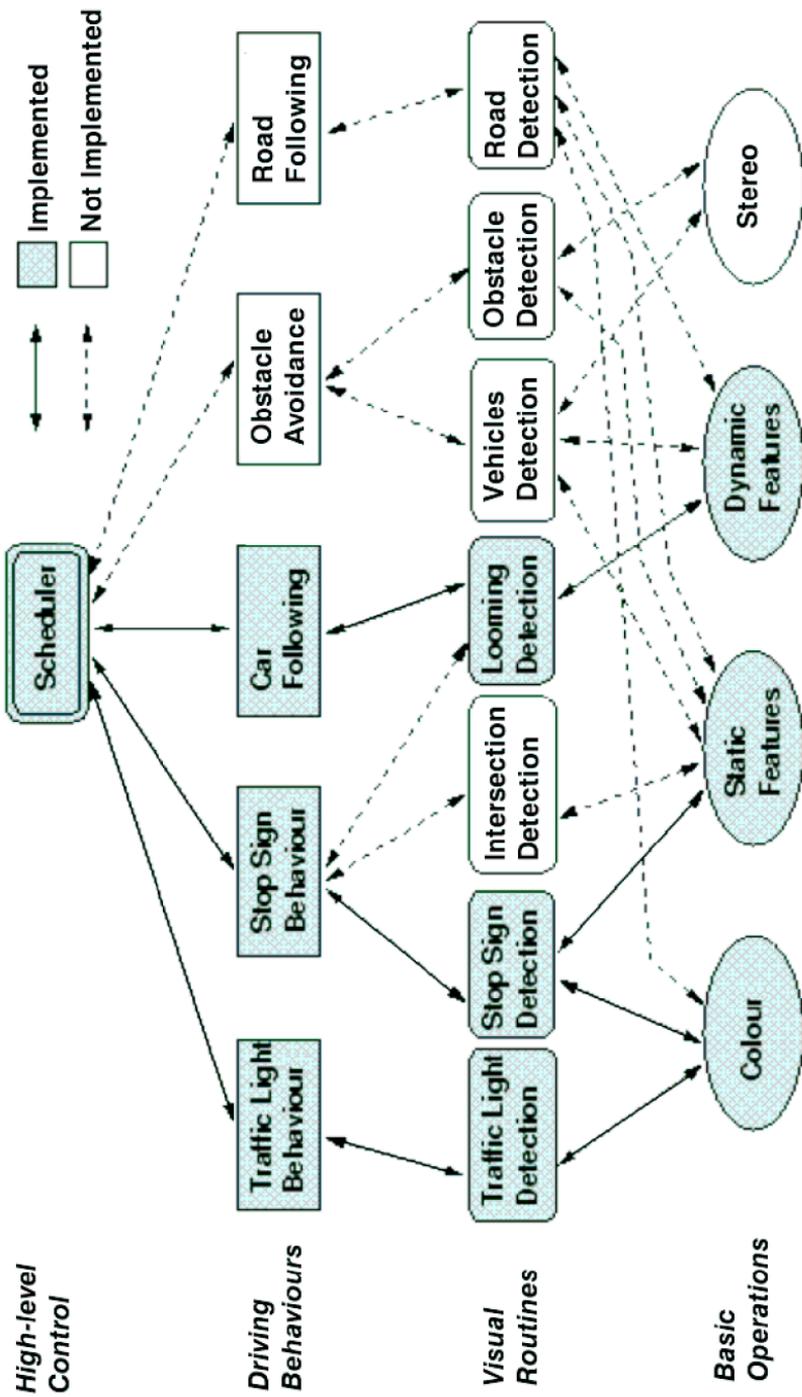


FIG. 5. Representation of Salgian and Ballard's (in press) autonomous vehicle as a collection of specialized visual routines using unsegmented image data. Behaviours use one or more routines and the vehicle alternates between the various behaviours.

(1996) has also shown that current visual information controls steering for the next 800msec, so the alternation rate is close to the time demands of steering. Presumably observers learn an appropriate schedule of behaviours for different contexts or scenarios. The work of McCallum (1995), described earlier, shows that an autonomous agent can learn a minimal set of perception–action sequences appropriate for freeway driving. This can be modelled as a partially observable Markov decision process, where the state transitions reflect the behaviour and the underlying structure of the Markov process reflects the organization of the driving schema. In a similar way, a learnt schema for making a sandwich may be thought of as a Markov process where a state such as “get the bread” would have a high transition probability into the state: “get the peanut butter”. Understanding the way routines are scheduled and composed into behaviour is a critical issue for a functional account of vision.

### HOW LIMITED ARE VISUAL REPRESENTATIONS?

What, then, can we conclude about the content of visual representations from the change blindness work and the results described here? Much of the early work on the nature of the representations that span different fixations was done by Rayner and colleagues in the context of reading, and is summarized by Pollatsek and Rayner (1992). The general consensus in these studies and in later ones, (O’Regan, 1992; Simons & Levin, 1997) is that the visual information retained after a change in fixation position is very limited. Experiments by Irwin revealed strict capacity limits on memory for the patterns across saccadic eye movements (Irwin, 1991; Irwin, Zacks, & Brown, 1990). He suggested that only information that has been the focus of attention is retained across saccades and that this has the usual capacity limits associated with short-term visual memory. The suggestion of O’Regan and Irwin has been that only a sparse “post categorical” description of the objects and locations in a scene is preserved, with new information being actively acquired by gaze changes (Irwin, 1991; O’Regan, 1992; O’Regan & Levy-Schoen, 1983). A more explicit suggestion was made by Ullman (1984), who proposed that the visual routines operated on a “base representation” resulting from early visual transformations such as spatial and chromatic filtering, and that the results of the routines were added to the base representation to give an “incremental representation”. In this case, the ongoing visual goals determine both what is computed within a gaze position and across different gaze positions. That is, the task determines the selection of the appropriate visual routines and how they are composed into behaviour. This is also the implication of the task specificity revealed in the block copying experiments.

However, this seems like an incomplete answer to the question of what is represented. There are a variety of behavioural demands that suggest the need for representations that serve similar functions in most behavioural contexts.

Most critically, the visual system must maintain some level of responsiveness to environmental information that is unexpected or irrelevant to the ongoing task, and the visual system is clearly well designed to do this. Thus, there must be some kind of ongoing examination of the stimulus properties of the unattended visual field. Ullman describes this as “the initial access problem”, since there must be some means by which the current routines can be superseded by new events. To solve this problem, he proposed a set of “universal routines” that could be applied to a wide range of scenes, such as a description of the layout of prominent objects.<sup>4</sup> This is similar to the sparse representation suggested by others, but with the distinction that it is not bottom up, but controllable by the observer. An advantage of this feature is that observers can modulate the application of the routines in a way that would mirror the variations in overall awareness or attentiveness to parts of a scene that are revealed in ordinary behaviour. Another example of a universal routine would be the construction of some kind of representation of the current space and the observer’s position in that space. Such a representation is necessary for any kind of co-ordinated action.<sup>5</sup> (See also Chun & Nakayama, this issue.) For example, in the course of making the peanut butter and jelly sandwich, observers often initiated a movement of the left hand to pick up an object like the jelly jar while the right hand completed another operation such as spreading the peanut butter. Indeed, in about half the reaches, the hand preceded the eye movement, which is almost invariably required for guiding the final pickup. Thus, there must sufficient information in the internal representation to programme the ballistic phase of the hand movement without an initial fixation on the target.

Another important consideration in trying to estimate the extent of visual representations is the disparity between the perceptual reports and the fixation durations in the block copying experiment. Fernandez-Duque and Thornton (this issue) also present evidence that conscious reports underestimate the extent of sensitivity to change. This disparity means that perceptual report underestimates the extent of visual representations. Task relevance does not necessarily lead to conscious awareness. This does not mean that an item is not represented, however, since it is necessary for performing the task. It is possible that the detection of a change is itself a separate task, since it requires a comparison of specific information before and after an eye movement, or other manipulation that bypasses the normal sensitivity to transients, such as backward

---

<sup>4</sup>The selectivity of visual routines is not confined to a particular location. It can be based on properties as well. For example, visual search based on the appearance or features of an object requires analysis of the entire visual field.

<sup>5</sup>Many gaze changes can be initiated by a visual search using information currently presented on the peripheral retina. Such search is based on appearance. However, objects can be coded by spatial location with respect to the current reference frame, and gaze changes are almost certainly programmed on this basis also.

masking (Rensink, O'Reagan, & Clark, 1996). In this context it may be the memory that is limited, rather than the representations during a fixation. Wolfe (1998b) argues that visual representations may be extensive within a fixation, but only some of these last beyond the time that the stimulus is present on that retinal region, and that the change blindness experiments reflect the limitations of visual memory. The normal backward masking of the pre-saccadic stimulus by the post-saccadic display is an argument for something like this. It is also possible that the representation at any moment reflects a lot of information about a scene that has been accrued over many instances of exposure to that scene, and well-learned representations may require only partial information in order to evoke them. Many of the changes made in change blindness experiments are unlikely in a natural environment, and thus may be treated as insufficient evidence for revising the current representation. Current models of cortex suggest that higher cortical areas store abstract representations that are fit to image data coming in from lower areas (Rao & Ballard, 1997),<sup>6</sup> and these models may increase in complexity with perceptual learning. Although one cannot rule out the possibility of extensive representations within a fixation, the task specificity revealed in the blocks paradigm suggests they are limited in some fundamental way, since the representation of even foveally attended blocks depends on the immediate task demands. It is hard to make stronger statements than this in the absence of more extensive data.<sup>7</sup>

## ROUTINES AND ATTENTION

The idea that vision can be thought of as the execution of task-specific routines is closely related to the traditional conceptualization of visual attentive processes. The visual system is usually thought of as a set of transformations of simple sensory properties at early levels of the system, such as the spatial, chromatic, and contrast transformations accomplished by the retina. The extraction of information about certain features such as colour, motion, and texture has also been thought of as "pre-attentive", whereas, at some higher level, perception is thought to require attention. The issue of what can be done pre-attentively has been addressed in a large body of work on visual search, where the distinction between parallel and serial search was thought to correspond to pre-

---

<sup>6</sup> In the Rao and Ballard model, only the residual information that is not fit by stored representations is passed onto higher areas. A low residual corresponds to a match to the stored representation, or recognition. A high residual will result from a failure to match, as would occur when the stimulus was unexpected. This mismatch might initiate a change of task, to deal with the novel stimulus.

<sup>7</sup> Part of the difficulty is that the use of the term "represented" is ambiguous, and is frequently used to refer to computations at different time scales, as described earlier. It is also unclear how the term maps onto the neural events.

attentive vs. post-attentive vision. A recent review by Wolfe (1998a) reveals the difficulty of making this division, however, since high-level properties (segmentation, form, complex conjunctions) can support very rapid, putatively pre-attentive, search, and low-level features such as orientation and colour require slower and putatively serial search. The controlling factor in search time appears, rather, to be one of the difficulty of the discrimination or quality of the signal (Geisler, 1995; Palmer, 1995). This does not break down naturally to a pre-attentive/post-attentive division. The suggestion in this paper is that an explanation in terms of visual routines is more straightforward, since a variety of task-driven routines may require both high- and low-level information. Specification of what is done pre- vs. post-attentively cannot be done on the basis of the stimulus. One compelling example of this is Joseph, Chun, and Nakayama (1997) finding that, in a RSVP task, even orientation discrimination, typically considered pre-attentive, can be shown to require attentional resources when put in competition with another (particularly demanding) task (Joseph et al., 1997). Similarly, even a simple action such as making a saccadic eye movement requires some attention (Kowler, Anderson, Doshier, & Blaser, 1995). Thus it appears that the speed of computation depends on the complexity of the routine, and the strength of the signal,<sup>8</sup> and the particular information extracted by the routine is driven from the ongoing activity.

## SUMMARY

The change blindness phenomena suggest that visual representations may be more limited than previously thought. To take this a step further and understand exactly what constitutes the visual representation of a scene it is necessary to consider the ongoing behavioural demands on the visual system. Ordinary behaviour, and the block copying experiments described here, reveal that visual representations are dynamic and driven by the immediate task demands. This raises the more general issue, that vision should be thought of as an active process whereby the effects of visual stimuli depend on the current state of the system. Specifically, it is argued that vision can be thought of as the ongoing execution of special purpose "routines" that depend critically on the immediate behavioural context. An example of this is given by modelling an autonomous agent in a driving context. In this framework, understanding the nature of the routines and how they are composed into extended behavioural sequences become the central issues in understanding visual representations. It is also argued that it is necessary to distinguish between visual processes that operate at different functional levels and at different time scales. This paper

---

<sup>8</sup>This is similar to the distinction made by Norman and Bobrow (1975) in terms of resource-limited and data-limited processes.

focuses on aspects of vision on the time scale of several hundred milliseconds to a few seconds that come into play for visual processes that span fixations. The conclusion from all this work is that change blindness results from the task specificity of the visual routines but perceptual awareness reflects only events that operate at the time scale of the task. Visual routines themselves are not normally accessible to consciousness, but can be accessed by using other measures such as fixation duration. Consequently, these measures and conscious report may differ.

## REFERENCES

- Bajcsy, R. (1988). Active perception. *Proceedings of the Institute of Electrical and Electronic Engineers*, 76, 996–1005.
- Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Cognitive Neuroscience*, 7, 66–80.
- Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1998). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20, 723–767.
- Ballard, D.H. (1991). Animate vision: An evolutionary step in computational vision. *Journal of the Institute of Electronics, Information, and Communication Engineers*, 74, 343–348.
- Ballard, D.H., Hayhoe, M.M., Li, F., & Whitehead, S.D. (1992). Eye hand coordination in a sequential task. *Proceedings of the Royal Society of London, B.*, 337, 331–339
- Beauchamp, M., Cox, M., & DeYoe, E. (1997). Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *Journal of Neurophysiology*, 78, 516–520.
- Bensinger, D. (1997). *Visual working memory in the context of ongoing natural behaviors*. Unpublished PhD. dissertation, University of Rochester, New York.
- Brooks, R.A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2, 14–22.
- Chun, M.M., & Nakayama, K. (this issue). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition*, 7, 65–81.
- Corbetta, M., Miezin, F., Dobmeyer, S., Schulman, G., & Petersen, S. (1991). Selective and divided attention during visual discriminations of shape, color, and speed: Functional autonomy by positron emission tomography. *Journal of Neuroscience*, 11, 2383–2402.
- Crist, R., Ito, M., Westheimer, G., & Gilbert, C. (1997). Task dependent contextual interactions in the primary visual cortex of primates trained in hyperacuity discrimination. *Society of Neuroscience Abstracts*, 23, 269.
- Fernandez-Duque, D., & Thornton, I.M. (this issue). Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition*, 7, 323–344.
- Geisler, W. (1995). Separation of low level and high level factors in complex tasks: Visual search. *Psychological Review*, 102, 356–378.
- Gilbert, C.D. (1998). Adult cortical dynamics. *Physiological Review*, 78(2), 467–485
- Gottlieb, J., Kusunoki, M., & Goldberg, M.E. (1998). The representation of visual salience in monkey posterior parietal cortex. *Nature*, 391, 481–484.
- Hayhoe, M., Bensinger, D., & Ballard, D. (1998). Task constraints in visual working memory. *Vision Research*, 38, 125–137.

- Irwin, D.E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology*, 23, 420–456.
- 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.
- Joseph, J., Chun, M., & Nakayama, K. (1997). Attentional requirements in a “pre-attentive” feature search task. *Nature*, 387, 805.
- Just, M., & Carpenter, P. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, 8, 441–480.
- Kowler, E., Anderson, E., Doshier, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, 35, 1897–1916.
- Land, M. (1996). The time it takes to process visual information while steering a vehicle. *Investigative Ophthalmology & Visual Science*, 37, S525.
- Land M., & Furneaux S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London, B*, 352, 1231–1239.
- Land, M., & Lee, D. (1994). Where we look when we steer. *Nature* 369, 742–744.
- Land, M., Mennie, N., & Rusted, J. (1998). Eye movements and the role of vision in activities of daily living: Making a cup of tea. *Investigative Ophthalmology and Visual Science*, 39, S457.
- McCallum, A.K. (1995). *Reinforcement learning with selective perception and hidden state*. Unpublished Ph.D. dissertation, University of Rochester, New York.
- Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.
- Norman, D., & Bobrow, D. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44–64.
- O'Regan, J.K. (1992). Solving the “real” mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, 46, 461–488.
- O'Regan, J.K., Deubel, H., Clark, J.J., & Rensink, R.A. (this issue). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7, 191–211.
- O'Regan, J.K., & Levy-Schoen, A. (1983). Integrating visual information from successive fixations: Does trans-saccadic fusion exist? *Vision Research*, 23, 765–769.
- O'Regan, J.K., Rensink, R., & Clark, J.J. (1996). Mud splashes render picture changes invisible. *Investigative Ophthalmology and Visual Science*, 37, S213.
- Pollatsek, A., & Rayner, K. (1992). What is integrated across fixations? In K. Rayner (Ed.) *Eye movements and visual cognition: Scene perception and reading* (pp. 166–191). New York: Springer-Verlag.
- Palmer, J. (1995). Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, 4, 118–123
- Rao, R., & Ballard, D.H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78, 461–505.
- Rao, R., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9, 721–763.
- Rensink, R., O'Regan, J.K., & Clark, J.J. (1996). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8, 368–373.
- Salgian, G., & Ballard, D. H. (1998). Visual routines for autonomous driving. *Proceedings of the 6th International Conference on Computer Vision, Bombay* (pp. 876–882).
- Salgian, G., & Ballard, D.H. (in press). Visual routines for vehicle control. In D. Kreigman, G. Hager, & S. Morse (Eds), *The confluence of vision and control*. Springer-Verlag.
- Simons, D., & Levin, D. (1997). Change blindness. *Trends in Cognitive Science*, 1, 261–267.
- Snyder, L.H., Batista, A.P., & Andersen, R.A. (1997). Coding of intention in the posterior parietal cortex. *Nature*, 386, 167–169.
- Ullman, S. (1984). Visual routines. *Cognition* 18, 97–157.
- Wallis, G., & Bülthoff, H. (this issue). What's scene and not seen: Influences of movement and task upon what we see. *Visual Cognition*, 7, 175–190.

- Wolfe, J.M. (1998a). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–74). Hove, UK: Psychology Press.
- Wolfe, J.M. (1998b). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories*. Cambridge, MA: MIT Press.