CHAPTER 11

Eye Movement Control in Reading: An Overview and Model

Keith Rayner, Erik D. Reichle and Alexander Pollatsek

University of Massachusetts

Abstract

In this chapter, we review experiments dealing with eye movement control in reading, evaluating the factors that control the decisions about (a) where the eyes move and (b) when the eyes move. We briefly review models attempting to account for aspects of these data and then outline a computational model that we have implemented. This model provides a good fit to the eye movement data at the level of predicting which words are fixated and the individual fixation times on words.
Introduction

There has recently been considerable debate about the characteristics of a model of eye movement control in reading. Two general categories of models have been proposed: (a) those that assign lexical processing or other ongoing comprehension processes a major role in influencing eye movements, versus (b) those that maintain that eye movements are mainly controlled by oculomotor factors and are only indirectly related to ongoing language processing. The first category, which we will refer to as *processing* models, includes a model proposed by Morrison (1984), with various modifications (Henderson and Ferreira, 1990; Kennison and Clifton, 1995; Pollatsek and Rayner, 1990; Sereno, 1992), as well as one proposed by Just and Carpenter (1980). The second category, which we will refer to as *oculomotor* models, includes the strategy-tactics model (O’Regan, 1990, 1992), as well as proposals of Kowler and Anton (1987) and McConkie et al. (1989).

In this chapter, we will initially briefly review what we take to be the primary empirical facts that exist about eye movement control in reading. We will then discuss the models proposed by Morrison and by O’Regan. We will argue that a major limitation associated with each model is that they are qualitative, verbal descriptions that lack sufficient power to be precisely tested. We will then briefly describe some recent attempts to provide more quantitative models of eye movement control. The remainder of the chapter will then focus on a quantitative model of eye movement control (Reichle et al., 1998) that we recently developed.

Empirical data

Understanding how eye movements are controlled in reading is important in devising a model of skilled reading (see Rayner and Pollatsek, 1989). Thus, it is not surprising that there has been considerable interest in this topic. Twenty years ago, many studies focused on the extent to which eye movements were controlled on a moment-to-moment basis (Bouma and deVoogd, 1974; Hochberg, 1975; O’Regan, 1979; Pollatsek and Rayner, 1982; Rayner and McConkie, 1976; Rayner and Pollatsek, 1981). The general consensus that emerged was that where readers look next (fixation location) and when they move to a new location (fixation duration) are relatively independent processes, but they are both on-line decisions (Rayner and Pollatsek, 1987).

In this chapter, our primary focus is on the decision regarding when to move the eyes. However, we will first review research dealing with the decision about where to move the eyes. One important feature that distinguishes the processing models from the oculomotor models is the stance they take on the relationship between these two decisions. Processing models view the *where* and *when* decisions as
relatively independent. In particular, they assume that the decision of when to move the eyes is primarily affected by linguistic variables and that fixation durations reflect on-line cognitive processing of language, whereas the decision of where to fixate (especially where within a word to fixate) is largely determined by low-level visual computations. In contrast, oculomotor models posit that lower-level oculomotor or visuomotor factors are the primary determinant of both the when and where decisions. For example, where the eyes are fixated in a particular word is viewed as an important factor affecting both how long the eyes remain fixated and where the eyes will go next. The oculomotor models do not necessarily posit that the where and when decisions are made at the same time; usually, the when decision is thought to depend on the outcome of the where decision, and it is in this sense that they are highly dependent. We will discuss these two classes of models in greater detail after presenting the research findings concerning the where and when decisions.

Where to fixate next

As the above characterization of the models indicates, there seems to be relatively widespread agreement that low-level visual information obtained on the prior fixation (much of it in parafoveal vision) is an important factor determining where to fixate next during reading. Specifically, most theories view word boundaries, defined by the spaces surrounding the fixated word and the next word in the text, as important visual information guiding eye movements. One piece of evidence for this is that when this low level visual information is not available (e.g., when spaces between words are removed), readers move their eyes a shorter distance than when such information is available (McConkie and Rayner, 1975; Morris, Rayner and Pollatsek, 1990; Pollatsek and Rayner, 1982; Rayner and Bertera, 1979; Rayner, Fischer and Pollatsek, 1998; Rayner and Pollatsek, 1981).

Epelboim, Booth and Steinman (1994), however, recently challenged the assertion that we just made — that space information is important in guiding the eyes — and argued that the primary reason that unspaced text interferes with reading is that removal of spaces interferes with word identification. Their argument is largely based on the assumption that if readers normally relied on space information to guide the eyes during reading, they should be virtually helpless when spaces are removed, and the pattern of saccades should be altered drastically. Instead, they claim that the pattern of saccades is only minimally altered and that reading unspaced text is surprisingly easy. (However, removal of spaces does cut reading speed in half for most readers.) We agree with part of their claim — that word identification is interfered with when spaces are absent (Pollatsek and Rayner, 1982; Rayner and Pollatsek, 1996) and that one has to be careful about inferring that removing spaces is only interfering with eye guidance. However, recent experiments from our lab (Rayner et al., 1998) and data we review below make it clear that where
the eyes move is guided by space information. For example, the length of the parafoveal word has been shown to strongly influence where the eyes initially fixate on a word (the initial landing position) and hence the length of the saccade into that word (Blanchard, Pollatsek and Rayner, 1989; O'Regan, 1979, 1980, 1981; Rayner, 1979).

More generally, oculomotor theorists have viewed the initial landing position on a word as perhaps the most important variable mediating reading. In their view, (a) a major oculomotor goal of the reader is to land on the middle of each word, and (b) the resultant oculomotor behavior on that word is chiefly determined by the success of the initial eye movement in achieving the goal of landing in the middle. The data on initial landing positions are reasonably consistent with the first part of the claim. Initial landing positions are quite systematic: readers tend to fixate about halfway between the beginning and middle of words (Dunn-Rankin, 1978; McConkie et al., 1988; O'Regan, 1981; Radach and Kempe, 1993; Rayner, 1979; Rayner and Fischer, 1996; Vitu et al., 1995; Vutu, O'Regan and Mittau, 1990). Rayner (1979) called this prototypical location the preferred viewing location. It should be quickly pointed out, however, that the preferred viewing location is a mean, and there is considerable variability in the initial landing position and that the histograms of initial landing positions tend to look like truncated Gaussian distributions; moreover, readers fixate on the spaces between words about 10% of the time.

O'Regan and Lévy-Schoen (1987) subsequently distinguished between the preferred viewing location and what is now called the optimal viewing position, which they posited was the center of the word. Presumably, the optimal viewing position is optimal because of visual acuity considerations (i.e., letters are harder to identify, the further they are from fixation). Hence, fixating in the middle of a word may be "best" because it minimizes the maximum distance a letter in the word would be from fixation. McConkie et al. (1988) presented a detailed analysis of a large corpus of data (see also Radach and Kempe, 1993; Rayner, Sereno and Raney, 1996) that indicates that the discrepancy between the preferred and optimal viewing positions may be largely explained by the distance between the center of a target word and the launch site (i.e., the location of the prior fixation). That is, the further the launch site is from the center of a word, the further to the left the mean landing position is. In addition, the further the launch site is from the center of the target word, the more variability there is in the landing position. This pattern is consistent with the hypothesis that the intended target is the center of the word, as much of the motor programming literature indicates that actual motor movements tend to undershoot the target, with both the amount of undershoot and random variability increasing, the larger the movement.

The data reviewed so far are consistent with the hypothesis that the center of a word is the target for the initial fixation, but they don't compel the conclusion that the middle of a word is, in fact, the optimal viewing position. Extensive research
efforts have examined the consequences of making fixations at locations other than this optimal viewing position (McConkie et al., 1989; Nazir, 1993; O'Regan et al., 1984; Vitu, 1991; Vitu et al., 1990) and for words in isolation, two general effects have been found, which have been referred to (see Rayner et al., 1996) as refixation and processing-cost effects. The refixation effect is that the further the eyes are from the optimal viewing position, the more likely it is that a refixation will be made on the word. The processing-cost effect is that there is a cost in fixation time on the word associated with initial fixation locations other than the optimal viewing position. Presumably, both effects occur because the less optimal the initial viewing position, the less information from a word can be extracted from the initial fixation, and hence (a) the greater the need to fixate it again, and (b) the more time needed to process it. (We will defer discussion of the processing cost effect until the next section, because it is largely relevant to the when rather than the where decision.)

The reason that these effects have been chiefly studied in isolation is that the initial landing position can be experimentally manipulated (by calculating where a subject’s fixation is and then presenting the position of a target word contingent on that fixation point). These experiments thus clearly establish that differences in initial landing position are causing these effects in these circumstances. What is less clear, however, is the importance of these findings for reading text. Both effects become quite attenuated when words in text have been examined, although the refixation effect still appears to be significant (Vitu et al., 1990). We will defer for the moment discussing why these effects attenuate in silent reading as the refixation effect still is significant in reading and indicates a second way in which low-level visual information guides the eyes. That is, from the evidence discussed earlier, it appears that most interword saccades are calculated to be directed to the middle of a word. The refixation effect suggests that refixations on a word may also be targeted to some place on a word without direction from cognitive processes. O'Regan (1992) posits that the target is the end of the word farthest from the current fixation, and there is some evidence that many refixations tend to be near the ends of words (Hyönä and Pollatsek, 1998; Pynte, 1996; Rayner et al., 1996). However, many refixations also tend to be directed to the middle of a word as well. The refixation data raise a problem for a strong form of the oculomotor view. That is, what is a sensible strategy for refixating? If one has obtained information from the initial part of a word, then it makes sense to refixate near the end. However, if one has not obtained any useful information on the first fixation, then it makes sense to fixate in the middle. But if either strategy can be adopted, how does the reader decide where to refixate? Presumably, this would be on the basis of some cognitive operation.

To this point, the data seem to suggest that where a reader’s eyes land is completely guided by calculations of where words are and how long they are (presumably indicated by the locations of spaces) and the content of words appears to be irrelevant. In fact, Vitu et al. (1995) essentially attempted to demonstrate this
by examining eye movements when readers scanned nonsense "text" (in which every letter in a coherent text was replaced with a "z" but where the spaces between words were preserved). They found that global aspects of the eye movement record (such as frequency distributions of fixation durations and saccade lengths) were quite similar to those when the same readers read text. They argued, on the basis of these data, that the similarity of the patterns suggested that predetermined oculomotor strategies were an important factor in eye movement control in reading. The problem with these analyses, however, is that they are too gross, and do not attempt to examine whether linguistic variables play a part. In a follow-up experiment, Rayner and Fischer (1996) replicated their findings, but showed that there were important differences between the two situations (reading text and scanning z-strings).

There are several consistent findings about where the eyes land in real text that would not be observed in z-string text. One is that words that are low frequency in the language tend to be refixated more than high frequency words, even when word length is controlled (Rayner and Fischer, 1996; Rayner et al., 1996). Moreover, this finding holds even when initial landing position is controlled (Rayner et al., 1998). Thus, it appears that the decision of whether to refixate a word or to move on is clearly controlled by cognitive variables. Another is that the frequency with which a word is skipped is dependent on both its frequency in the language and on how predictable it is from the prior text (Balota, Pollatsek and Rayner, 1985; Ehrlich and Rayner, 1981; Rayner and Well, 1996). Thus, it appears that the decision of whether to fixate a word is also dependent on cognitive variables in addition to low level variables.

To summarize, the data on where the eyes land seem to argue for the following factors. First, cognitive variables appear to be influencing gross decisions, such as which word to fixate (e.g., "Do I refixate this word?", "Do I skip this word?"), but the details of where to fixate on words appear to largely be left to lower level processors. The initial fixation location appears to be primarily determined by a low-level oculomotor strategy — fixating the center of a word (though there is bias and random error in carrying out the motor program). In addition, there is some evidence that the targeting of refixations may also be made on a similar basis. However, we need to stress that all the evidence indicates that there is considerable inaccuracy in programming saccadic eye movements. As a result, it is hard to pinpoint the causes of any particular eye movement. For example, there are undoubtedly words skipped because the eyes overshot their target; this would most commonly occur for short words (Blanchard et al., 1989; Rayner, 1979; Vitu et al., 1995). Likewise, there are undoubtedly words refixated because the eyes undershot their target; this would most commonly occur for long words (Hyönä and Pollatsek, 1998; Rayner and Morris, 1992). (For a review and an alternative account of word skipping, see Chapter 6).

A final issue for which the data are less clear is whether cognitive variables influence where on a word the eyes land. One question is whether semantic
variables influence the initial landing position on a word. Several studies have attempted to get at this issue by using long words (10 or more letters) and varying whether the beginning or the end of a word is “informative” (i.e., the word part is low frequency in the language) or “redundant” (i.e., the word part is high frequency in the language). Some of these studies have suggested that the eyes move farther into a word when the informative portion is located at the end rather than the beginning of the word (Everatt and Underwood, 1992; Hyönä, Niemi and Underwood, 1989; Underwood, Bloomfield and Clews, 1988; Underwood, Clews and Everatt, 1990). However, neither Rayner and Morris (1992) nor Hyönä (1995) replicated the effect. Moreover, such effects, even if replicable, do not necessarily reflect semantic preprocessing: they could be due to either orthographic or lexical factors. Indeed, there is a suggestion that some type of orthographic processing of the initial letters of words may influence where readers initially fixate (Beavillain, Doré and Baudoiun, 1996; Hyönä, 1995). A second question is whether refixation locations are influenced by cognitive variables. Hyönä and Pollatsek (1998) demonstrated that the length of the initial morpheme in long compound words influenced the location of the second fixation on a word even when the total length of the word was controlled. They also demonstrated that the frequency of the initial morpheme influenced the pattern of refixations on the word. Thus, it appears that even decisions about where to fixate on a word are influenced by cognitive variables. Whether these influences are limited to very long words in languages that have productive systems for compounding (e.g., Finnish) is still not clear, however.

When to move the eyes

As indicated earlier, the question of what determines when we move our eyes is central to the distinction between models of eye movement control in reading. Because part of the motivation for the oculomotor models is that cognitive operations are relatively slow and thus unlikely to play a major role in deciding when the eyes move, we will first review some data arguing that this is not necessarily so.

One paradigm that has been used to address the issue of how long it takes to encode the visual information involves presenting a visual mask at different intervals during reading (Ishida and Ikeda, 1989; Rayner et al., 1981; Slowiaczek and Rayner, 1987). These studies demonstrated that if text is exposed for longer than 50–70 ms on each fixation before being masked, reading proceeds quite normally. If the mask occurs earlier, however, reading is disrupted. This finding of course does not imply that words are identified within the first 50–70 ms of a fixation; it merely means that sufficient visual information can be extracted within the first 50–70 ms of a fixation so that cognitive operations can proceed normally even when all visual information is removed after that. However, when the visual information during a fixation is changed after the first 50–70 ms, there is disruption (Blanchard et al., 1984).
The above evidence indicates that the completion of early stages of visual processing is quite rapid. There is another reason to expect that full word identification may be completed quite rapidly after a word is initially fixated: word identification often begins before the word is fixated. We have termed this phenomenon *preview benefit*, and it occurs both in situations where words are presented in isolation (Rayner, McConkie and Ehrlich, 1978; Rayner, McConkie and Zola, 1980) and in reading (Balota et al., 1985; Rayner et al., 1982). In isolated situations, this has usually been assessed by examining the time to name a word, and the findings are that word naming time can be speeded by up to 20–60 ms by seeing a preview of a word before it is fixated. Moreover, preview of a word that is either orthographically similar or phonologically identical to the target word also produces preview benefit (Pollatsek et al., 1992; Rayner et al., 1978); this indicates that preview benefit is not merely a result of full processing of the parafoveal word prior to fixating it and is instead largely some sort of integration of the processing of the word in the parafovea with processing of the word when it is later fixated. (We will come back to preview benefit in reading after we discuss the data on when decisions.)

Above, we argued that it is not implausible that cognitive operations are fast enough to guide the decision of when to move the eyes. Perhaps more importantly, a considerable body of data has accumulated over the past twenty years which demonstrates that various lexical, syntactic, and discourse factors influence fixation times on words (for recent reviews, see Rayner, 1995; Rayner and Sereno, 1994). There is evidence that the fixation time on a word is affected on-line by the following psycholinguistic variables: word frequency, lexical ambiguity, semantic relationship, contextual constraint, syntactic complexity, anaphora, and coreference. The most common measure involved in establishing the above relationship is *gaze duration*, which is the sum of the fixation durations on a word before a saccade is made off of the word. This is clearly a composite measure, which is a function both of the durations of individual fixations and the number of fixations on a word. However, a second measure *first fixation duration*, which is the mean fixation of the initial fixation on a word, is also influenced by virtually all of the above variables. (Both measures are conditional on the word being fixated.)

Thus, it appears that linguistic variables are involved in the decision about when to move the eyes. However, others (see O'Regan, 1990, 1992) have argued that lower level oculomotor factors are primarily responsible for controlling eye movements and hence fixation durations. As we have indicated earlier, one of the pieces of evidence for this assertion is that the probability of making a refixation is influenced by the initial landing position. As the number of refixations influences the mean gaze duration on a word, this suggests that oculomotor factors also influence fixation times on a word. More directly, Vitu et al. (1990) found that there was about a 20 ms cost in the gaze duration on a word for each letter that the initial
landing position was from the optimal viewing point. However, this estimate came from examining words in isolation, and when words were examined in silent reading, there was no significant difference in gaze duration as a function of initial landing position (although there did appear to be a small residual cost for initial landing position being distant from fixation).

To summarize, it has been shown that cognitive variables indexing psycho-linguistic processing of the text have definite effects both on the time spent processing a word and the duration of individual fixations. This indicates that at least some cognitive operations are fast enough to influence the decision of when to move the eye. Moreover, a parafoveal preview of a target word reduces fixation times (both gaze durations and first fixation durations) on the word; thus one factor helping to speed these cognitive operations relative to the beginning of the initial fixation on a word is that processing for many words starts before they are fixated. There are some data, however, that indicate that lower level factors, such as the initial landing position, influence the duration of fixations. We will discuss many of these issues further in the next section.

Models of eye movement control

**A process model: Morrison**

In Morrison’s (1984) model, at the beginning of each fixation, eye location and covert visual attention are oriented to the same location: the foveal word (word \( n \)). After foveal processing has reached some criterial level (such as some stage of lexical access), attention shifts to word \( n + 1 \) during the fixation. This shift of attention allows processing of word \( n + 1 \) to begin and also signals the eye movement system to prepare a motor program for an eye movement to word \( n + 1 \). Once the motor program is completed, it is executed and the eyes then make a saccade to that word. Because there is a lag between the attention shift and saccade execution due to programming latency, information continues to accumulate from word \( n + 1 \) before it is directly fixated. If word \( n + 1 \) is identified quickly, attention shifts again to word \( n + 2 \) before the eye movement is fully programmed. In this case, the eyes saccade to word \( n + 2 \), skipping word \( n + 1 \). Because a later program has cancelled an earlier one, the duration of the fixation prior to a skip is inflated compared to when the next word is not skipped (see Hogaboam, 1983; Pollatsek, Rayner and Balota, 1986). If the motor program to word \( n + 1 \) is more advanced, however, there will be either (1) a short fixation on word \( n + 1 \) followed by a longer fixation of word \( n + 2 \) or (2) a fixation located at an intermediate position between words \( n + 1 \) and \( n + 2 \). The model can thus explain two rather puzzling aspects of eye movement behavior in reading: (1) the fact that there are fixations that are much
shorter than the 175–200 ms saccade latency in simple oculomotor tasks (Rayner et al., 1983) and (2) unusual landing positions (e.g., the space between words).

One problem with Morrison’s model is that there is no explanation for why words are sometimes refixated. That is, if lexical access is the trigger for attention shifts and hence eye movements, words should never be refixated. Some recent modifications of the model (e.g., Henderson and Ferreira, 1990; Sereno, 1992) incorporate a deadline for programming a saccade; if lexical processing has not reached a criterion level by the deadline, attention does not shift from the current word and it may be refixated. (We will indicate another possible explanation for refixations below.) Two other problems are: (a) that it cannot explain why there are regressions, and (b) it does not explain how “higher-order” psycholinguistic processes, such as anaphora, can influence eye movements. It also doesn’t attempt to explain where people fixate on words.

An oculomotor model: O’Regan

According to O’Regan’s (1990, 1992; O’Regan and Lévy-Schoen, 1987) strategy-tactics model, the initial landing position in a word chiefly determines how long the reader will remain fixated and where the next fixation is made. O’Regan proposed that readers adopt a global strategy (e.g., careful or risky reading) that coarsely influences fixation time and saccade length. He also proposed that readers implement local, within-word tactics that are based on lower-level nonlexical information available early in a fixation. It is the operation and control of these within-word tactics that are most relevant to the issue of when to move the eyes. If the initial landing position is optimal (near the word’s middle), there will be a single fixation. However, if the initial landing position is in a non-optimal position, a refixation will generally occur and the fixation time for this refixation time is short and unaffected by any linguistic variable. Moreover, the probability that a word will be refixated does not depend on its lexical status, but on lower level visual factors such as the landing position in the word. Linguistic factors are thought to be slow and thus only can influence (a) the duration of single long fixations (presumably 300 ms or longer, see O’Regan, 1992), or (b) the second of two fixations in a refixated word. Thus, fixation times according to this scheme are mainly determined by oculomotor constraints.

As our discussion of the eye movement data indicates, there are some problems with the model. Most specifically, the claim about linguistic effects being limited to long single fixations or second fixations on a word is wrong: (a) the first (of two or more) fixations on low frequency words are longer than on high frequency words (Rayner et al., 1989, 1996); and (b) frequency effects on fixation durations do not only show up in the upper tails of the distributions. However, the model does point to certain ways in which lower-level oculomotor variables influence eye movements. As
we indicated earlier, however, the fact that most of these effects are substantially weaker in reading than when people examine individual words suggests that they are less important in reading than the model would suggest. We think there are several possible reasons for why these oculomotor variables are less important in reading than in more controlled studies. First, we think that attentional strategies may be very different in the two situations. In the more controlled studies, people are asked to make (and maintain) a fixation to a particular location prior to presentation of the target word. This could produce a narrowing of the attentional focus around fixation that is quite atypical of reading. Second, there is no parafoveal preview information in such studies. This would slow down lexical processing and make oculomotor factors more likely to predominate. More generally, if the precise location of eye fixations (e.g., the initial landing position on a word) were as important for reading as posited by these models, it would seem unlikely that they would be as variable as they are.

Which model is right?

Although we are clearly in greater sympathy with the process model approach, it should be obvious that neither type of model has a monopoly on truth. First, neither type of model is adequate to explain all of the data. Morrison's model, for example, simply does not attempt to explain where readers land on words. On the other hand, as we have argued above, O'Regan's models give an extremely unsatisfactory account of how linguistic variables affect reading. At present, neither model gives an entirely satisfactory account of the details of fixations on words. For example, Vitu and O'Regan (1995) presented data which they argued are inconsistent with the modification of Morrison's model proposed by Henderson and Ferreira (1990), which claims that refixations on words occur because a deadline is reached and lexical access has not occurred yet. Such a model predicts that the first of multiple fixations should be longer than single fixations because the latter should be the result of eye movements programmed prior to the deadline. Instead, Vitu and O'Regan found that the duration of a single fixation is longer than the first of two fixations (see also Rayner et al., 1996). On the other hand, there are word frequency effects that are independent of where the reader initially fixated on the word (Rayner et al., 1996, 1998), which is inconsistent with predictions of the oculomotor model. In addition, both models ignore syntactic effects (e.g., "garden path" effects) or discourse-level effects (e.g., resolution of anaphora) on eye movements in reading. More critically, we would like to argue that neither type of model is sufficiently precise for comprehensive testing. In the remainder of this chapter, we will focus on recent endeavors in our laboratory to provide a more formal model of eye movement control. Prior to doing so, we will briefly mention other such attempts.
Attempts to produce more quantitative models

In addition to the qualitative models described above, there have been some recent attempts to produce more quantitative models of the characteristics of eye movements during reading. For example, models by Legge, Klitz and Tan (1997) and Suppes (1990) have focused on low level aspects of reading and have not concerned themselves with the duration of fixations. The two models, however, are quite different. Legge et al.'s model, called Mr. Chips, attempts to explain the details of where readers fixate cognitively, assuming an intelligent guiding mechanism that is controlled by lexical access and the details of which letters can be processed due to acuity limitations and other considerations. In contrast, Suppes assumes rather "dumb" mechanisms and focuses on the stochastic properties of the variability of saccadic eye movements in reading. One other model that deserves mention is that of Thibadeau, Just and Carpenter (1982), which is a formal production system that provides a more quantitative account of the Just and Carpenter (1980) model. Like its predecessor, it focuses on a composite gaze duration measure and ignores important details like the probability of refixations and of word skipping.

These models are all reasonable attempts to explain part of the eye movement record. However, our model goes a significant step beyond them by trying to account simultaneously for the details of individual fixation durations and the location of individual fixations (at the level of which word is fixated). To the best of our knowledge, there are no extant models which successfully account for eye movements in reading at both of these levels. However, we should make clear that our model has two clear deficiencies. First, it does not attempt to explain where on a word a reader fixates. This is because, at present, it is completely a process model; however, we think that adding oculomotor stages to handle the where question in more detail is not a conceptually difficult next step, although it would complicate the formal modeling. Second, we do not attempt to explain how syntactic and discourse level processes influence eye movements. We have two major reasons for this. First, we think that this would be too difficult given that there are no theories of syntactic or discourse-level processing that are even close to providing a precise enough theory to incorporate into a model of eye movements. Second, we think it is not an unreasonable hypothesis that these variables may intervene to control eye movements primarily when the reader in fact runs into trouble (such as in “garden path” sentences) and needs to suspend normal processing until the problem is repaired (usually involving regressions and/or very long fixations). Thus, we think that our model, which assumes that word identification is the “engine” that drives the eyes forward may explain a good deal of the eye movement record in reading.
The E-Z Reader Model

In the remainder of this chapter, we provide an overview of the E-Z Reader model. More detail on the underlying mathematical equations and formal modelling is provided in Reichle et al. (1998). E-Z Reader\(^1\) is similar to Morrison's (1984) model with two notable exceptions. First, E-Z Reader decouples the signal to shift attention from the signal to program a saccade\(^2\). Second, E-Z Reader is better specified in that it has been implemented as a computer simulation program. The basic goal of E-Z Reader is to predict (a) the probability a word is fixated, (b) the number of times it is fixated, and (c) the duration of individual fixations on it.

There are five basic processes in E-Z Reader: (1) a familiarity check on a word; (2) completion of lexical access; (3) an early, labile stage of saccadic programming which can be canceled by subsequent saccadic programming; (4) a later, nonlabile, stage of saccadic programming; and (5) the actual saccadic eye movement. Figure 1 is a schematic diagram showing how these process interact to control eye movements during reading.

The first pair of processes correspond to different stages of lexical access, and thus can be conceptualized as products of a single cognitive module (Fodor, 1983) that is responsible for word identification. The first process, the familiarity check on a word, signals a point in time when lexical access is imminent, and thereby cues the system mediating eye movements to begin planning a saccade. The second process, the completion of lexical access, corresponds to a stage where a word's identity has been determined and this information can be passed on to other systems. As in Morrison's (1984) model, the completion of lexical access causes attention to shift to the next word. However, as already mentioned, our model differs from Morrison's in that E-Z Reader decouples covert shifts of attention from saccadic programming; whereas the familiarity check initiates the programming of a saccade, the completion of lexical access causes attention to shift to the next word.

Although the division of lexical access into two discrete stages is partly a modeling convenience, it is not without precedent. For instance, the distinction

---

\(^1\) Our previous modelling efforts (see Reichle et al., 1998) were directed towards finding the minimal set of assumptions that could account for eye movement control in reading. As a result, our final model, E-Z Reader 5, was preceded by several approximations. In this chapter, we will ignore the earlier versions of the model; “E-Z Reader” therefore refers to E-Z Reader 5.

\(^2\) We realize that the decision to decouple covert shifts of spatial attention from eye movements is somewhat controversial. However, most of the studies that have shown tight coupling between the two (specifically that covert attention can not be deployed to places other than where a saccade is programmed to) use paradigms in which exogenous signals are driving the eyes. Moreover, a recent study by Stelmach, Campsall and Herdman (1997) has found independence between attention shifting and saccades even in these paradigms.
between (a) a rapid feeling of familiarity and (b) a slower retrieval of information from memory is consistent with previous empirical (Atkinson and Juola, 1973) and theoretical work (Hintzman, 1988; Gillund and Shiffrin, 1984). In addition, there have been several two-stage "verification" models of lexical access (e.g., Paap et al., 1982; Van Orden, 1987) that would be reasonably consistent with our model.

In E-Z Reader, the time necessary to complete the familiarity check (in the absence of top-down influences) is assumed to be a linear function of the logarithm of word frequency (as tabulated by Francis and Kucera, 1982)\(^3\). The additional time necessary for the completion of lexical access is assumed to be a constant multiple of the familiarity check time. Both process durations are also assumed to be affected by top down influences, captured in a predictability value obtained from off-line data obtained from subjects. The time required for both the familiarity check and completion of lexical access are assumed to be reduced (multiplicatively)

---

\(^3\) In our modeling, to minimize the number of parameters, we did not distinguish between frequency and word length effects. Thus "frequency effects" in our model are really a combination of frequency and word length effects because the two are highly correlated in our sample of text as in printed English in general. Similarly, the assumption that the duration of the familiarity check stage is a constant fraction of the duration of lexical access (in the absence of top-down influences) is to minimize free parameters. For a complete explanation of the mathematical equations and details of the formal modeling, see Reichle et al. (1998).
Eye movement control

by predictability. Finally, two additional parameters control the familiarity check and completion of lexical access process durations so that the rate of processing becomes slower as eccentricity (i.e., the distance between the word being processed and fixation) increases. The eccentricity assumption was added to increase the psychological plausibility of the model because (a) retinal acuity rapidly decreases as distance from the fovea increases, and (b) words are identified more rapidly in the fovea than the parafovea (Rayner and Morrison, 1981).

The three remaining processes (i.e., the labile and nonlabile stages of saccadic programming and the actual saccades) can be viewed as components of a single module that programs and executes eye movements. The division between labile and nonlabile stages of saccadic programming was motivated by Morrison's model, which in turn was motivated by Becker and Jürgens (1979). Their data indicated that the computations that are necessary to move the eyes to a target location can be canceled or modified if a new target location is presented early during programming, but cannot be interrupted if a new location is presented late in programming.

As already mentioned, the programming of an interword saccade is initiated in E-Z Reader when the familiarity check on the previous word has been completed. And, consistent with Becker and Jürgens (1979), the early, labile stage of programming is canceled whenever the labile program for another saccade is initiated (as in Morrison's model). This process is illustrated schematically in Fig. 2. In the first example, the labile program to move the eyes to word \( n + 1 \) completes before the program to move the eyes to word \( n + 2 \) is initiated, so that both saccades are executed. (This is the mechanism that could produce very short fixation durations.) In contrast, in the second example, the labile program to word \( n + 1 \) is canceled when the program to word \( n + 2 \) is initiated; consequently, only the saccade to word \( n + 2 \) is executed, and word \( n + 1 \) is skipped. The latter sequence of events (i.e., the cancellation of one labile program by another) allows E-Z Reader to generate skips.

In the case of intraword saccades (or refixations), the signals that initiate and terminate the labile stage of programming are somewhat different: The labile program is initiated at the moment that a word has been fixated and it will produce a refixation unless it is canceled by the completion of the word's familiarity check (i.e., the event that signals the eye movement system to plan an interword saccade). This assumption is consistent with the following "dumb" default strategy: Fixate each word from more than one viewing location unless the word's familiarity indicates that a refixation is unnecessary. (This assumption is not very different from that of O'Regan.) Finally, as Fig. 2 indicates, the nonlabile stage of saccadic

---

4 In the model, the impact of predictability on the familiarity check and lexical completion stages was somewhat different. This was handled with a single multiplicative parameter that attenuated the predictability effect for the familiarity check stage. The rationale for this was that top-down processes should affect early stages of lexical access less than later stages.
Example 1:

labile program to word_{n+1}  nonlabile program to word_{n+1}  execute saccade to word_{n+1}

labile program to word_{n+2}  nonlabile program to word_{n+2}  execute saccade to word_{n+2}

Example 2:

labile program to word_{n+1}  abort program and cancel saccade to word_{n+1}

labile program to word_{n+2}  nonlabile program to word_{n+2}  execute saccade to word_{n+2}

Fig. 2. Schematic diagram showing the relationship between labile and nonlabile stages of saccadic programming.

programming immediately follows the labile stage, and a saccade is executed once the nonlabile stage has completed. For simplicity, the durations of the labile and nonlabile stages of saccadic programming and for the actual saccadic movements were set equal to fixed values.

E-Z Reader was applied to a corpus of data collected by Schilling, Rayner and Chumbley (1998). In their study, subjects read 48 sentences while their eye movements were monitored. Each sentence was 8–14 words in length. Eight values were tabulated for each word in the corpus: (1) the natural frequency of occurrence (from Francis and Kucera, 1982); (2) the predictability (obtained from a rating study); (3) the mean gaze duration; (4) the mean first fixation duration; (5) the mean single fixation duration (i.e., the mean duration of fixations when there was exactly one fixation on a word); (6) the proportion of times a word was skipped; (7) the proportion of times a word was fixated once; and (8) the proportion of times a word was fixated twice. The mean fixation durations and proportion of times skipped, fixated once, and fixated twice were then averaged within five frequency classes of words to produce 30 means. E-Z Reader was fitted to these data by conducting

---

5 The first and last words in each sentence were excluded from the analyses because (a) the first word was initially fixated by a reading-irrelevant movement from a fixation cross, and (b) the fixation on the last word coincided with a button push. Sentences with interword regressions were also excluded since they generally reflect difficulty with higher-order processing (Ehrlich and Rayner, 1983; Frazier and Rayner, 1982) and are, therefore, beyond the scope of E-Z Reader. However, intraword regressions were included in the data.
Eye movement control  

Table 1

Observed and predicted values of gaze durations, and individual fixations, probability of skipping, making a single fixation, and making two fixations for five frequency classes of words

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–10</td>
<td>3</td>
<td>293 291</td>
<td>248 251</td>
<td>265 274</td>
</tr>
<tr>
<td>2</td>
<td>11–100</td>
<td>45</td>
<td>272 271</td>
<td>234 253</td>
<td>249 263</td>
</tr>
<tr>
<td>3</td>
<td>101–1,000</td>
<td>347</td>
<td>256 257</td>
<td>228 246</td>
<td>243 252</td>
</tr>
<tr>
<td>4</td>
<td>1,001–10,000</td>
<td>4,889</td>
<td>234 226</td>
<td>223 223</td>
<td>235 224</td>
</tr>
<tr>
<td>5</td>
<td>10,001+</td>
<td>40,700</td>
<td>214 211</td>
<td>208 210</td>
<td>216 210</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–10</td>
<td>3</td>
<td>0.10 0.09</td>
<td>0.68 0.73</td>
<td>0.20 0.17</td>
</tr>
<tr>
<td>2</td>
<td>11–100</td>
<td>45</td>
<td>0.13 0.16</td>
<td>0.70 0.76</td>
<td>0.16 0.07</td>
</tr>
<tr>
<td>3</td>
<td>101–1,000</td>
<td>347</td>
<td>0.22 0.27</td>
<td>0.68 0.68</td>
<td>0.10 0.04</td>
</tr>
<tr>
<td>4</td>
<td>1,001–10,000</td>
<td>4,889</td>
<td>0.55 0.49</td>
<td>0.44 0.50</td>
<td>0.02 0.01</td>
</tr>
<tr>
<td>5</td>
<td>10,001+</td>
<td>40,700</td>
<td>0.67 0.68</td>
<td>0.32 0.32</td>
<td>0.01 0.00</td>
</tr>
</tbody>
</table>

multiple grid-searches of the parameter space to determine the best values for the model parameters. The data and simulation results using the best-fitting parameter values are presented in Table 1.

As can be seen in Table 1, there is a close correspondence between the observed and predicted means for the five frequency classes of words. The most serious discrepancies are: (a) the predicted first fixation and single fixation durations for the lower frequency words are a bit long; (b) there is some non-monotonicity in the predicted first fixation durations, with the predicted means for Frequency Class 2 being larger than those for Frequency Class 1; (c) the percent of refixations for Frequency Class 2 is also underestimated. These results further indicate...
complex because first fixation durations are complex: They are a mixture of single fixation durations and the durations of first fixations that are followed by refixations. As can be seen in Table 1, the locus of the problem is not the predicted single fixation durations because they are monotonic and reasonably close to the observed values. Instead, the non-monotonicity stems from the fact that the model generates the right number of refixations for Frequency Class 1 but tends to underpredict the number of refixations for Frequency Classes 2 and 3. Additional simulations indicated that this anomaly can be rectified by softening the assumption of equal process durations for the labile stages of interword and intraword saccade programming.

Additional simulations

Pattern of variability

The above results indicate that E-Z Reader predicts the mean behavior of readers quite well. However, one can often be misled about a model’s utility and/or psychological validity by simply evaluating aggregate properties (Hintzman, 1991). Consequently, we decided to examine the patterns of variability predicted by the model. Figure 3 shows histograms of gaze durations that were observed by Schilling et al. (1998) and those predicted by E-Z Reader. In both cases, each datum represents the gaze duration on a single word for a single subject. As can be seen in Fig. 3, the absolute ranges and shapes of the observed distributions are in reasonably close agreement to those predicted by E-Z Reader, with the main discrepancy being that the observed distributions are less variable than those predicted by the model.

Frequency effects

The 48 sentences were used by Schilling et al. (1998) to examine word frequency effects during reading. Half contained high-frequency target words (over 46 per million, mean = 141; Francis and Kucera, 1982) and half contained low-frequency

6 Similar histograms were constructed for the observed and predicted first fixation durations, which also show a close correspondence (see Reichle et al., 1998).
Frequency Class 1 Gaze Durations

Frequency Class 2 Gaze Durations

Frequency Class 3 Gaze Durations

Frequency Class 4 Gaze Durations

Frequency Class 5 Gaze Durations

Eye movement control
target words (less than 4 per million, mean = 2). One obvious question, therefore, is whether E-Z Reader predicts the frequency effects reported by Schilling et al. on their target words. For the gaze durations, the observed means for the high- and low-frequency target words were 248 ms and 298 ms, respectively, resulting in a 50 ms frequency effect. E-Z Reader predicted means for the high- and low-frequency target words of 260 ms and 298 ms, respectively, giving a 38 ms frequency effect. Thus, the model accurately predicts both the absolute values for the means of the high- and low-frequency words, and the frequency effect.

We were also interested in whether the model could predict frequency effects on spillover (i.e., increased durations of fixations immediately subsequent to fixating a target word). Unfortunately, the corpus did not allow a particularly good empirical test because the high- and low-frequency words were in different sentences. As a result, we substituted the mean values for Schilling et al.’s (1998) high- and low-frequency target words (141 and 2 per million, respectively) into each of the 48 designated target positions, so that predicted effects of frequency on fixation times on the target word and on spillover durations could be calculated uncontaminated by differences in sentence frame. The mean frequency effect on gaze duration on the target words was close to that reported in the previous section: 35 ms. The mean spillover frequency effect (predicted increase in the gaze duration on word \( n + 1 \) was 22 ms, which is a bit smaller than the values observed in prior studies (which range from 30 to 50 ms), but not unreasonable. E-Z Reader predicts a spillover effect because, as the frequency of word \( n \) decreases, the time required to complete lexical access on the word increases, thereby reducing any preview benefit on word \( n + 1 \) that might otherwise occur while word \( n \) is being fixated. Less preview benefit on word \( n + 1 \) translates into increased processing times and hence longer fixation durations.

**Preview benefit**

We simulated the preview benefit from parafoveal processing on the 48 target words used by Schilling et al. (1998). In the simulated control condition (i.e., normal parafoveal preview), the target word was left unchanged in the parafovea (i.e., the rate of processing was moderated by the eccentricity parameters, as in previous simulations). In the simulated experimental condition (i.e., no parafoveal preview), lexical processing of the target words began when they were fixated. The predicted preview benefit (i.e., the difference between the two conditions) on the gaze duration on the target words was 40 ms, which corresponds well to observed values in prior studies that range from 40 to 60 ms.

---

7 We also examined frequency effects on first fixation and single fixation durations and found close correspondences between the E-Z Reader predictions and the results reported by Schilling et al. (1998).
Conclusions

We think that E-Z Reader is a step forward in understanding eye movements. In our view, it is likely to be the simplest model that can explain much of the lawfulness in the eye movement record. As with any model, it focuses on certain aspects of the data and not others. Thus, like Morrison, we did not attempt to explain where readers fixated within words. Although we believe that it would not be difficult to graft a module (based on McConkie et al.'s equations) that would simulate initial landing positions reasonably well, we think it would be quite difficult to evaluate the success of such an endeavor (i.e., evaluating whether the predictions would be sensitive to changes in the assumptions about cognitive processes, as opposed to low-level visual processes). In addition, as indicated above, the model is incomplete (as are all the models) in ignoring the demonstrated effects of higher-order processes on reading and interword regressions.

Even within this more limited scope, our model is clearly far from perfect. Among other things, we suspect that our assumptions about the causes of refixations are too simple. For example, the morphemic effects obtained by Hyönä and Pollatsek (1998) could not be predicted by the model. In addition, our assumptions about how word frequency and predictability affect lexical processing are clearly just zero-order approximations to the truth. However, we believe it is an important tool that helps us to make sense of the eye movement record.

Acknowledgements

Preparation of this chapter was supported by Grant HD 26765 from the National Institute of Health. The first author was supported by a Research Scientist Award from the National Institute of Mental Health (MH 01255) and the second author was supported by a Traineeship from the National Institute of Mental Health (MH 16745).

References


Eye movement control

Performance, 18, 163–172.
CHAPTER 12

Eye Movements During Scene Viewing: An Overview

John M. Henderson and Andrew Hollingworth
Michigan State University

Abstract

How do the semantic and visual characteristics of local scene regions influence the placement and duration of eye fixations during scene viewing? First, we review research on eye movement behaviour during scene viewing, focusing particularly on the influence of semantic information on eye movement behaviour. Second, we identify a number of factors that may influence eye movement behaviour in scenes, and suggest directions for future research. Finally, we propose a descriptive model of eye movement control in complex scenes.
Overview

In this chapter our goal is to provide an overview of eye movement patterns during scene viewing. There are at least three important reasons to understand eye movements in scene viewing. First, eye movements are critical for the efficient and timely acquisition of visual information during complex visual-cognitive tasks, and the manner in which eye movements are controlled to service information acquisition is a critical question. More generally, the interaction between vision, cognition, and eye movement control can be seen as a scientifically tractable testing ground for theories of the interaction between input, central, and output systems (Henderson, 1996). The vast majority of our current knowledge of eye movement control in complex visual-cognitive tasks derives from studies of reading, but a complete theory will require generalization to other ecologically valid tasks like scene viewing. Second, how we acquire, represent, and store information about the visual environment is a critical question in the study of perception and cognition. The tradition in the study of scene perception (and in perception and visual cognition generally) has been to study performance in tasks that use static, briefly presented images as stimuli. However, vision is a dynamic process in which representations are built up over time from multiple eye fixations. The study of eye movement patterns during scene viewing contributes to an understanding of how information in the visual environment is dynamically acquired and represented. Finally, eye movement data provide an unobtrusive, online measure of visual and cognitive information processing. In order to capitalize on this measure, it will be necessary to develop a more complete understanding of the manner in which visual-cognitive processing is reflected by eye movement behaviour.

This chapter is divided into three sections. First, we briefly review the literature on eye movement behaviour during scene viewing, with particular emphasis on where the eyes tend to fixate in a scene, and how long they tend to stay at a particular location. Our focus here is on static scenes. Reports of recent investigations of eye movements during the viewing of dynamic scenes can be found in Chapters 17–19. Second, we identify some largely unexplored factors that may affect the placement and duration of eye fixations in a scene. Finally, we offer a tentative descriptive model of eye movement control during scene viewing.

Review of eye movements during scene viewing

Eye movement behaviour during scene viewing can be divided into two relatively discrete temporal phases, fixations, or periods of time when the point of regard is relatively (though not perfectly) still, and saccades, or periods of time when the eyes are rotating at a relatively rapid rate to reorient the point of regard from one spatial
position to another. Useful pattern information is acquired during the fixations, with little useful pattern information taken in during the saccades due to a combination of visual masking and central suppression (Matin, 1974).

During fixations, the quality of the information acquired falls off rapidly and continuously from the center of the point of regard (fixation position) due to the optical properties of the eyes and the neural structure of the retina and visual cortex, with the highest quality visual information acquired from the spatial area immediately surrounding that point. Two important issues for understanding eye movement control during scene viewing are where the fixation position tends to be centered during scene viewing, and how long the fixation position tends to remain centered at a particular location in a scene. We will address these issues of fixation position and fixation duration next.

*Where do viewers look in a scene?*

**Effects of general region informativeness on fixation position**

The first systematic exploration of fixation positions in scenes was reported by Buswell (1935), who asked 200 participants to look at 55 pictures of different types of artwork under a variety of viewing instructions. An important result was that fixation positions were found to be highly regular and related to the information in the pictures. For example, viewers tended to concentrate their fixations on the people rather than on background regions when examining *Sunday on the Island of La Grande-Jatte* by Georges Seurat. These data thus provided some of the earliest evidence that eye movement patterns during complex scene perception are related to the information in the scene, and by extension, to perceptual and cognitive processing of the scene. Buswell concluded that “Eye movements are unconscious adjustments to the demands of attention during a visual experience. The underlying assumption in this study is that in a visual experience the center of fixation of the eyes is the center of attention at a given time.” (Buswell, 1935, pp. 9–10).

Buswell’s finding that informative scene regions tend to receive more fixations has been replicated many times. In the first study to explore this relationship analytically, Mackworth and Morandi (1967) divided each of two colour photographs into 64 square regions, and a group of participants then rated the informativeness of each region based on how easy it would be to recognize on another occasion. A new group of viewers then examined the pictures to decide which one of the two they preferred. Fixation density (the number of discrete fixations) in each of the 64 regions in each scene was found to be related to the informativeness rating of the region, with regions rated more informative receiving more fixations. Regions that received low informativeness ratings were often not fixated at all, suggesting that the scenes were filtered by peripheral vision and that uninformative regions could be rejected as potential fixation sites based on peripheral information alone. Mackworth
and Morandi (1967) also found that viewers were as likely to place their fixations on informative regions in the first two seconds of scene viewing as in other two-second intervals, providing evidence for relatively early, peripherally-based scene analysis.

The two pictures used by Mackworth and Morandi (1967) were visually simple: One depicted a pair of eyes within a hooded mask, and the other was a coastal map. Using images of more complex scenes taken predominantly from the Thematic Apperception Test, Antes (1974) provided additional evidence that region informativeness affects fixation position. Like Mackworth and Morandi (1967), Antes (1974) asked one group of viewers to rate each scene region according to the degree to which it contributed to the total amount of information conveyed by the whole picture. A different group of viewers then examined the scenes while their eye movements were recorded. Their task was to decide which scene they preferred. There were two main results relevant to fixation position. First, the density of fixations in a scene region was highly correlated with that region's informativeness, with regions rated more informative receiving more fixations, replicating Mackworth and Morandi (1967). Second, the first fixation position selected by a viewer (following the experimenter-induced initial fixation position at the center of the scene) tended to be within an informative region of a scene, suggesting rapid control of fixation position by scene characteristics.

In summary, the studies reviewed in this section suggest that the positions of individual fixations in a scene, including the position of the fixation after the first saccade, are determined in part by the informativeness of scene regions, with more fixations being directed to more informative regions. However, because region informativeness was determined by experimenter intuition (Buswell, 1935; Yarbus, 1967) or by viewer ratings (Antes, 1974; Mackworth and Morandi, 1967), visual and semantic informativeness were probably correlated in these studies. Therefore, it is not possible to determine whether there is an independent effect of semantic informativeness (i.e., the meaning of a region) beyond visual informativeness (i.e., the presence of discontinuity in texture, colour, luminance, and depth) on the positions of fixations in a scene. This issue is important because it is related to the question of whether fixation positions reflect cognitive operations as well as perceptual processes during scene viewing. If so, then semantically informative regions should be more likely to receive fixations during scene viewing, holding visual informativeness constant. We turn to this issue next.

Effects of semantic informativeness on initial fixation positions
In perhaps the first study to investigate the influence of semantic informativeness on fixation location, Loftus and Mackworth (1978) presented viewers with line drawings of scenes in which a manipulated target object was either high or low in semantic informativeness. Semantic informativeness was defined as the degree to
which an object was predictable within the scene, with the logic that an object unlikely to be found in a scene is more informative than an object likely to be found there. Importantly, visual informativeness was controlled by exchanging objects across scenes. For example, a farm scene could contain either a tractor (low informativeness) or an octopus (high informativeness). An underwater scene contained the same two objects, so that the semantic informativeness of the target objects was reversed. The two target objects occupied the same position in each scene. Participants viewed the scenes for four seconds each in preparation for a later recognition test. There were two main findings with respect to fixation location. First, viewers tended to fixate the inconsistent objects earlier during the course of scene viewing. Second, and more interestingly, viewers were more likely to fixate the semantically informative objects immediately following the first saccade within the scene. Because the distance of the saccade to the target objects averaged 6.5–8° of visual angle, these data suggest that viewer’s could determine in a single fixation the semantic informativeness of an object based on peripheral information, and that semantic informativeness could then exert an immediate effect on eye movement control.

De Graef, Christiaens and d’Ydewalle (1990) investigated the influence of semantic informativeness on eye movement patterns during scene viewing using a visual search task: Viewers searched line drawings of scenes for object-like figures that were not associated with any identifiable real-world object ("non-objects"). Using the same manipulation as had Loftus and Mackworth (1978), pre-specified target objects were placed in the scenes, and these objects were either semantically consistent or inconsistent (referred to by De Graef et al. as probability violations) with the scene. (Other types of violations were used as well, but we will focus on the semantic consistency here.) In contrast to Loftus and Mackworth (1978), De Graef et al. (1990) found no evidence that semantically inconsistent objects were fixated earlier than consistent objects. In fact, when De Graef et al. (1990) plotted the cumulative proportion of targets fixated as a function of informativeness, they found that viewers were no more likely to fixate the inconsistent than the consistent objects for the first 8 fixations. Our examination of this cumulative probability distribution (De Graef et al., 1990, Fig. 2) suggests to us that after the first 8 fixations in a scene, there was even some tendency for viewers to fixate the consistent objects sooner than the inconsistent objects. Clearly, these data do not support the view that the eyes are immediately drawn to semantically informative objects.

We recently conducted two new experiments to provide additional evidence concerning the role of semantic informativeness on eye movement patterns during scene perception (Henderson, Weeks and Hollingworth, 1999). In the first experiment, we attempted to replicate and extend Loftus and Mackworth (1978). We constructed 24 line drawings of real-world scenes generated from photographs (De
Graef et al., 1990). Semantically uninformative (consistent) target objects were drawn independently for each scene. Pairs of objects were then inserted into two yoked scenes to create two scenes in which the objects were informative and two in which the objects were uninformative, as shown in Fig. 1. The two target objects in a pair were always placed in the same location in a given scene so that the distance from the initial fixation point and lateral masking from surrounding contours would be controlled. During the experiment, we asked viewers to look at the scenes in preparation for a later memory test (which was, in fact, never given). The viewers were shown each of the 24 scenes once, half containing the informative target object for that scene and half containing the uninformative target object. Whether a given scene contained the informative or uninformative object was counterbalanced across viewers.

In contrast to Loftus and Mackworth (1978) but similar to De Graef et al. (1990), we found that viewers were no more likely to fixate the more informative target object than the less informative object early during scene viewing. First, viewers were no more likely to fixate the informative than the uninformative target after the first saccade in the scene, fixating the target immediately on about 10% of the trials in both conditions. Viewers were also no more likely to fixate the informative target after two saccades, fixating both types of objects after the first or second saccade in about 20% of the trials. Second, viewers initially landed on a target object after an average of about 11 fixations in the scene regardless of the semantic informativeness of the object. Third, the magnitude of the initial saccade to the target object was about 3°, and there was no evidence that these saccades were longer to the informative targets. These data suggest that the eyes are not initially driven by peripheral semantic analysis of individual objects.

In a second experiment, we introduced a visual search task to provide additional evidence concerning the relationship between semantic informativeness and initial fixation placement. During each trial, viewers were provided the name of a target object and then shown a line drawing of a scene. The viewer's task was to determine as quickly as possible whether the target object was present in the scene. Because of the instructions, viewers should have been highly motivated to find the targets as quickly as possible. If semantically informative objects can draw the eyes from peripheral regions of the scene, informative objects should be found more quickly than uninformative objects. As in the first experiment, however, viewers were no more likely to fixate the informative than the uninformative target after the first saccade in the scene. Instead, uninformative targets were fixated sooner (after about 3 fixations) than informative targets (after about 3.5 fixations). This finding presumably resulted from the fact that the positions of the uninformative objects were more constrained by the scenes, and so they were easier to find. For example, a blender in a kitchen is likely to appear on a counter-top rather than on the floor or elsewhere in the scene. A blender in a farmyard, by comparison, might appear just about
Fig. 1. Pairs of objects inserted into two yoked scenes (bar and laboratory) to create two scenes in which the objects were informative and two scenes in which the objects were uninformative.
anywhere, and would thus be more difficult to find. Finally, it is of interest that viewers moved their eyes to the targets more quickly in the second experiment (after about 3 fixations) than in the first (after about 11 fixations), suggesting that they could use peripheral visual information to guide their search. Even so, there was no evidence in either experiment that the eyes were drawn to semantically informative objects.

In summary, four experiments have examined the effects of semantic informativeness on initial fixation placement. Of these, one experiment has shown that the eyes are drawn to inconsistent object (Loftus and Mackworth, 1978), while three have shown that they are not (one experiment reported by De Graef et al., 1990, and two experiments reported by Henderson et al., 1999). Why might Loftus and Mackworth (1978) have found that viewers' initial fixations were drawn to semantically informative objects? One possible explanation is simply that the Loftus and Mackworth (1978) result was due to statistical error. This explanation seems possible given the relatively low spatial and temporal resolution of the eyetracking equipment that was available at the time of that study.

If we assume that the Loftus and Mackworth result was not due to statistical error, there are at least three other potential explanations for the inconsistency across studies. First, semantic informativeness and visual informativeness may have been correlated in the Loftus and Mackworth experiment (De Graef et al., 1990; Rayner and Pollatsek, 1992). This problem might arise if, for example, the consistent target objects were initially drawn in the scenes, and then the target objects were swapped across scenes. If this were true, then the result of semantic informativeness on initial fixations may actually have been due to visual factors. While we cannot say for certain whether this was a problem in the Loftus and Mackworth experiment, we do know that it was not a problem in our study: All scenes were created in the same way, and target objects were drawn independently of the scene backgrounds. Second, it could be that the scenes used by Henderson et al. (1999) and by De Graef et al. (1990) were visually more complex than those used by Loftus and Mackworth (1978). For example, if there were fewer contours in the Loftus and Mackworth scenes, then there may have been less lateral masking of individual objects, and so it may have been easier for viewers to semantically analyse peripheral objects. A third and related possibility is that the difference in results across studies may have been due to a difference in the size of the scenes used across studies. Larger scenes might lead to greater peripheral semantic analysis because the objects in the scenes would potentially be larger. In our study, the scenes subtended 10x14.5° while the Loftus and Mackworth (1978) scenes subtended 20x30°. Contrary to this hypothesis, however, De Graef et al. (1990) used scenes that subtended 20x30°, but as discussed above, they observed no influence of peripheral object semantics on early fixation placement.
There is an additional point that leads us to believe that the Loftus and Mackworth (1978) result was anomalous. Loftus and Mackworth (1978) observed an average saccadic amplitude of over 7° in their study. This average is roughly twice as large as the average saccadic amplitude typically observed in scene viewing experiments. For example, viewers in both of our experiments moved their eyes to the target objects from about 3–4° away, and very few saccades were in the 6–8° range (Henderson et al., 1999). (We report further evidence concerning distributions of saccadic amplitudes during scene viewing below.) The smaller saccadic amplitudes observed in our study were not due to the size of our scenes. Antes (1974) presented scenes that subtended 20x20° and observed average saccadic amplitudes in the same range as we did. Saida and Ikeda (1979) had participants view 14.4x18.8° pictures in preparation for a later memory test. In their control condition in which the entire scene was visible throughout the trial, the modal saccade length was under 2° and very few saccades were greater than 4°. Shiori and Ikeda (1989) reported that the median saccade size in a non-degraded viewing condition of their study was about 3° in 15x15° pictures, with 75% of all saccades between about 1.5 and 5.5° (estimated from Shiori and Ikeda, 1989, Fig. 10). Van Diepen, De Graef and d’Ydewalle (1995) found average saccadic amplitudes of about 3.4° when viewers searched for “non-objects” in line drawings of scenes that subtended 16x12°. (We ignore here conditions in the Saida and Ikeda (1979), Shiori and Ikeda (1989), and van Diepen et al. (1995) studies in which the amount of the scene that was visible during each fixation was manipulated using a window or mask that moved contingent on eye position; see Chapter 15 for information on these manipulations.) Overall, then, the saccadic amplitudes observed by Loftus and Mackworth (1978) appear to be anomalous given the remainder of the picture viewing literature.

Effects of semantic informativeness on fixation density
Fixation density can be defined as the number of discrete fixations within a given region. As reviewed above, viewers tend to cluster their fixations within informative regions of a scene (Antes, 1974; Buswell, 1935; Mackworth and Morandi, 1967; Yarbus, 1967). An examination of the figures presented by Buswell (1935) and Yarbus (1967) suggest that these clusters are not entirely determined by visual factors, but instead that viewers tend to concentrate their fixations on regions that are semantically interesting. Other evidence for an influence of scene semantics on fixation density comes from the manipulation of viewing instructions by Yarbus (1967). Yarbus found that when looking at a picture of I.E. Repin’s An Unexpected Visitor, viewers tended to concentrate their fixations on the people in the picture and particularly on their faces when they were attempting to determine the ages of the people, but tended to distribute their fixations more widely over the scene when they were attempting to estimate the material circumstances of the family.
Fixation densities in a scene region can be influenced both by the number of fixations made within that region each time it is examined (including the first time), and by the number of times viewers look back to that region. The figures presented by Buswell and Yarbus provide some qualitative evidence that both the number of initial fixations and the number of looks back to a scene region are affected by the informativeness of the region. There is also quantitative evidence supporting these conclusions. First, we have shown that the number of fixations viewers make in a region when that region is first fixated is affected by scene semantics (Henderson et al., 1999). In addition, there are two studies that provide quantitative evidence that viewers tend to return their gaze to semantically informative regions over the course of scene viewing (Loftus and Mackworth, 1978; Henderson et al., 1999). In our study, we found that viewers looked to informative objects about 3.3 times and to uninformative objects about 2.6 times on average over the course of 15 seconds of scene viewing.

In contrast to the results reported by Loftus and Mackworth (1978) and Henderson et al. (1999), Friedman (1979) found no effect of informativeness (likelihood) on the number of discrete looks to an object from a position beyond that object (Friedman and Liebelt, 1981). In that study, Friedman (1979) used a correlational approach to investigate the relationship between semantic consistency and eye movement patterns. Participants viewed line drawings of real-world scenes in preparation for a memory test in which "they would have to later be able to distinguish between the original pictures and new pictures in which, for example, only a small detail on one object would be different." Each scene contained objects that had been rated for their likelihood within the scene by a separate group of participants. A likely explanation for the lack of effect of semantic informativeness in the Friedman (1979) study is that the overall manipulation of informativeness was relatively weak; objects ranged continuously from very likely to somewhat likely in the scenes, with no truly unlikely objects. In our study (Henderson et al., 1999) as well as that of Loftus and Mackworth (1978), when a scene contained a semantically inconsistent object, that object was highly anomalous in the scene. Thus, the effect of semantic informativeness on fixation density was probably easier to detect in these latter studies.

Summary
In summary, the results of the past scene viewing studies indicate that the positions of fixations within a scene are non-random, with fixations clustering on informative scene regions (Antes, 1974; Buswell, 1935; Henderson et al., 1999; Mackworth and Morandi, 1967; Yarbus, 1967). However, the specific effect of semantic informativeness beyond that of visual informativeness on fixation position is less clear. Loftus and Mackworth (1978) observed that viewers tended immediately to fixate semantically informative objects, but neither De Graef et al. (1990) nor Henderson
et al. (1999) were able to replicate this effect. At the same time, both Loftus and Mackworth (1978) and Henderson et al. (1999) observed that viewers tended to look back more often to semantically informative than to uninformative scene regions, while Friedman (1979) did not observe this effect.

How long do viewers look at different scene regions?

While initial studies of eye movement patterns during scene viewing did not report viewing time measures (Antes, 1974; Buswell, 1935; Mackworth and Morandi, 1967; Yarbus, 1967), later research provides good evidence that the amount of time viewers fixate a scene region is dependent on the informativeness of that region. At a macro level of analysis, the total time that a region is fixated in the course of scene viewing (the sum of the durations of all fixations in that region) is correlated with the number of fixations in that region. Because, as discussed in the preceding section, fixation density is higher for visually and semantically informative scene regions, total viewing time spent on those regions also tends to be longer.

At a micro level of analysis, one can ask whether the durations of individual fixations and temporally contiguous clusters of fixations in a region (rather than the sum of all fixations) are also affected by region informativeness. Several commonly used micro-level measures of fixation time include first fixation duration (the duration of the initial fixation in a region), first pass gaze duration (the sum of all fixations from first entry to first exit in a region), and second pass gaze duration (the sum of all fixations from second entry to second exit in a region). In a recent series of experiments, van Diepen and colleagues have manipulated the quality of the visual information available during each fixation using a moving mask paradigm (see Chapter 15). Viewers searched for non-objects in real-world scenes, and the image at fixation was normal or was degraded. Image degradation was manipulated by reducing the contrast or overlaying a noise mask on the fixated region. When the image was degraded beginning at the onset of fixation, first fixation durations were longer than in a control condition, suggesting that the duration of the initial fixation is controlled, at least in part, by the acquisition of visual information from the fixated region. The van Diepen et al. study (Chapter 15) is the only direct exploration of the influence of visual factors on fixation duration during scene viewing that we are aware of, and there is currently no direct data concerning whether first fixation durations or gaze durations in a scene are affected by other correlates of visual informativeness such as contour density or contrast.

The effects of semantic informativeness on micro measures of fixation time during scene viewing have been studied more extensively. Loftus and Mackworth (1978) found that first pass gaze durations were longer for semantically informative (i.e., inconsistent) objects. Friedman (1979) similarly showed that first pass gaze duration on an object was correlated with the rated likelihood of that object in the
scene, with longer gaze durations on objects that were less likely to be found in a particular scene. Using the non-object counting task, De Graef et al. (1990) also found that first pass gaze durations were longer for semantically inconsistent objects, though this difference appeared only in the later stages of scene viewing. Finally, Henderson et al. (1999) found that first pass gaze duration and second pass gaze duration, as well as total fixation duration, were longer for semantically informative than uninformative objects.

The influence of semantic informativeness on the duration of the very first fixation on an object is less clear. De Graef et al. (1990) found that overall, first fixation durations on an object did not differ as a function of semantic informativeness. However, when first fixation duration was examined as a function of fixation moment (whether an object was fixated during the first or second half of all the fixations on the scene within which it appeared), first fixation durations on objects that were first encountered relatively late during scene exploration (following the median number of total fixations) were shorter on semantically uninformative (consistent) objects. We have recently analysed the first fixation duration data from our study (Henderson et al., 1999). Overall, we did not observe an effect of semantic informativeness, with mean first fixation durations of 317 ms in the informative condition and 314 ms in the uninformative condition, \( F < 1 \). In a subsequent analysis, we used a median split to divide the data into first fixations that occurred during the first versus second half of scene exploration, and again found no effect of semantic informativeness for either fixation moment. It appears that if they exist, effects of region semantics on first fixation durations during scene viewing are fragile.

Factors that may influence eye movement patterns during scene viewing

While there has been reasonable consistency in the eye movement patterns that have been observed across scene viewing studies, there are also some notable differences, as discussed above. It is often difficult to determine the cause of these differences because there are a number of potentially important factors that vary from study to study. These factors include image size, viewing task, viewing time per scene, image content, and image type. Table 1 summarizes the values of these factors used in the studies reviewed above. These factors could each produce main effects and could also interact with each other in complex ways to influence dependent measures of eye movement behaviour such as saccadic amplitudes, fixation positions, and fixation durations.

1 We note that both Loftus and Mackworth and Friedman called their measures duration of the first fixation, though the measures are equivalent to what has commonly been called gaze duration. Their eyetracking equipment did not have the spatial resolution to allow these investigators to examine the true first fixation duration.
### Table 1

Summary of methods for eye movement scene studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Image size</th>
<th>Viewing task</th>
<th>Viewing time per scene</th>
<th>Image type/content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buswell (1935)</td>
<td>varied</td>
<td>generally choose which images are pleasing</td>
<td>self-paced</td>
<td>colour paintings and images of other works of art</td>
</tr>
<tr>
<td>Mackworth and Morandi (1967)</td>
<td>16x16°</td>
<td>decide which image preferred</td>
<td>10 s</td>
<td>colour photographs of a mask and a coastline</td>
</tr>
<tr>
<td>Yarbus (1967)</td>
<td>varied</td>
<td>varied</td>
<td>varied; up to 30 min</td>
<td>colour paintings and images of other works of art</td>
</tr>
<tr>
<td>Antes (1974)</td>
<td>no more than 20°</td>
<td>decide which image preferred</td>
<td>20 s</td>
<td>monochrome shaded drawings (mostly from TAT test)</td>
</tr>
<tr>
<td>Loftus and Mackworth (1978)</td>
<td>20x30°</td>
<td>prepare for a later recognition memory test</td>
<td>4 s</td>
<td>black and white line drawings of real-world environments</td>
</tr>
<tr>
<td>Friedman (1979)</td>
<td>20x30°</td>
<td>prepare for memory test in which a small detail of one object may have changed</td>
<td>30 s</td>
<td>black and white line drawings (with some shading) of real-world environments</td>
</tr>
<tr>
<td>Friedman and Liebelt (1981)</td>
<td>20x30°</td>
<td>prepare for memory test in which a small detail of one object may have changed</td>
<td>30 s</td>
<td>black and white line drawings (with some shading) of real-world environments</td>
</tr>
<tr>
<td>De Graef, Christiaens and d’Ydewalle (1990)</td>
<td>20x30°</td>
<td>count non-objects</td>
<td>8 s</td>
<td>black and white line drawings of real-world environments</td>
</tr>
<tr>
<td>Henderson, Weeks and Hollingworth (1999) Exp. 1</td>
<td>10x14.5°</td>
<td>prepare for memory test in which a small detail of one object may have changed</td>
<td>15 s</td>
<td>black and white line drawings of real-world environments</td>
</tr>
</tbody>
</table>

(continued)
Table 1 (continuation)

<table>
<thead>
<tr>
<th>Study</th>
<th>Image size</th>
<th>Viewing task</th>
<th>Viewing time per scene</th>
<th>Image type/content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henderson, Weeks and Hollingworth (1999) Exp. 2</td>
<td>10x14.5°</td>
<td>search for a pre-specified target object</td>
<td>until response</td>
<td>black and white line drawings of real-world environments</td>
</tr>
<tr>
<td>Henderson and Hollingworth (1997)</td>
<td>10x14.5°</td>
<td>prepare for memory test in which a small detail of one object may have changed</td>
<td>15 s</td>
<td>black and white line drawings of real-world environments</td>
</tr>
</tbody>
</table>

**Image size**

One example of how variation in one factor can make interpretation across studies difficult arises in the case of the effect of the semantic informativeness of a scene region on the amplitude of a saccade to that region. As discussed above, it is possible that the amount of the visual field subtended by a depicted scene affects saccadic amplitudes, and that the influence of semantics on amplitude is mediated by this factor. While our review above led us to conclude that mean saccadic amplitudes range between about 2 and 4° despite scene size when the scene subtends between 10 and 20°, there are no published studies that were designed to directly examine saccadic amplitudes as a function of scene size. It is possible that cross-experiment comparisons are misleading because other factors have not been held constant, and that saccadic amplitudes do scale with scene size. Only studies designed to directly test these possibilities will be able to answer this question.

**Viewing task**

Another important variable in scene viewing is the task given to the viewer. Buswell (1935) and Yarbus (1967) both presented evidence that viewers place their fixations in a scene differently depending on the viewing task. However, these studies were descriptive in that the conclusions were based on a qualitative analysis of the viewing behaviour of particular individuals on specific scenes under differing viewing instructions. In the study described above, we compared eye movement patterns in a memory preparation task and a visual search task (Henderson et al., 1999). This is, to our knowledge, the only study that has held the nature of the stimulus image constant and quantitatively examined the influence of viewing task on eye movement patterns. As discussed above, our results showed that participants
made fewer fixation in a scene prior to fixation on a particular object when they were searching for that object than when they were trying to memorize the scene. We must point out, however, that in this study the comparison of eye movement patterns across tasks was accompanied by variations in other factors. For example, different groups of viewers took part in each task, and the amount of time viewers were allowed to look at the scenes differed in the two viewing tasks, with 15 seconds of viewing time in the memory task and self-termination of the scenes in the visual search task.

**Viewing time per scene**

Scene viewing time is another potentially important factor in determining viewing patterns over scenes. In the studies reviewed above, scene viewing time has ranged from a minimum of 4 seconds per scene (Loftus and Mackworth, 1978) to a maximum of 30 minutes per scene (Yarbus, 1967), though one has to wonder at the patience of the viewer in the latter case. Other studies have used scene presentation durations that fall between these extremes, as shown in Table 1. In Buswell's study, scene viewing time was determined by the viewer (Buswell, 1935), and there were large individual differences in the length of time viewers wanted to look at the scenes. Thus, it is not clear what the appropriate duration for scene presentation should be. In addition, there has been some evidence that viewing patterns change over time. For example, Friedman (1979) found that the difference in gaze durations on high and low probability objects decreased from 342 ms on first entry (first pass gaze duration) to 78 ms on the third and higher re-entries. This change appeared to be due to a decrease in gaze durations on each entry for low probability objects, but not for high probability objects. Given that viewing patterns and eye movement measures may change over the course of scene viewing, these measures may also change depending on how long viewers are allowed to look at a picture. There are currently no data available on this topic.

**Image content**

The content of the images presented to viewers in eye movement studies has varied markedly, as can be seen in Table 1. At one extreme, Buswell (1935) and Yarbus (1967) obviated the problem of image content by using a wide variety of types of scenes, while at the other extreme, Mackworth and Morandi (1967) presented only two images, one of a hooded face, and the other of an aerial view of a coastline. Both of these latter images contained large areas of uniform background. It is not clear what effect the use of such a restricted set of images has on viewers’ eye movements. It may be that scenes with different content (e.g., outdoor versus indoor, large-scale spaces versus small-scale spaces) produce systematic effects on eye movement patterns. Currently, this is another unexplored issue.
The final potentially important factor that we will discuss here is the manner in which a scene is depicted. As shown in Table 1, scene depiction has varied from line drawings (e.g., Friedman, 1979; Henderson et al., 1999; Loftus and Mackworth, 1978) to monochrome shaded drawings (Antes, 1974) to colour paintings (Buswell, 1935) and colour photographs (Mackworth and Morandi, 1967). All of the work that has so far been conducted to examine the influence of semantic informativeness on eye movement patterns has used line drawings as stimuli (De Graef et al., 1990; Friedman, 1979; Henderson et al., 1999; Loftus and Mackworth, 1978). It is not yet clear to what extent the results generated from one type of image type will generalize to other image types. Furthermore, it will ultimately be important to determine whether the results that are derived from images that depict real-world scenes generalize to the visual world itself, that is, to the situation in which the viewer is looking at the actual visual environment. The introduction of viable head-mounted eyetracking equipment in the last few years should help to encourage the exploration of this latter issue.

In order to begin to get a feel for the influence of image type on eye movement patterns, we recently conducted a study in which we contrasted viewing behaviour on line drawings, colour photographs, and computer-rendered 3-D colour images of real-world scenes (Henderson and Hollingworth, 1997). Eight viewers examined each of 30 scenes for 15 seconds each. The viewing instructions were the same as those used by Henderson et al. (1999): Viewers were told that after they had viewed all of the scenes, they would be given a memory test in which they would have to discriminate the test scenes from new scenes in which only a small detail of a single object might be changed. Ten exemplars of each of three image types (colour photographs, line drawings, and computer-rendered 2-D images from 3-D models) were presented. All images depicted common real-world scenes. The colour photographs and line drawings depicted the same 10 categories of scenes, while the rendered images depicted rooms in a house. All of the images were viewed by the same set of 8 participants. The images were presented in a random order determined individually for each participant, so that participants would be less likely to develop different viewing strategies for different image types. The three image types were presented on the same SVGA display system at the same resolution (800x600 pixels) and visual angle (10x14.5°). Eye movement data were collected using a Fourward Technologies Generation 5.5 dual-Purkinje image eyetracker. Further details of our general method can be found in Henderson et al. (1999).

The eye movement data were analysed using analysis software developed in our laboratory (see Henderson et al., 1999). Colour Plate 1a shows the eye movement pattern of one viewer on a colour photograph of a kitchen, and Colour Plate 1b shows the pattern for that same viewer on a line drawing depicting a similar scene. In the figure, the green dots represent fixations, the red numbers indicate the ordinal
Colour Plate 1a. Viewing pattern for one participant viewing a photograph of a kitchen. Green dots represent discrete fixations, ordinally numbered in red. Green lines represent saccadic vectors.

Colour Plate 1b. Viewing pattern for the same participant viewing a line drawing of a kitchen. Green dots represent discrete fixations, ordinally numbered in red. Green lines represent saccadic vectors.
number of the associated fixation, and the green lines represent (straightened) saccade vectors. As can be seen in the figure, this viewer tended to distribute her fixations over a relatively large area of the scene, with more fixations concentrated on the more distant counter top where there were many objects than on the closer but empty counter top. Colour Plate 2 presents a contour plot of total fixation time summed across the eight viewers for the kitchen photograph (Plate 2a) and line drawing (Plate 2b). In this figure, cooler colours represent less total fixation time, and hotter colours represent more fixation time, with the colours ordered from dark blue through bright red. This figure illustrates that the viewers as a group spent the majority of their time fixating the informative regions of the scenes. A particularly striking example of this tendency can be seen by comparing the total fixation times (Colour Plate 2) on the closer counter top for the photograph and line drawing. In the colour photograph where the close counter was empty, very little fixation time was spent in that region of the scene. In contrast, in the line drawing the close counter contained a rolling pin, and fixation time clearly was devoted to that object. This contrast points out very nicely how fixation time is directed to informative scene regions.

In a quantitative analysis of these data, we found small but reliable differences in eye movement parameters as a function of image type. First, viewers made an average of 36.5 fixations in each scene. They tended to fixate the photographs reliably fewer times (34.8) than the line drawings (36.8) or rendered scenes (37.9). Second, the duration of each fixation was on average 327 ms. Offsetting the reduced number of fixations on the photographs, the duration of the average fixation in the scene was reliably longer for photographs (336 ms) than for line drawings (324 ms) or rendered scenes (321 ms). Given the instructions to prepare for a memory test, it is possible that fixations were longer on the photographs because more visual information was available to commit to memory in photographs than in the other, more schematic stimuli. Finally, the mean saccade length was 2.4° and did not differ as a function of image type.

Despite the small differences in eye movement parameters across image type, the consistency of the viewing patterns is quite striking, as can be seen in Colour Plates 1 and 2. Also, the specific nature of the fixations and saccades in the different scene types was very similar. For example, Fig. 2 shows the frequency distributions of the fixation durations (top panel) and saccadic amplitudes (bottom panel) as a function of image type for all participants. As can be seen in the figure, fixation duration and saccadic amplitude distributions were remarkably similar for the different image types, with modal fixations durations of about 220 ms and modal saccadic amplitudes of about 0.5°.

For the purposes of comparison, we have also plotted the data from a reading study in Fig. 2. In this study, conducted by Fernanda Ferreira and Melissa Johnson, 36 participants each read 20 paragraphs of text for comprehension. The text was
Colour Plate 2a. Contour plot of fixation times summed across eight viewers for a photograph of a kitchen. The scene was originally presented in colour. On this image, cooler colours represent less total fixation time, and hotter colours represent more fixation time, with the colours ordered: dark blue, light blue, dark green, light green, yellow, red.

Colour Plate 2b. Contour plot of fixation times summed across eight viewers for a line drawing of a kitchen. Cooler colours represent less total fixation time, and hotter colours represent more fixation time, with the colours ordered: dark blue, light blue, dark green, light green, yellow, red.
Fig. 2. Frequency distributions of the fixation durations (top panel) and saccadic amplitudes (bottom panel) for participants viewing scenes (line drawings, colour photographs, and colour 3D renderings), and reading text.

presented in graphics mode on the same display system, and subtended the same visual angle as the scenes we used as stimuli. Importantly, the eye movement data were collected using the same eyetracking system, and were analysed using the same software parameters for determining the onset of a fixation and saccade as we used in the scene study. To our knowledge, this is the first report of a direct comparison of eye movement behaviour in reading and image viewing. As can be seen in the top panel of Fig. 2, the modal fixation duration in reading and scene
viewing was the same. However, fixation duration was considerably less variable for reading, with fewer fixations lasting longer than 340 ms. The greater number of longer fixations in scene viewing appears to account for the common finding that mean fixation duration is longer in scene viewing than in reading. The bottom panel of Fig. 2 shows that the modal saccadic amplitude is longer but less variable in reading than in scene viewing. A generalization that can be extracted from Fig. 2 is that both fixations and saccades are more variable in scene viewing than in reading. Of course, these comparisons must be viewed with some caution; a task that emphasizes memory (as used in the scenes viewing study) may differ in important ways from a task that emphasizes comprehension (as used in the reading study).

A saliency map framework for eye movement control in scene viewing

In this final section we want to outline a model of eye movement control in scene viewing (Henderson, 1992; Henderson et al., 1999). This framework is, at this point, descriptive rather than quantitative, but we believe that it is specific enough to generate new predictions. The framework is also couched in such a way that it could be computationally modelled, and several of the proposed components have been modelled as independent modules for other purposes (e.g., Mahoney and Ullman, 1988). The framework is also in the spirit of the computational model proposed by Reichle et al. (1997; see also Chapter 11). The saliency map framework expands on ideas originally discussed by Henderson (1992), which in turn extended the model of eye movement control in reading proposed by Morrison (1984) and elaborated upon by Henderson and Ferreira (1990; Henderson and Ferreira, 1993) and Rayner and Pollatsek (1989; see Chapter 11). The framework is meant to account for fixation placement and fixation duration for those fixations that are directed in the service of visual analysis and cognitive processing. The framework is not meant to account for more fine-grained eye movement behaviour such as micro-saccades and ocular drift, and also ignores other oculomotor phenomena like the global effect (Findlay, 1982) and the optimal viewing position effect (O'Regan, 1992; Vitu et al., 1995), though of course we do not deny their existence.

In the saliency map framework, a representation of potential saccade targets is generated from an early parse of the scene into regions of potential interest and a background that is relatively undifferentiated. This initial parse is derived from a fast early analysis of the low frequency information available during an initial fixation in the scene. The positions of the regions of potential interest are coded in a representation of visual space and are assigned a saliency weight. The combination of spatial position and saliency weight is the saliency map (Mahoney and Ullman, 1988). Initially, region salience is determined by visual factors such as luminance, contrast, texture, colour, contour density, and so on, because this is the only information that is available about each region. Salience may also initially be
modified by top-down factors such as the viewer’s task, but only if the task can be based on these visual factors. For example, the salience of scene regions that are the same shape as a search target (e.g., rectangular) would be increased, leading to relatively efficient search, as found by Henderson et al. (1999). Further, a global semantic analysis of the scene could contribute to search by constraining the likely position of semantically consistent targets.

According to the saliency map framework, the visual information acquisition system follows two simple rules: (1) Allocate visual-spatial attention to the scene region with the highest saliency weight (Koch and Ullman, 1985), and (2) Try to keep the eyes fixated on the attended scene region (Henderson, 1992; Henderson and Ferreira, 1990; Henderson and Ferreira, 1993). Because initially saliency weights are determined only by visual factors, initial attention allocation and initial fixation placement will be determined by visual rather than semantic characteristics of the scene. When the eyes are in fixation, the amount of time they remain stationary will primarily be determined by the amount of time needed to complete perceptual and cognitive analysis of that region. Once processing is complete, the saliency weight for that region will be reduced and attention will be released. Attention is then reallocated to the region that now has the highest saliency weight, and the eyes are programmed to move to that region (Henderson, 1992; Henderson, Pollatsek and Rayner, 1989). If perceptual and/or cognitive analysis of the currently fixated region is taking too long given the present fixation position (i.e., the rate of information acquisition is too low), then visual-spatial attention will be reallocated within the current region to optimize information acquisition, and a refixation will be programmed to the new locus of attention (Henderson, 1993; see also McConkie, 1979, for a similar explanation of refixations in reading). Selecting a sub-region within a region is assumed to be based on constructing a saliency map at a finer scale of resolution. The reallocation of attention within a scene region accounts for the finding that scene regions that are difficult to analyse are more likely to receive refixations (Henderson et al., 1999). Refixations may also be programmed based on oculo-motor factors alone (Henderson, 1993; O’Regan, 1992).

In the saliency map framework, initial movements of the eyes during scene viewing should be controlled by stimulus features rather than by cognitive features. However, as individual scene regions are fixated and cognitively analysed, saliency weights will be modified to reflect the relative cognitive interest of those regions. In other words, we assume that the source of the saliency weight for a given scene region will change from primarily visual to primarily cognitive interest as regions are fixated and understood. As scene viewing and understanding progresses, region salience will become heavily determined by factors such as semantic informativeness. The eyes will then be more likely to be sent to regions of cognitive salience rather than drawn by regions of visual salience, leading to greater fixation density and total fixation time on semantically interesting objects and scene regions.
In contrast to initial fixation placement, the amount of time the eyes initially remain fixated in a region, and the number of initial refixations in that region, should be affected by semantic aspects of the region right from the first time the region is fixated, because these aspects of eye movement behaviour are determined by the amount of time required to complete cognitive analysis of that region. In other words, the length of time the eyes remain in a region is controlled primarily by the needs of perceptual and cognitive analysis of the region. In addition, to the extent that additional looks back to a region are needed for additional cognitive analysis of that region, fixation times during these additional looks should also be influenced by the same factors that influence initial fixation times.

Conclusion

There had been something of a hiatus in the exploration of eye movements during scene viewing following the studies that were conducted in the 1960s and ‘70s. Now, after 20 years of relative inactivity, there has been a resurgence of interest in this topic, as exemplified by many of the chapters in this volume. We see this renewed interest as positive and necessary: while a great deal has been learned about eye movement behaviour during scene viewing, there are still a large number of unresolved questions. Ultimately, answers to these questions will provide a more complete understanding of the interface between perception and action, will contribute to our knowledge of scene perception, and will allow eye movement monitoring to fulfill its promise as a noninvasive, on-line measure of visual-cognitive processing.

Acknowledgements

We would like to thank Femanda Ferreira for her lively discussions of the issues raised here, and several anonymous reviewers for their comments. The work described in this chapter was supported by grants from the U.S. Army Research Office and the National Science Foundation to John M. Henderson, and by a National Science Foundation graduate fellowship to Andrew Hollingworth. The contents of this article are those of the authors and should not be construed as an official Department of the Army position, policy, or decision.

References


