Published in final edited form as:
Psychol Sci. 2009 January ; 20(1): 6–10. doi:10.1111/j.1467-9280.2008.02243.x.

Eye Movements and Visual Encoding During Scene Perception

Keith Rayner[1], **Tim J. Smith**[2], **George L. Malcolm**[2], and **John M. Henderson**[2]
[1] University of California, San Diego

[2] University of Edinburgh

Abstract

The amount of time viewers could process a scene during eye fixations was varied by a mask that appeared at a certain point in each eye fixation. The scene did not reappear until the viewer made an eye movement. The main finding in the studies was that in order to normally process a scene, viewers needed to see the scene for at least 150 ms during each eye fixation. This result is surprising because viewers can extract the gist of a scene from a brief 40- to 100-ms exposure. It also stands in marked contrast to reading, as readers need only to view the words in the text for 50 to 60 ms to read normally. Thus, although the same neural mechanisms control eye movements in scene perception and reading, the cognitive processes associated with each task drive processing in different ways.

The neural mechanisms that underlie oculomotor activity do not vary as a function of the task viewers engage in; there is not one oculomotor system for looking at scenes, another for visual search, and another for reading. Eye movements are essential in these tasks because the eyes must be placed on the part of the scene or text viewers want to process in detail in foveal vision (Henderson, 2003; Rayner, 1998, in press). Does the oculomotor system react in the same way to stimuli in these different tasks?

In the present studies, we utilized a gaze-contingent display change paradigm (Henderson & Hollingworth, 1999; McConkie & Rayner, 1975; Najemnik & Geisler, 2005; Rayner, 1975; Rayner & Bertera, 1979) to precisely vary when a visual mask obscured a scene that viewers examined. In reading, it has been demonstrated that, if readers are allowed to examine text for 50 to 60 ms on each eye fixation before a visual mask appears (which makes further visual encoding of text impossible on that fixation), they read quite normally (Liversedge et al., 2004; Ishida & Ikeda, 1989; Rayner, Inhoff, Morrison, Slowiaczek, & Bertera, 1981; Rayner, Liversedge, & White, 2006; Rayner, Liversedge, White, & Vergilino-Perez, 2003). Given that it is also well-known that viewers can obtain the gist of an entire scene from a brief exposure of 40 to 100 ms (Biederman, 1972; Biederman, Mezzanotte, & Rabinowitz, 1982; Castelhano & Henderson, 2008; Potter, 1975; Rousselet, Joubert, & Fabre-Thorpe, 2005; Schyns & Oliva, 1994; Thorpe, Fize, & Marlot, 1996), it would be tempting to think that the amount of time viewers need to glimpse a scene on each fixation should likewise be in the range of 50 to 60 ms. We explicitly tested this hypothesis by masking scenes 25, 50, 75, 150, 200, and 250 ms after the beginning of each fixation.

EXPERIMENT 1

In Experiment 1, participants were asked to find a specific target object in a scene. Thus, for example, in a warehouse scene, viewers were asked to locate a broom. Eye movements were

Address correspondence to Keith Rayner, Department of Psychology, University of California, San Diego, La Jolla, CA 92093–0109, e-mail: krayner@ucsd.edu.

recorded, and on each fixation, a mask appeared after a specified interval from the beginning of the fixation. Once the mask appeared, the scene did not reappear until the viewer made a saccade to another location.

Method

Participants—Ten University of Edinburgh undergraduate students with normal or corrected-to-normal vision participated. They were naive concerning the purpose of the experiment.

Experimental Apparatus and Procedure—Eye movements were monitored via a SR Eyelink 1000 eye-tracker, with a spatial resolution of less than 1/4 degree (eye position was sampled every millisecond). Saccades were defined with a 50 deg/s velocity threshold using a nine-sample saccade-detection model. Viewing was binocular, but only the right eye was tracked. The images were presented on a 21-in. cathode ray tube monitor at a viewing distance of 90 cm with a refresh rate of 140 Hz. The computer kept a complete record of the duration, sequence, and location of each eye fixation.

The viewers' task was to locate the target object as quickly and accurately as possible. At the onset of each trial (see Fig. 1), a target word was presented for 800 ms, followed by a fixation cross for 400 ms and, then, the scene. Presentation of the scene was interrupted after a predefined viewing time (25, 50, 75, or 150 ms) during each fixation by the sudden presentation of a contrast-matched color noise mask. This sequence continued until either the viewer made a response or 20 s had elapsed. In addition to the mask conditions, a control condition was included in which the scene was presented entirely without any mask.

Materials—Sixty unique full-color 800 × 600 pixel photographs of real-world scenes[1] from a variety of scene categories were used in the experiment.

Results

An analysis of variance (ANOVA) on each of the measures shown Table 1 yielded an effect of mask onset on search time, $F(4, 36) = 12.30$, $p < .001$; fixation duration, $F(4, 36) = 30.94$, $p < .001$; saccade length, $F(4, 36) = 5.90$, $p < .01$; and search accuracy, $F4, 36) = 52.36$, $p < .001$. For search time, pair-wise comparisons between the different mask-onset conditions revealed that all masking conditions yielded significantly longer times than the control condition, all ps $< .001$ (p_{rep}s $\geq .99$) except for the 150-ms mask-onset condition ($p = .077$, $p_{rep} = .88$, $d = 0.85$). For fixation duration, all mask conditions produced significantly longer fixations than the control condition (all ps $< .001$, $p_{rep} \geq .99$). For saccade length, all mask-onset conditions yielded significantly shorter saccade amplitudes (all ps $< .05$, $p_{rep} \geq .95$) except for 150-ms mask-onset condition ($p = .063$, $p_{rep} = .95$, $d = -0.90$). Finally, for search accuracy, the probability of correctly responding was much lower for the 25-, 50-, and 75-ms mask-onset conditions than for the control condition (ps $< .001$, $p_{rep} \geq .99$); the 150-ms mask-onset condition (91% correct) was much closer to the control condition (99%), but the difference was significant ($p < .05$, $p_{rep} = .94$, $d = -1.14$).

Discussion

Although viewing text for 50 to 60 ms prior to mask onset seems to be sufficient for reading to proceed effectively (Rayner et al., 1981, 2003), viewers needed much longer than this to effectively encode the scene. Indeed, even with the 150-ms mask onset, performance did not

[1]The scenes were drawn from the pool used by Castelhano and Henderson (2008). Although the tasks in Castelhano and Henderson were different from those used in this study, we used these images because they supported very fast (40–50 ms) scene gist extraction.

reach the level of the no-mask control condition. To determine more precisely how long viewers need to view the scene so that the mask onset is not disruptive, we carried out a second experiment in which the mask onset was delayed for longer time intervals. We also varied the task to determine whether the longer viewing time needed in Experiment 1 was a peculiarity of visual search.

EXPERIMENT 2

In Experiment 2, mask onset delays were 75, 150, 200, and 250 ms. Half of the viewers were again asked to search for a specific target item in the scene (search task), and the other half examined each scene in anticipation of a recognition memory test given at the end of the experiment (memory task).

Method

Participants—Twenty naive University of Edinburgh undergraduate students with normal or corrected-to-normal vision participated.

Experimental Apparatus and Procedure—The apparatus was identical to Experiment 1, as was the procedure for half of the viewers. The remaining viewers were instructed to examine the scenes in anticipation of a recognition memory test. In the search task, the scene remained until either a response occurred or 20 ms elapsed; in the memory task, the scene was only presented for 6 s (see Fig. 1). The mask-onset delays were 75, 150, 200, and 250 ms, and a control condition was again included in which a mask did not appear on each fixation. After the encoding phase, participants in the memory task were presented with 120 randomly mixed scenes, 60 of which were previously presented (old) and 60 were new. Participants were instructed to identify as quickly as possible whether the scenes were either "old" or "new" and then rate the confidence of their response on a scale from 0 (*no confidence*) to 3 (*full confidence*). Confidence ratings were uninformative and therefore are not presented.

Materials—The materials were identical to those used in Experiment 1 except for the addition of the 60 new scenes in the memory task.

Results

Although fixation durations and saccade amplitudes were longer in the memory task than the search task, there were no interactions between task and mask onset delay. Hence, we discuss the data collapsed over the two tasks.[2] Table 2 shows the measures as a function of mask onset. Baseline performance, when no mask appeared (i.e., the scene appeared normally and the viewer had to find the search target or examine the scene in anticipation of a memory test), can again be judged from the control condition.

As in Experiment 1, ANOVAs on the measures in Table 2 yielded significant effects of mask onset on search time, $F(4, 36) = 3.53$, $p < .05$; fixation duration, $F(4, 72) = 24.95$, $p < .001$; saccade length, $F(4, 72) = 28.09$, $p < .001$; search accuracy, $F(4, 36) = 16.94$, $p < .001$; and recognition accuracy, $F(5, 40) = 7.13$, $p < .001$. In the search task, search time and search accuracy only differed significantly from the control in the 75-ms mask-onset condition (search time: $p < .05$, $p_{rep} = .99$, $d = 0.99$; search accuracy: $p < .001$, $p_{rep} = .99$, $d =$

[2]The accuracy measures are shown separately for the search and memory tasks; it is not appropriate to collapse over them because they measure different things. Accuracy in the search task refers to the probability of correctly identifying the location of the target, whereas accuracy in the memory task refers to performance on a recognition memory task in which viewers had to indicate whether a given scene was old or new. The accuracy measure of correctly identifying a new scene as new was. 97. There is no equivalent to search time in the memory task because all scenes were presented for the same duration during the memory encoding period.

−2.02). Accuracy on the recognition memory test was only significantly worse than the control for the 75-ms mask-onset condition (p <.01, p_{rep} =.94, d = −1.18). Over both tasks, all mean fixation durations and saccade lengths differed significantly from the control condition in all conditions (ps <.01, p_{rep}s ≥.90) but the 250-ms mask-onset condition (fixation duration, p =.053, p_{rep} =.76, d = 0.34; saccade length, p =.138, p_{rep} =.70, d = −0.26).

Discussion

A number of results from Experiment 2 are striking. First, as in Experiment 1, the 75-ms mask-onset delay did not provide viewers enough time to process the scenes; this condition significantly increased search time and average fixation duration on each scene, and also reduced saccade length. This result, along with the results in Experiment 1 in which 25- and 50-ms mask-onset delays resulted in considerable disruption to scene processing, clearly demonstrates that it takes longer for viewers to encode the stimulus material in scene perception than it takes for readers to encode words in reading.[3] It is also clear that acquiring gist alone is not sufficient for normal scene processing.

Second, in terms of the search time and the accuracy measures, there were no significant differences between the control condition and the other mask-onset delays beyond the 75-ms delay. Thus, it would seem that 150 ms is needed to encode the scene material prior to the onset of the mask for processing to occur relatively normally. Again, this is much longer than the time needed to encode the material during reading, and is interesting in light of the well-known finding that viewers can encode the gist of a scene very quickly. Although they can perhaps know the gist from a brief exposure, the present results suggest that the details extracted from the scene take longer to accumulate.

Third, although the search time and accuracy measures reached asymptote at 150 ms, this was not the case for either saccade length or fixation duration. For saccade length, performance reached asymptote at 250 ms. For fixation duration, there was a steady decrease in fixation duration with each level of mask delay from 150 to 250 ms, which was on the order of 25 ms for each 50-ms increase in the mask onset. Likewise, there was a 24-ms decrease in fixation duration from the 250-ms mask-onset condition to the control condition. We suspect that the reason for the differences is saccade inhibition associated with the onset of the mask (Henderson & Pierce, 2008; Reingold & Stampe, 2002).

GENERAL DISCUSSION

The present studies demonstrate that viewers need at least 150 ms to encode stimulus properties during eye fixations in scene perception. This finding indicates that the 40 to 100 ms needed to acquire sufficient information to understand the gist of a scene is not adequate for the type of complete scene analysis undertaken during typical scene viewing. This finding also stands in marked contrast to similar studies in which text is masked during reading, which have demonstrated that readers need only 50 to 60 ms to encode words and read normally.[4] This conclusion is reinforced by Figure 2, which shows the fixation-duration data from Experiments 1 and 2 along with data from a comparable reading study (Rayner et al., 1981).

[3]Earlier, van Diepen, Ruelens, and d'Ydewalle (1999) used a masking technique like that used here and reported that visual information in scene perception is encoded within 45 to 75 ms. However, they used very simple line drawings that were not as complex as the color photographs we used.

[4]In most reading studies, only the fixated word was masked, whereas the entire scene was masked in our study. Thus, viewers might have been less certain about where to move next in the scene experiments than in the reading experiments. However, saccade size was fairly large in even the 50- and 75-ms onset conditions. Also, Rayner et al. (1981) included a condition in which the entire line was masked, and it was still the case that 50 ms was sufficient for reading to proceed normally.

What is it about scene viewing that makes it different from both reading and gist processing, and why does the scene need to be presented for a longer time before the mask onset? First, perhaps it takes longer than 50 ms to encode the general meaning of the scene. However, as already noted, the gist can be understood from a 40- to 50-ms scene exposure. Second, perhaps it takes more presentation time to encode fixated objects in scenes than it does to encode words in text. Contrary to this hypothesis, studies have shown that objects can be encoded from very brief presentations (~50 ms), even when the object appears in a scene (Davenport & Potter, 2004; Li, Iyer, Koch, & Perona, 2007; Rousselet, Macé, & Fabre-Thorpe, 2003; Thorpe, Fize, & Marlot, 1996). Third, perhaps it takes more display time to acquire the spatial information needed to find a saccade target. Again, there is evidence that spatial structure can be encoded very rapidly from scenes (Castelhano & Henderson, 2008; Li et al., 2007; Schyns & Oliva, 1994). Thus, it appears that each of the component processes taking place within a fixation (understanding the meaning of the scene, identifying the object being looked at, and locating potential places to look next) can all operate effectively given 50 ms of scene presentation.[5] From this perspective, it is surprising that three times that value is needed.

Our results are consistent with other data (Rayner, Li, Williams, Cave, & Well, 2007) demonstrating that eye movement parameters in reading do not correlate well with those in scene perception, face perception, and visual search. Although the neural mechanisms controlling the oculomotor system are invariant across tasks, the cognitive processes associated with the tasks manifest themselves in different ways. Specifically, in the present case, the encoding of the scene properties takes longer than the encoding of words in reading.

Acknowledgments

This research was supported by National Institutes of Health Grant HD26765 and by Economic and Social Sciences Research Council Grant RES-062–23–1092. We thank Geoff Loftus and Marvin Chun for excellent comments.

References

Biederman I. Perceiving real world scenes. Science. 1972; 177:77–80. [PubMed: 5041781]

Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology. 1982; 14:143–177. [PubMed: 7083801]

Castelhano MS, Henderson JM. The influence of color on the activation of scene gist. Journal of Experimental Psychology: Human Perception and Performance. 2008; 34:660–675. [PubMed: 18505330]

Davenport JL, Potter MC. Scene consistency in object and background perception. Psychological Science. 2004; 15:559–564. [PubMed: 15271002]

Henderson JM. Human gaze control during real-world scene perception. Trends in Cognitive Sciences. 2003; 7:498–504. [PubMed: 14585447]

Henderson JM, Hollingworth A. The role of fixation position in detecting scene changes across saccades. Psychological Science. 1999; 5:438–443.

Henderson JM, Pierce GL. Eye movements during scene viewing: Evidence for mixed control of fixation durations. Psychonomic Bulletin & Review. 2008; 15:566–573. [PubMed: 18567256]

Ishida T, Ikeda M. Temporal properties of information extraction in reading studied by a text-mask replacement technique. Journal of the Optical Society A: Optics and Image Science. 1989; 6:1624–1632.

Li FF, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene. Journal of Vision. 2007; 7:1–29.

[5]Another factor may be that information is acquired from a wider region in scene perception than in reading (Rayner, in press).

Liversedge SP, Rayner K, White SJ, Vergilino-Perez D, Findlay JM, Kentridge RW. Eye movements when reading disappearing text: Is there a gap effect in reading? Vision Research. 2004; 44:1013–1024. [PubMed: 15031094]

McConkie GW, Rayner K. The span of the effective stimulus during a fixation in reading. Perception & Psychophysics. 1975; 17:578–586.

Najemnik J, Geisler WS. Optimal eye movement strategies in visual search. Nature. 2005; 434:387–391. [PubMed: 15772663]

Potter MC. Meaning in visual search. Science. 1975; 187:965–966. [PubMed: 1145183]

Rayner K. The perceptual span and peripheral cues in reading. Cognitive Psychology. 1975; 7:65–81.

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124:372–422. [PubMed: 9849112]

Rayner K. Eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology. in press.

Rayner K, Bertera JH. Reading without a fovea. Science. 1979; 206:468–469. [PubMed: 504987]

Rayner K, Inhoff AW, Morrison RE, Slowiaczek ML, Bertera JH. Masking of foveal and parafoveal vision during eye fixations in reading. Journal of Experimental Psychology: Human Perception and Performance. 1981; 7:167–179. [PubMed: 6452494]

Rayner K, Li X, Williams CC, Cave KR, Well AD. Eye movements during information processing tasks: Individual differences and cultural effects. Vision Research. 2007; 50:2714–2726. [PubMed: 17614113]

Rayner K, Liversedge SP, White SJ. Eye movements when reading disappearing text: The importance of the word to the right of fixation. Vision Research. 2006; 46:310–323. [PubMed: 16085229]

Rayner K, Liversedge SP, White SJ, Vergilino-Perez D. Reading disappearing text: Cognitive control of eye movements. Psychological Science. 2003; 14:385–388. [PubMed: 12807416]

Reingold EM, Stampe DM. Saccadic inhibition in voluntary and reflexive saccades. Journal of Cognitive Neuroscience. 2002; 14:371–388. [PubMed: 11970798]

Rousselet GA, Joubert OR, Fabre-Thorpe M. How long to get to the “gist” of real-world natural scenes? Visual Cognition. 2005; 12:852–877.

Rousselet GA, Macé MJ, Fabre-Thorpe M. Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. Journal of Vision. 2003; 3:440–455. [PubMed: 12901715]

Schyns P, Oliva A. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychological Science. 1994; 5:195–200.

Thorpe SJ, Fize D, Marlot C. Speed of processing in the human visual system. Nature. 1996; 381:520–522. [PubMed: 8632824]

van Diepen PMJ, Ruelens L, d’Ydewalle G. Brief foveal masking during scene perception. Acta Psychologica. 1999; 101:91–103. [PubMed: 10100455]

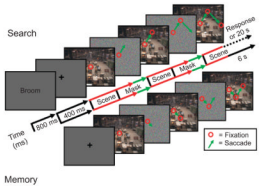

Fig. 1.

The sequence of events on a trial. In the search task, the name of a target object appeared for 800 ms, followed by a fixation cross. Participants fixated on the cross, which remained in view for 400 ms, and then the scene appeared. At the designated mask onset, the mask appeared; it remained present until the beginning of a new fixation. The mask then reappeared at the designated mask onset. This sequence continued until either the participant made a response or 20 s had elapsed. In the memory task, the sequence started with the fixation cross, but the sequence thereafter was the same as in the search task. However, the trial ended after 6 s in the memory task.

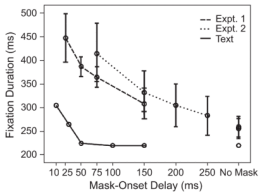

Fig. 2.

Fixation duration as a function of mask onset in Experiments 1 and 2, and in the full-line masking condition from Rayner, Inhoff, Morrison, Slowiaczek, and Bertera (1981). Error bars represent the 95% confidence intervals for the data from Experiments 1 and 2.

TABLE 1

Mean Search Time, Fixation Duration, Saccade Length, and Search Accuracy as a Function of Mask-Onset Delay in Experiment 1

	Mask-onset delay (ms)				
Variable	**25**	**50**	**75**	**150**	**No mask**
Search time (s)	10.3	7.8	6.4	4.8	3.8
Fixation duration (ms)	447	387	364	308	256
Saccade length (deg)	3.6	3.6	3.6	3.9	4.3
Search accuracy	.30	.56	.74	.91	.99

Note. Search accuracy is the probability of correctly identifying the location of the target within the scene.

TABLE 2

Mean Search Time, Fixation Duration, Saccade Length, and Search and Recognition Accuracy as a Function of Mask-Onset Delay in Experiment 2

	Mask-onset delay (ms)				
Variable	**75**	**150**	**200**	**250**	**No mask**
Search time (s)	6.5	4.9	4.7	4.5	4.5
Fixation duration (ms)	414	332	305	284	260
Saccade length (deg)	3.4	3.8	4.0	4.4	4.5
Search accuracy	.61	.90	.95	.93	.95
Recognition accuracy	.72	.82	.83	.87	.87

Note. Search accuracy indicates the probability of correctly identifying the location of the target. Recognition accuracy indicates performance on a recognition memory task in which viewers had to indicate whether a given scene was old or new.