Eye Movements and Visual Encoding during Scene Perception

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Psychological Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>PSCI-08-0193.R2</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Research report</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Rayner, Keith; University of Massachusetts, Psychology Department Smith, Tim; University of Edinburgh, Psychology Malcolm, George; University of Edinburgh, Psychology Henderson, John; University of Edinburgh</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Eye Movements, Perception, Visual Attention</td>
</tr>
</tbody>
</table>
Eye Movements and Visual Encoding during Scene Perception

Keith Rayner¹, Tim J. Smith², George Malcolm², and John M. Henderson²

¹ Department of Psychology, University of California, San Diego, USA
² Department of Psychology, University of Edinburgh, Scotland

Correspondence:

Keith Rayner
Department of Psychology
University of California, San Diego
858-822-7816
Email: krayner@ucsd.edu

Word Count = 2499 words (text, notes, and acknowledgments)

Running Head: Scene encoding
Abstract

The amount of time viewers could process a scene during eye fixations was varied as a mask appeared at a certain point in each eye fixation. The scene didn’t 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 150ms during each eye fixation. This result is surprising since viewers can extract the gist of a scene from a brief 40-100ms exposure. It also stands in marked contrast to reading where readers need only to view the words in the text for 50-60ms to read normally. Thus, although the same neural mechanisms control eye movements in scene perception and reading, the cognitive processes associated with each task drives processing in different ways.
The neural mechanisms that underlie oculomotor activity do not vary as a function of the task viewers engage in; there isn't one oculomotor system for looking at scenes, another for visual search, and another for reading. Eye movements are essential in these tasks since 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, 2008). 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 50-60ms 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, Rayner, White, Vergilino-Perez, Findlay, & Kentridge, 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 40-100ms (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 50-60ms. We explicitly tested this hypothesis by masking scenes 25, 50, 75, 150, 200, and 250ms after the beginning of each fixation.
Experiment 1

In Experiment 1, the task was to find a pre-specified target object in a scene. Thus, for example, in a warehouse scene viewers were asked to locate a broom. Eye movements were 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 naïve concerning the purpose of the experiment.

Experimental Apparatus and Procedure. Eye movements were monitored via a SR Eyelink1000 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/second velocity threshold using a 9-sample saccade detection model. Viewing was binocular, but only the right eye was tracked. The images were presented on a 21 inch cathode ray tube monitor at a viewing distance of 90 cm with a refresh rate of 140Hz. 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 Figure 1), a target word was presented for 800ms, followed by a fixation cross for 400ms. Then the scene was presented. Presentation of the scene was interrupted after a pre-defined viewing time (25, 50, 75, and 150ms) 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 20sec 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 x 600 pixel photographs of real-world scenes\(^1\) from a variety of scene categories were used in the experiment.

**Results**

Figure 1 and Table 1 here

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

**Discussion**

Unlike reading (Rayner et al., 1981, 2003) where viewing text for 50-60ms prior to mask onset seems to be sufficient for reading to proceed effectively, viewers needed
much longer than this to effectively encode the scene. Indeed, even with the 150ms mask onset performance did not reach the level of the no-mask control condition. In order 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 if 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 250ms. Half of the viewers were again asked to search for a specific target item in the scene (Search task), while the other half examined each scene in anticipation of a recognition memory test given at the end of the experiment (Memory task).

**Participants.** Twenty naïve 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. Whereas the scene remained until either a response occurred or 20ms elapsed in the Search task, in the Memory task the scene was only presented for 6sec (see Figure 1). The mask onset delays were 75, 150, 200, and 250ms, and a control condition was again included in which a mask did not appear on each fixation. Following 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 0 (no confidence) to 3 (full confidence) scale.

Confidence ratings were uninformative and therefore will not be presented.

**Materials.** These were identical to Experiment 1 except for the addition of the 60 new scenes in the Memory task.

**Results**

While 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 will 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.

Table 2 here

As in Experiment 1, ANOVAs on the measures in Table 2 yielded significant effects of mask onset: 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 75ms mask condition (search time: p<.05, Prep=.99, d=0.99; search accuracy: p<.001, Prep=.99, d=-2.02). Accuracy on the recognition memory test was only significantly worse than the Control for the 75ms mask condition (p<.01, Prep=.94, d=-1.18). Over both tasks, all mean fixation durations and saccade lengths differed significantly from the
Control (ps <.01, $Prep \geq .90$) except for the 250ms mask condition (fixation duration, $p=.053$, $Prep=.76$, $d=0.34$; saccade length, $p=.138$, $Prep=.70$, $d=-0.26$).

**Discussion**

A number of results from Experiment 2 are very striking. First, as in Experiment 1, the 75ms 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 50ms mask delays also 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 75ms delay. Thus, it would seem that 150ms 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. While 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 150ms, this was not the case for either saccade length or fixation duration. For saccade length, performance reached asymptote at 250ms. For fixation duration, there was a steady decrease in fixation duration with each level of mask delay from 150ms to 250ms.
which was on the order of 25ms for each 50ms increase in the mask onset. Likewise, there was a 24ms decrease in fixation duration from the 250ms mask onset condition to the control condition. We suspect that the reason for the differences is due to 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 150ms to encode stimulus properties during eye fixations in scene perception. This finding indicates that the 40-100ms needed to acquire sufficient information to understand scene gist 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 wherein text is masked during reading which have demonstrated that readers need only 50-60ms 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 the most comparable reading study (Rayner et al., 1981).

Insert Figure 2 about here

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 prior to the mask? First, perhaps it takes a longer presentation time than 50 ms to encode the general meaning of the scene. However, as already noted, the gist can be understood from 40-50 ms scene exposure. Second, perhaps it takes more presentation time to encode fixated objects in scenes than words in text. Contrary to this hypothesis, studies have shown that objects can be encoded from very brief presentations (on the order of 50 ms), even when
the object appears in a scene (Davenport & Potter, 2004; Li, Iyer, Koch, & Perona, 2007; Rousellet et al., 2003; Thorpe & 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 a 50ms of scene presentation⁵. 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. While 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 encoding of words in reading.
References


Liversedge, S.P., Rayner, K., White, S.J., Vergilino-Perez, D., Findlay, J.M., &


Acknowledgments

This research was supported by grants HD26765 from the National Institute of Health (USA) and RES-062-23-1092 from the Economic and Social Sciences Research Council (UK). We thank Geoff Loftus and Marvin Chun for excellent comments.
Footnotes

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

2 The accuracy measures are shown separately for the Search and Memory tasks as it is not appropriate to collapse over them since 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 where viewers had to indicate if 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 since all scenes were presented for the same duration during the memory encoding period.

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 45-75 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 while here the entire scene was masked. Thus, viewers might be less certain about where to move next in the scene experiments than the reading experiments. However, saccade size was fairly large in even the 50 and 75ms onset conditions. Also, Rayner et al. (1981) included a condition in which the entire line was masked and it was still the case that 50ms was sufficient for reading to proceed normally (see Figure 2).
Another factor may be that information is acquired from a wider region in scene perception than reading (Rayner, 2008).
Table 1. Mean search time, fixation duration, saccade length, and search accuracy as a function of mask onset delay.

<table>
<thead>
<tr>
<th>Mask Onset Delay</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>150</th>
<th>No Mask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search Time (sec)</td>
<td>10.3</td>
<td>7.8</td>
<td>6.4</td>
<td>4.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Fixation Duration (ms)</td>
<td>447</td>
<td>387</td>
<td>364</td>
<td>308</td>
<td>256</td>
</tr>
<tr>
<td>Saccade Length (deg)</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
<td>3.9</td>
<td>4.3</td>
</tr>
<tr>
<td>Search Accuracy</td>
<td>.30</td>
<td>.56</td>
<td>.74</td>
<td>.91</td>
<td>.99</td>
</tr>
</tbody>
</table>
Table 2. Mean search time, fixation duration, and saccade length as a function of mask onset delay. Accuracy in finding the target in the Search task and recognition memory accuracy in the Memory task is also shown.
Figure Caption

Figure 1: The sequence of events on a trial in the experiments. In the Search task, the name of a target object appeared for 800 ms, followed by a fixation cross. Subjects fixated on the fixation cross which remained in view for 400 ms and then the scene appeared. At the designated mask onset, the mask appeared and remained present until the beginning of a new fixation, with the mask again appearing at the designated mask onset. This sequence continued until either the subject made a response or 20 sec 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 sec in the Memory task.

Figure 2: Fixation durations as a function of mask onset in Experiments 1 and 2, and the full-line masking condition from Rayner et al. (1981). Error bars represent the 95% Confidence Intervals for the data from Experiments 1 and 2.