PERCEIVING THE ROADWAY IN THE BLINK OF AN EYE – RAPID PERCEPTION OF THE ROAD ENVIRONMENT AND PREDICTION OF EVENTS

Benjamin Wolfe[1,2], Lex Fridman[1], Anna Kosovicheva[3], Bobbie Seppelt[1,4], Bruce Mehler[1], Ruth Rosenholtz[2,5], Bryan Reimer[1]

[1]AgeLab, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Department of Psychology, Northeastern University, Boston, MA, USA
[4]Touchstone Evaluations Inc., Detroit, MI, USA
[5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Summary: This study investigated how quickly participants could develop a functional mental representation of a real-world road scene, based on briefly viewed recorded video. Using Amazon Mechanical Turk, we recruited 27 participants and collected 25k individual trials assessing the development of a percept of the road environment. This was operationalized as the duration of road video required for participants to predict which of two temporally spaced images would happen next. We found that participants could begin to build a representation of the road environment with as little as 100 ms of viewed road video and that the representation improved with additional video. These results suggest that drivers may begin to construct robust, predictive mental representations of the road environment with the briefest of glances, and the more information available to them, the more robust these representations are. While 100 ms of eyes-on-road time is insufficient to ensure safe driving, comprehension of the road environment begins in the blink of an eye.

INTRODUCTION

How quickly can a driver glean task-relevant information from the road ahead? To frame this in the context of on and off-road glances, NHTSA (2013) guidelines suggest that a driver should not remove their gaze from the forward roadway for more than two seconds, and it seems likely that several seconds of viewing time are necessary to fully comprehend the road environment. While it is undoubtedly safer for a driver to keep their eyes on the road, how long does it take to develop a simple but useful perception of the road environment? This is a fundamental question critical to the successful deployment of technologies aiming to assist drivers in the management of their attention to the roadway. In order to understand how a driver becomes ready to respond to changing events, we must first understand how they build the foundation for subsequent action.

Our interest grew out of previous basic research in vision science, particularly in the area of scene gist, the ability to comprehend the basics of a natural scene in the blink of an eye (Oliva, 2005; Oliva & Torralba, 2006). Research in this area has shown observers to be capable of making complex determinations about natural scenes with very brief exposure to static images; e.g., the category that a scene belongs to (Greene & Oliva, 2009b), whether the scene is navigable (Greene & Oliva, 2009a) in less than 100 ms of exposure to a static image.

We extend this basic vision science work on perception of static scenes to dynamic driving environments. In an experiment run on Mechanical Turk, we asked participants to watch brief clips of forward-facing road video recorded in and around Boston, MA and then to report which of two subsequently shown still images was most likely to follow what they had been shown. This design allowed us to focus exclusively on the perceptual element of this question, rather than on the subsequent actions a driver might need to perform. Knowing that another vehicle might be about to swerve into the lane of travel is useful, but the driver must first perceive the basics of the road environment. Based on previous work with static scenes, we hypothesized that participants would be able to develop a sufficient percept of the driving scene to begin to build a foundational understanding of events with very brief video durations.

METHODS

Participants

Participants were recruited through Amazon Mechanical Turk, and provided informed consent to participate (as per review by MIT's Committee on the Use of Humans as Experimental Participants). Data was collected from a total of 31 participants; data from 27 of these was retained in the final analysis. Data from four participants was removed due to these participants' data failing a binomial test (e.g., overall performance being no better than chance). Age and gender data were not recorded, per MIT policy for Mechanical Turk experiments. Each participant performed 898 trials across all conditions and trial types. Participants were compensated at a rate of 1¢ per completed trial, with a $10 bonus for completion of the total set of 898 trials. The experiment took approximately one hour, if done in a continuous session.

Stimuli

Stimuli were extracted from 16 four-minute segments of forward-facing road video, recorded in and around Boston, MA. Expert observers classified these four minute videos as either urban/suburban or highway driving. The video resolution was 720p (1080x720) at 29.97 frames per second. There were an equal number of segments classified as urban/suburban or highway driving. These video segments were subsequently divided into shorter clips for use in individual trials (Figure 1). Video clips for the viewing phase of the trial ranged from 100 ms to 4000 ms duration (100, 233, 500, 1000, 2000 and 4000 ms durations were used). In addition, for the testing phase of the trial we extracted still frames from after the end of each clip, with a minimum of 500 ms between the end of the clip and the first still frame. Two still frames were extracted for every trial, with the two frames temporally separated by 100, 233, 500, 733, 1000, 2000, 3000 or 4000 ms. On any given trial, participants viewed a single clip, and a single pair of still frames.

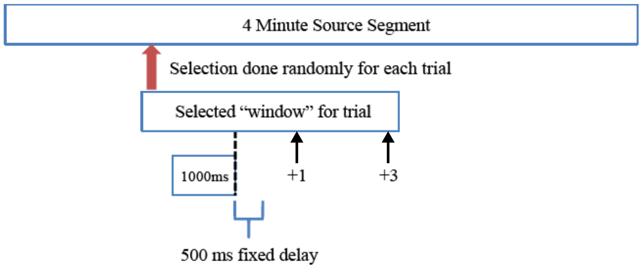
4 Minute Source Segment
Selection done randomly for each trial
Selected "window" for trial
1000ms
+1
+3
500 ms fixed delay

Figure 1: Visualization of stimulus extraction for an individual trial from a longer source segment. Given a start point, randomly selected for each trial, we then extracted a stimulus clip for the viewing phase of the trial, as well as two still frames following that clip. In this example, the stimulus clip is 1000 ms long, and the two still frames (solid black arrows) are from +1 sec (1000 ms) and +3 sec (3000 ms) after the end of the stimulus clip (dotted vertical line), which gives the still frames a separation of 2000 ms.

Trial Types

The study consisted of three conditions, which were randomly interleaved during the experiment. In all cases, participants' task was to view two still images and to click on the one they believed to be correct, depending on the instruction provided.

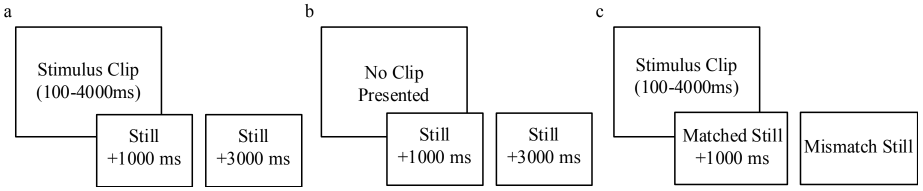

Figure 2: Schematic of conditions. (a) *Scene perception* condition, consisting of a stimulus clip and two matched but temporally separated stills. (b) *No-Video* condition, where no clip was presented. (c) *Attentional Catch* condition, where one of the two stills did not match the clip.

The core of the study was the *scene perception* condition (Figure 2a), where participants viewed a short clip (100 to 4000 ms), and were subsequently shown two still frames from video following that clip (temporally separated by 100 to 4000 ms). In this condition, participants were asked "Which of these images would come first after the video you watched?" and were instructed to click on the image which they believed best answered the question. This condition tested the core hypothesis that participants would be able to perceive the road environment and make predictions about subsequent events, based on brief views of the road scene.

Conceivably, participants could, to some extent, guess which video came first simply by looking at the two test frames, without even looking at the video clip for context. To quantify this effect, we included a *no-video* condition (Figure 2b); participants were shown two still images, separated by 100 to 4000 ms, sourced identically as in the scene perception condition, but without any video. Participants were asked to identify which still image came first, as in the scene perception condition, but without the preceding video. Participants' ability to perform this task, compared to the scene perception condition, allows us to distinguish between their use of information in the still images and information in the video.

We also included an *attentional catch* condition (Figure 2c) as part of our automated data quality procedure for Mechanical Turk to ensure that participants were watching the videos and performing the task as instructed in the absence of experimenter monitoring. In this condition, participants were shown a short clip (100 to 4000 ms), and then shown two still images. One still image came from the same video sequence as the clip, 1000 ms after the end of that clip. The other frame was a total mismatch (e.g., a drive along a beach when no such scene existed in our stimulus set; the mismatched images were hand-selected to be easily discriminable from the correct answer in this condition). Participants were again asked “which of these images would come first after the video you just watched.” These trials allowed us to determine whether participants were attending to the task throughout the duration of the experiment. Participants who failed more than two catch trials throughout the study were automatically ejected from the study and paid for the trials they had completed to that point. No participants whose data was included in the final sample missed more than one catch trial in the entire experiment.

Analysis

We performed two main analyses. The first examining performance directly, and the second examining how performance changed as a function of still frame separation. In the initial analysis, we performed a 2 (road type) × 8 (still-frame separation) x 7 (clip duration) repeated-measures ANOVA on participants' accuracy (% correct). To understand how still frame separation changed performance for different clip durations, we applied a linear fit within each clip duration, and subsequently performed a 2 (road type) x 7 (duration) repeated-measures ANOVA on the fitted slopes to determine whether these slopes changed as a function of clip duration and road type. As still frame separation increases, the still frames become more discriminable and the participants’ task becomes easier.

RESULTS

In the analysis of participants’ performance in the scene discrimination condition, we found a main effect of road type; *F*(1,26)=22.86, *p*<.001 (Figure 3). Participants are better at the task with urban/suburban road scenes than with highway scenes. This may be attributable to increased visual density in the urban/suburban scenes than the highway scenes, facilitating correct discrimination and therefore increased performance on the task.

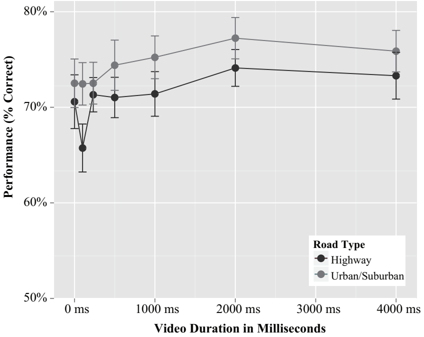

Figure 3: Performance (percent correct responses) by Clip Duration and Road Type. Error bars this and subsequent figures represent the standard error of the mean. Note the 0 ms label corresponds to the No Video condition.

There was also a main effect of clip duration, *F*(6,156)=7.17, *p*<.001, indicating an improvement in performance with longer views of the road environment, as well as a main effect of still frame separation, F(7,182)= 23.74, *p*<.001, indicating that temporally separated still frames were more easily discriminated by participants. There were significant 2-way interactions between road type and still frame separation *F*(7,182)=6.21, *p*<.001, clip duration and still frame separation, *F*(42,1092)=4.002, *p*<.001, but not between road type and clip duration (*p*=.219). A significant 3-way interaction appeared between clip duration, still frame separation and road environment, *F*(42,1092)=4.53, *p*<.001. However, after performing pairwise comparisons within clip duration and across still frame separation (using a corrected α = 0.0003), we only found significant differences in performance for longer clips (2000 ms or longer) and longer separations (beginning at 1000 ms separation).

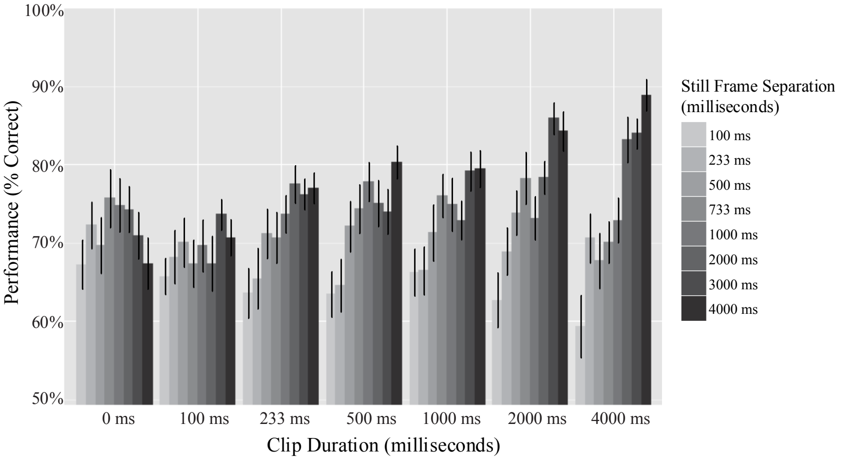

Figure 4: Performance (percent correct responses) by Clip Duration and Still Frame Separation. Each grouping on the X-Axis is a given clip duration; bars within a group are different separations between the two test stills. Note the change in slope within each grouping, moving from shorter to longer separations.

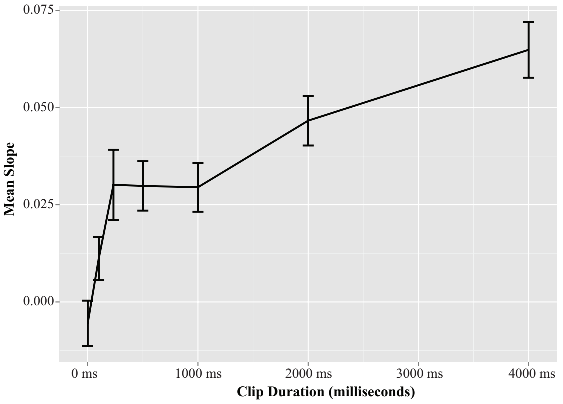

Figure 5: Mean slope (change in performance within trials for a given clip duration) as a function of clip duration. Note that the only duration condition to not exhibit a positive slope is the 0 ms (or no video) condition, suggesting that even very little information facilitates discrimination of the still frames, and the increase in mean slope with increasing clip duration indicates that additional information facilities this.

To understand, within each clip duration, how participants' ability to discriminate the test frames changed, we fit performance within each clip duration to a line, and then estimated the slope. This measure, critically, allows us to understand how the difficulty of the discrimination task changed when more information was available. Slopes at or near zero would indicate that increases in frame separation were having no meaningful effect on task performance, whereas positive slopes would indicate that increases in separation were making the task easier. In this analysis, we found no significant effect of road type, *F*(1,26)=0.115, *p*=.738, however, we found a main effect of clip duration, *F*(6,156)=17.47, *p*<.001, and an interaction between clip duration and road type, *F*(6,156)=5.755, *p*<.001. Notably, we find slopes significantly above zero in all duration conditions, including the 100 ms condition, with the exception of the no-video trials (indicated by the 0 ms duration on the plots in Figures 4 and 5). Given the positive slopes, increasing still frame separation does increase discriminability, but clip duration is also a factor in performance, suggesting that additional information, from longer clips, has a significant influence on performance.

DISCUSSION

Building on the scene gist literature discussed in the introduction, which has shown that observers can classify static images of natural scenes when those scenes are only shown for 50 ms, we show that as little as 100 ms of driving video is sufficient to facilitate discrimination of temporally subsequent still frames. Making this discrimination is inherently a prediction task, insomuch as we were asking participants to indicate which still frame would follow the clip they had viewed. In order to do this, participants needed to develop a mental representation of the scene they had just viewed, and be able to use that mental representation to predict what would happen next in the scene.

To some degree, this is facilitated by the road environment being comparatively predictable. One may be able to guess which frame occurs first simply through understanding the statistics of road scenes, much as in a child's puzzle the picture with the half-eaten apple comes after the picture with a whole apple. With this in mind, we included the no-video trials as a control. Had participants been able to perform the task at a high level of accuracy in the no-video trials, with increasing accuracy as a function of still frame separation, this would have indicated that the clips did not provide additional information. In fact, performance on the no-video trials did not change appreciably as a function of still frame separation, suggesting that increased separation did not add additional information to support accurate discrimination and prediction. In addition, performance increased as a function of clip duration, indicating that the more information participants were provided (as conveyed in longer clips, from which the participant would be able to construct a more robust representation of the scene), the better they were able to discriminate the earlier from the later still frame following the video, and thereby better perform the task. Together, these results indicate that participants used the additional information provided by the clips, rather than just relying on cues in the still frames.

While the brevity of the video required to accurately identify the correct temporal ordering of events is surprising, it should not be taken to be at odds with current best practice for safe driving. We are not, in any way, suggesting that drivers only need a tenth of a second glimpse of the forward roadway to be safe on the road. What our results show is that drivers start to build a useful representation of the road environment with very little visual information. However, our results also indicate that this representation gets progressively more detailed with increased viewing duration, as evidenced by the increasing slopes. However, this approach has certain limitations; participants performed the task in isolation, rather than while operating a vehicle, and were simply asked to report their predictions, rather than to react to a hazard in the road environment. That said, this work advances our theoretical understanding of drivers' perception of the world, and these limitations allowed us to focus on the core perceptual question, rather than the complexities of the operational environment. In summary, the idea that keeping the driver's eyes on the forward roadway is better for their understanding of the road environment is scarcely novel, but how quickly this process begins, and how much information is available even in brief glances is.

Asking what the driver can perceive might, at first glance, be an overly theoretical question, because the essential focus of much of the study of driving and driver behavior is on drivers' actions, rather than their internal representation of the road environment. However, in order to understand how the driver can act, and why the driver did or did not act in an expected manner, we need to understand how quickly they can begin to perceive the road environment, and what it takes for them to start to predict what might occur in the seconds to come. Information in this context may be critical in the development in systems that actively assist drivers in managing their attention, in short, how predictive a sensing system must be to alert drivers of a threat far enough in advance such that they can perceive and respond to an event or not. Our experiment, as described in this paper, is a step towards understanding the visual foundations that underlie safe driving. We found that even the barest glimpse of the road is useful, but that more information is always preferred over less.

ACKNOWLEDGEMENTS

Support for various aspects of this work was provided by the Advanced Human Factors Evaluator for Automotive Demand (AHEAD) Consortium, the US DOT's Region I New England University Transportation Center at MIT, Toyota Research Institute / MIT CSAIL joint partnership, and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs' class counsel.

REFERENCES

Greene, M. R., & Oliva, A. (2009a). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. http://doi.org/10.1016/j.cogpsych.2008.06.001

Greene, M. R., & Oliva, A. (2009b). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science*, *20*(4), 464–472. http://doi.org/10.1111/j.1467-9280.2009.02316.x

National Highway Traffic Safety Administration (2013). *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (Docket No. NHTSA-2010-0053)*. U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA), Washington, DC.

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, 251–256.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. http://doi.org/10.1016/S0079-6123(06)55002-2