Predicting road scenes from brief views of driving video

Benjamin Wolfe Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

Lex Fridman AgeLab, Massachusetts Institute of Technology, Cambridge, MA, USA

Anna Kosovicheva Department of Psychology, Northeastern University, Boston, MA, USA

Bobbie Seppelt AgeLab, Massachusetts Institute of Technology, Cambridge, MA, USA

Bruce Mehler AgeLab, Massachusetts Institute of Technology, Cambridge, MA, USA

Bryan Reimer AgeLab, Massachusetts Institute of Technology, Cambridge, MA, USA

Ruth Rosenholtz Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA; Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

If a vehicle is driving itself and asks the driver to take over, how much time does the driver need to comprehend the scene and respond appropriately? Previous work on natural-scene perception suggests that observers quickly acquire the gist, but gist-level understanding may not be sufficient to enable action. The moving road environment cannot be studied with static images alone, and safe driving requires anticipating future events. We performed two experiments to examine how quickly subjects could perceive the road scenes they viewed and make predictions based on their mental representations of the scenes. In both experiments, subjects performed a temporal-order prediction task, in which they viewed brief segments of road video and indicated which of two still frames would come next after the end of the video. By varying the duration of the previewed video clip, we determined the viewing duration required for accurate prediction of recorded road scenes. We performed an initial experiment on Mechanical Turk to explore the space, and a follow-up experiment in the lab to address questions of road type and stimulus discriminability. Our results suggest that representations which enable prediction can be developed from brief views of a road scene, and that different road environments (e.g., city versus highway driving) have a significant impact on the viewing durations drivers require to make accurate predictions of upcoming scenes.

Introduction

Drivers must generate an accurate representation of their environment in order to anticipate the future locations of other vehicles, cyclists, and pedestrians. In considering the relationship between perception and prediction, one must ask: How much visual information does a driver need to accurately predict future road scenes? In driving research, this question is usually framed in the context of the driver's need to develop

Citation: Wolfe, B., Fridman, L., Kosovicheva, A., Seppelt, B., Mehler, B., Reimer, B., & Rosenholtz, R. (2019). Predicting road scenes from brief views of driving video. *Journal of Vision*, *19*(5):8, 1–14, https://doi.org/10.1167/19.5.8.

https://doi.org/10.1167/19.5.8 Received September 26, 2018; published May 7, 2019 ISSN 1534-7362

situation awareness (Endsley, 1995), a cognitive model reliant on the driver's ability to perceive and comprehend their environment prior to prediction. Such awareness is frequently conceived as being built up over periods of seconds, and sometimes minutes, as the driver develops a detailed mental representation of key elements and interactions from a dynamically changing scene. In contrast, work in vision science has shown that subjects can perceive the essence, or gist, of static natural scenes in less than 100 ms (Greene & Oliva, 2009), although a gist-level representation may be insufficient for predicting a moving scene or operating a vehicle safely. Probing the differences between these accounts may be critical for understanding how quickly drivers can assess situations and make decisions. This work is particularly timely, given the rapidly expanding deployment of automated vehicle systems that, at variable frequency, require safe exchange of control between vehicle and driver under time pressure. Unlike drivers of largely manually controlled vehicles, drivers of more highly automated vehicles may not need (or be willing) to maintain a continuous representation of the environment while the vehicle is in motion. If we are to have automated vehicles where a handoff may be required, whether these handoffs are planned or not, we must understand how long the driver will need to perceive and predict the world after a period during which they have not been paying attention.

To put our work and approach in context, we will first discuss how the driver's representation of the environment has been considered in driving research. According to Endsley (1995), a driver's situation awareness refers to their understanding of the operating environment (the road and surroundings) and their ability to respond to changes in that environment while maintaining control over their vehicle. In this definition, situation awareness develops via a three-stage process, beginning with perception of visual elements in the scene, which are then processed in relation to the individual's goals to enable prediction of future events in the environment. The driver is thought to then be able to act on these predictions, based on their awareness of their operating environment in relation to their goal state. However, Endsley also discusses evidence that experts, as a consequence of their expertise, recognize the correct action relative to their goal state without deliberation. That said, Endsley's theory does not describe how this process might work, or what timescale it occurs on, but merely suggests that these stages might be a way to consider the problem.

In contrast, vision-science research on scene perception indicates that extracting the gist of a scene is a fast process (Larson & Loschky, 2009; Oliva, 2005; Oliva & Torralba, 2006). The gist, sometimes conceived of as the one or two sentences one would use to describe the scene, is often operationalized as the information available in a single glance. It consists of often-holistic properties of the scene, such as the broad category to which it belongs or the degree to which it can be readily navigated. Subjects are able to perceive the gist of a scene with less than 100 ms of viewing time (Greene & Oliva, 2009). However, while gist perception is fast, it is often probed with broad questions that are less applicable to driving—for example, "Is this scene navigable at all?" or "Is this a city or a highway scene?" Therefore, a gist-level understanding of the driving environment may not be enough to enable safe driving; the driver needs to know more than whether they are on a highway or an urban road. For that matter, this work has been done with static images of natural scenes rather than video, and it is unknown how these results will extend to dynamic scenes. We would argue that the ability to extract the gist of a scene on the timescale previously reported (e.g., in 100 ms or less) suggests that situation awareness is, to some degree, reliant on these fast processes. A more complete answer may be found between the time courses suggested, respectively, by the gist and situation-awareness literatures.

Something of a middle ground exists between these two views of visual perception and prediction when we look to the hazard-perception literature, where subjects are asked to view still images or videos of driving scenes and either assess the relative hazard (Pelz & Krupat, 1974) or report the presence of hazards in the stimulus (McKenna & Crick, 1994). In brief, this body of work suggests that drivers learn to search for hazards (Underwood, Crundall, & Chapman, 2002) in particular locations in the scene—for example, where they are more likely to be present (Mackenzie & Harris, 2015; Underwood, Phelps, & Wright, 2005) and where they may be occluded by other objects (Alberti, Shahar, & Crundall, 2014). Given the role of prior knowledge here, there are strong similarities to the idea of scene grammar (Draschkow & Võ, 2017) in visual search, where what belongs (or not) in the scene influences search speed and accuracy. However, hazard-perception studies explicitly probe the perception and prediction of only one visual category (road hazards), not the driver's visual input and representation more broadly.

Whereas classifying static scenes, as used in scene-gist studies, requires very little time, predicting changes as part of the driver's situation awareness is likely a slower process. That said, probing a driver's representation of their environment, the foundations of their situation awareness, is a difficult problem; the researcher must consider both the stimuli and the task used to probe questions of representation. Prior work on this question (Lu, Coster, & de Winter, 2017) has used simulated environments to maximize control over the stimuli, but given subjects tasks that were quite dissimilar from those drivers would encounter on the

road (e.g., asking about where other cars were on the road and their relative speed). Drawing closely from prior work in scene-gist research would also limit the relevance to natural tasks, as examining representation for prediction in driving requires different questions.

In order to probe subjects' representations of the road environment in a more realistic way, we developed a temporal-order prediction task based on video taken from drives around the Boston area. We asked our subjects to watch a brief segment of road video and to indicate which one of two subsequently presented still images would come next (similar to the What Happens Next task described by Jackson, Chapman, & Crundall, 2009, but asking subjects to choose between two images). This task required subjects to develop a sufficiently robust representation of the scene and to make general predictions of how the scene would change overall in the immediate future, distinguishing this from other, less temporally proximate images. This allowed us to probe subjects' representations at a broad level, without focusing on a single element or feature in the environment, as well as to probe both perception and prediction in a single task.

Experiment 1: Effects of duration and road type on prediction of subsequent events

To gain an understanding of drivers' ability to make predictions about future events on the roadway, we ran an experiment using short (100–4,000 ms) clips of forward-facing road video on Amazon Mechanical Turk and asked subjects to judge, using two still images from the source video, which image they believed would come next after the video they had seen. In contrast to previous efforts in driving research at probing representations of the driving environment, this design avoided explicit reliance on specific features in the scene (e.g., vehicle color, position, or speed) and tested drivers' overall ability to perceive the scene, similar to some tasks in the representational-momentum literature (see Blättler, Ferrari, Didierjean, & Marmèche, 2012). By manipulating both the duration of the clip and the temporal separation of the two test images, we probed the duration required for subjects to develop a sufficiently robust representation as a function of task difficulty. We also predicted that performance on the task would vary as a function of the contents of the scene, and therefore tested two road environments: highway and urban settings. We note that Experiment 1 was previously presented as a conference paper at Driving Assessment 2017 (Wolfe, Fridman, et al., 2017).

Participants

Subjects were recruited through Amazon Mechanical Turk, and provided informed consent prior to participating in the experiment (as required by MIT's Committee on the Use of Humans as Experimental Participants, and in accordance with the Common Rule). Subject age and gender were not recorded, in accordance with institute policy for Mechanical Turk experiments. Subjects were compensated for their time at a rate of 1 cent per completed trial, with a bonus of $10 for completion of the entire set of trials (and a total compensation of $20 for approximately 1 hr of testing). A total of 31 subjects completed the experiment, and data from 27 were retained in the final analysis. Three subjects were excluded from the final analysis because their overall performance on the task was no better than chance (binomial test, $p > 0.05$). A fourth subject was excluded from the analysis due to an error that allowed them to complete the task twice.

Stimuli

Videos were taken from random time points within 16 4-min segments of forward-facing road video, recorded from a centrally mounted camera inside a vehicle that was driven on both highways and surface roadways near Boston. Prior to the experiment, two expert observers classified these 16 videos into two road-type conditions—urban or highway—resulting in eight videos per condition. Urban videos typically had speeds below 40 mph and were recorded in urban or suburban settings, and highway videos consisted of multilane, high-speed driving. All videos contained uneventful, mundane driving, with no proximate hazards (e.g., unsignaled lane changes, near-collision events, collisions), in order to focus on subjects' ability to predict the scene as a whole rather than respond to immediate hazards. Our criteria were intentionally broad, controlling road type (urban vs. highway) and allowing more granular elements (e.g., number of lanes, traffic level) to vary, since the goal of this work was to probe representation and prediction broadly. Videos were recorded and presented at 720p (1,280 × 720) resolution at 29.97 frames/s.

For the experiment, the 4-min video segments were divided into shorter clips that were shown to subjects in individual trials (Figure 1a). Video clips for the preview stage of each trial had one of seven possible preview durations—0 (no video), 100, 233, 500, 1,000, 2,000, and 4,000 ms—selected from random time points within the clip. In addition, we extracted two still frames for the response screen on each trial. The first frame of each pair was always taken from 500 ms after the end of the preview clip for that trial, and the second

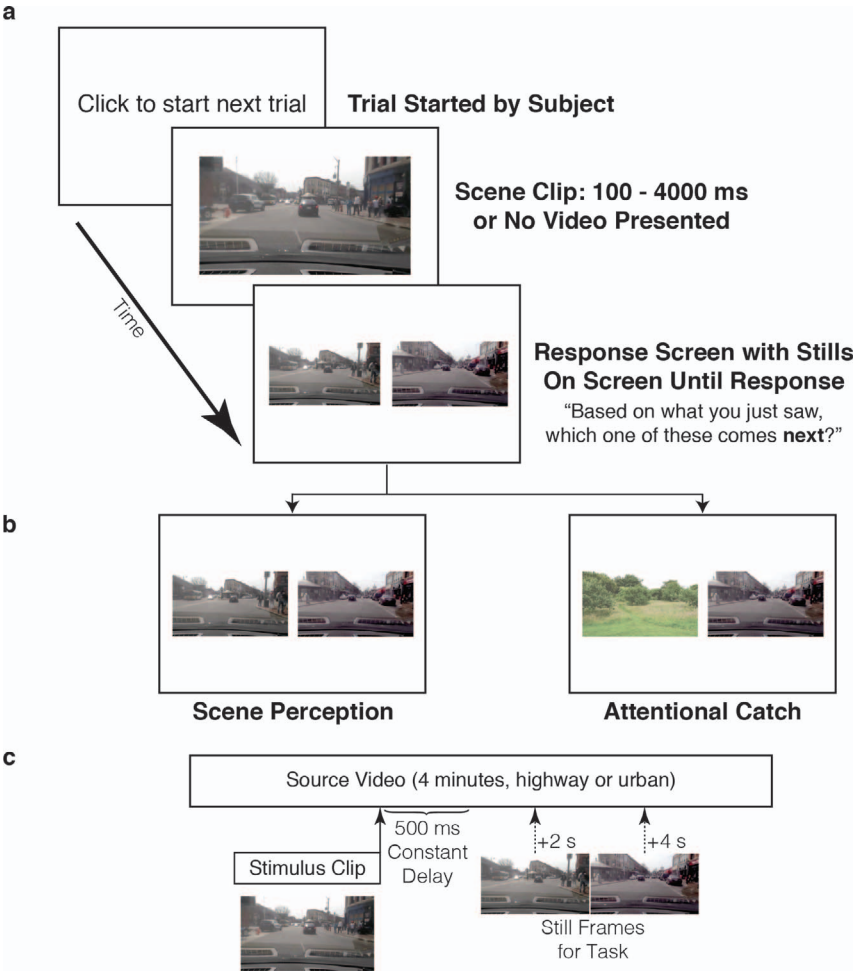

Figure 1. Stimulus sequence and video-clip extraction procedure for Experiment 1. (a) Stimulus sequence for a single trial. Subjects started each trial by clicking on a button in the Mechanical Turk interface and were shown either a stimulus clip (with a variable preview duration, 100–4,000 ms) or no video, followed by the response screen. Subjects' task was to report which of the two images on the response screen they believed came next. (b) Visualizations of the response screen. On most trials, subjects were shown two still images that came after the end of the video clip. In attentional catch trials, they were shown one still that matched the video and one image unrelated to the video. (c) Diagram of video-clip extraction for individual trials. Each stimulus clip was selected from a random time point within a longer (4-min) video. For the two images on the response screen, one of the still frames corresponded to 500 ms after the end of the clip and the other corresponded to a time point between 100 and 4,000 ms after the first still frame.

was selected from one of eight time points after the first frame: 100, 233, 500, 733, 1,000, 2,000, 3,000, or 4,000 ms. This resulted in eight frame-separation conditions, which were intended to manipulate the difficulty of the temporal-order discrimination task (i.e., answering the question "Which one comes next?"). We predicted that still frames that were more closely spaced in time (e.g., 100 ms, or three frames) would be harder to discriminate than ones more widely spaced in time (e.g., 2,000 ms, or 60 frames).

Procedure

The procedure on each trial is outlined in Figure 1a. On all trials (except the 0-ms preview-duration condition), subjects were first shown a video clip (720 pixels in height, between 100 and 4,000 ms in duration; see Stimuli) on a white background. Videos were presented without sound to avoid any extraneous auditory cues. The video was immediately followed by a response screen consisting of two still frames presented side by side (360 × 640 pixels each, horizontally separated by 100 pixels). In a two-alternative forced-choice task, subjects were asked to report which of the images would come first after the video they had watched. The two still frames were randomly assigned to the left and right locations on each trial. Note that both still images were taken from after the end of the clip, so subjects performed a temporal-order prediction task, mentally ordering the clips based on their internal representation of the scene

and reporting which one was temporally closer to the end of the clip. Subjects were instructed to click on the still frame they believed to be the one that came next, and were given an unlimited amount of time to respond. After responding, subjects were shown a screen with a button labeled "Click to start next trial," and initiated the next trial with a mouse click when ready to continue. Subjects were not given any feedback on the accuracy of their responses. Supplementary Movie S1 shows several trials similar to those used on Mechanical Turk.

In the 0-ms preview-duration condition, subjects were shown only two still images, with no preceding video. Subjects performed the same task here, answering which still frame came first in time, but without the benefit of a previously shown scene. Since the still images are themselves informative and contain cues about events and changes in the driving scene, we expected that subjects could perform above chance level (50% accuracy) in this condition by reporting the earlier of the two still frames. Therefore, data from these trials provides a baseline level of performance for comparisons to the trials that included video clips.

Each subject completed eight trials for each unique combination of road type (highway or urban), preview duration (seven durations), and frame separation (eight separations), presented in a random order. Since we performed this experiment on Mechanical Turk, we also added attentional catch trials to detect automated answering or failures to watch the video. These catch trials resembled the temporal-order prediction trials and were randomly intermixed with the rest of the experiment. In the catch trials, subjects were first shown a road-video clip with a randomly selected duration between 100 and 4,000 ms to match the rest of the experiment. Instead of the temporal-order prediction task, they were asked to indicate which of two still images matched the clip they had just seen. In the catch trials, one still image matched the clip and one had nothing in common with it (e.g., a beach scene), having been hand-selected from a Google image search to be trivially discriminable from the road-scene still images. Subjects who failed more than two of these attentional catch trials were automatically terminated from the study and compensated for the trials they had completed to that point. No subjects in the final sample of 27 missed more than one catch trial in the experiment. The full experiment, with catch trials, took approximately 1 hr for subjects to complete, and included 896 temporal-order prediction trials for all subjects who completed the experiment.

Analysis

Attentional catch trials were removed prior to any analysis, and we calculated subjects' performance (percentage of correct responses) in each condition in the remaining trials. To determine how performance on the temporal-order prediction task changed as a function of the variables we manipulated, we analyzed the percentage of correct responses with a 2 (road type) $\times$ 7 (preview duration) $\times$ 8 (frame separation) repeated-measures analysis of variance using R Version 3.50. Effect size is reported as partial eta-squared.

Results

Figure 2 shows mean group performance for each combination of frame separation, preview duration, and road type. We found a main effect of frame separation on subjects' performance, $F(7, 182) = 24.002$, $p < 0.001$, $\eta^2_p = 0.48$), indicating greater accuracy at larger frame separations (Figure 2a; upper region of heat map and upward trends in line plots in Figure 2c and 2d). As expected, stills that were spaced far apart in time were easier to discriminate than stills that were close together in time (64.1% vs. 78.3% overall accuracy in the 100-ms vs. 4,000-ms conditions, respectively). We also found a significant main effect of preview duration, $F(6, 156) = 7.39$, $p < 0.001$, $\eta^2_p = 0.22$, indicating better performance when subjects were shown a longer video clip before performing the temporal-order task (Figure 2a; right side of heat map and colored lines in Figure 2c and 2d). Overall, performance increased from 71.6% to 74.6% from the 0-ms (no-video) to the 4,000-ms preview-duration conditions (minimum and maximum performance were observed in the 100-ms and 2,000-ms conditions, with 69% and 75.7% accuracy, respectively). There was also a significant Preview duration $\times$ Frame separation interaction, $F(42, 1092) = 4.001$, $p < 0.001$, $\eta^2_p = 0.13$, indicating a relationship between how long subjects were able to view the scene and task difficulty, with the combination of large frame separations and long preview durations (upper right quadrant) producing the highest accuracy in the task. The highest accuracy (88.9%) was observed with a combination of a 4,000-ms preview duration and a 4,000-ms frame separation.

In addition, we found a significant main effect of road type, $F(1, 26) = 22.52$, $p < 0.001$, $\eta^2_p = 0.46$; mean accuracy was higher when subjects watched urban compared to highway videos (74.3% vs. 71.1%, respectively; Figure 2b and 2d). This difference between video types varied as a function of the other stimulus manipulations. Specifically, we observed a significant two-way Road type $\times$ Frame separation interaction, $F(7, 182) = 6.14$, $p < 0.001$, $\eta^2_p = 0.19$. Pairwise contrasts between the two road types at each frame separation indicated significantly higher accuracy in urban compared to highway videos, primarily at frame separations in the middle of the range that we tested (at

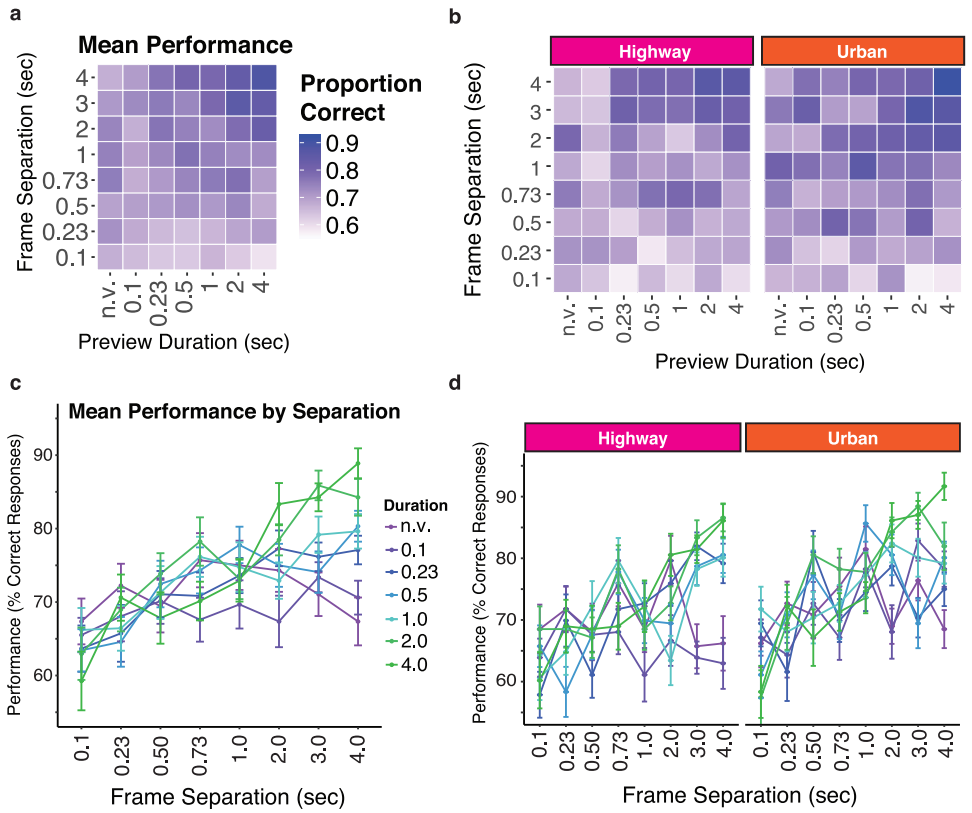

Figure 2. Results for Experiment 1, presented as heat maps (a–b) and line plots (c–d). (a) Mean performance on the temporal-order prediction task, separated by frame separation and preview duration (n.v. = no video, or the 0-ms condition) and collapsed across road type. Darker colors indicate better performance. Note the increase in performance with longer video durations, analogous to longer eyes-on-road durations, and longer frame separations, which make frame-to-frame discrimination easier. (b) Mean performance on temporal-order prediction task by preview duration and frame separation, separated by road type (highway and urban). Note the same overall pattern as in the collapsed data in (a). (c) Mean performance by separation, visualizing the same data as (a). (d) Mean performance by separation, split by road type as in (b). The same pattern is visible, but note that the prediction task becomes easier the greater the frame separation, since the frames become more easily discriminated. Error bars for (c) and (d) are ±1 standard error of the mean.

500, 1,000, and 2,000 ms, all p values < 0.001), with a smaller, nonsignificant difference at 3,000 ms ($p = 0.052$). All other comparisons were nonsignificant (all p values > 0.23). There was no significant Road type × Preview duration interaction, $F(6, 156) = 1.32$, $p = 0.25$, $\eta^2_p = 0.05$. However, there was a significant three-way Road type × Preview duration × Frame separation interaction, $F(42, 1092) = 4.51$, $p < 0.001$, $\eta^2_p = 0.15$, suggesting that the three factors are interrelated; and the difference in performance between the two road types (urban and highway) depended jointly on the frame separation and preview duration. However, any systematic patterns driving this interaction were difficult to ascertain from inspection of the data in Figure 2. We note that, given the complexity of these interactions, we simplified this design in our follow-up experiment, which measured frame-separation thresholds in a more immersive, controlled setting, testing three preview durations across two road types (urban and highway).

Discussion

Our goal in this experiment was to examine subjects' perceptions of brief driving scenes by probing their representation of the scene as a whole, rather than specific objects within the scene. By varying the preview duration, the difficulty of the task, and the contents of the scene, we examined the perceptual space here in a comprehensive manner, closer to the methods used in the gist literature than those used in studies of situation awareness. While this task is similar to the What Happens Next task (Jackson et al., 2009), there are two critical differences: Our subjects *chose* the proximate frame from the two presented options, rather than localizing, identifying, and predicting the hazard, and our stimuli *did not involve immediately hazardous situations*. In addition, manipulating preview duration while allowing subjects to freely view the scene acts as a proxy for—but not a direct measure of—the time a driver would need to keep their eyes on the road before

they could develop a representation of the environment. Varying frame separation allowed us to test a wide range of task difficulties; given the nature of the temporal-order prediction task, still frames separated by brief durations are much more difficult to discriminate than those separated by longer durations. By including both urban and highway environments, we were also able to assess the impact of scene content on subjects' ability to make predictions.

At one level, the temporal-order prediction task is comparatively simple, since even the still images contain information that subjects can use to build a cognitive model of the scene and make relevant predictions (Jackson et al., 2009; Ventsislavova et al., 2016), as shown by the results of our no-video trials. This is reminiscent of the paradigms used in the hazard-perception literature (Jackson et al., 2009; Pelz & Krupat, 1974), as well as experiments on representational momentum (e.g., Freyd, 1983), which have measured subjects' tendency to misremember the endpoint of a scene or video as being further in the direction of the implied motion. Much of this work has focused on subjects' biases in the remembered endpoint of a sequence (Freyd & Finke, 1984), rather than their ability to represent a dynamic natural scene and make predictions about it. More recent work by Blättler and colleagues (Blättler et al., 2012; Blättler, Ferrari, Didierjean, van Elslande, & Marmèche, 2010) on driving scenes has focused on whether subjects notice a change in direction of road video (following an interruption), rather than on their ability to predict the scene directly, as our subjects were required to. In contrast, the hazard-perception literature has focused on drivers' ability to report the presence of hazards (McKenna & Crick, 1994; Scialfa et al., 2012) or make a saccadic eye movement towards them to indicate detection (Crundall, 2016; Crundall & Underwood, 1998; Mackenzie & Harris, 2015; Underwood et al., 2005; Underwood, Ngai, & Underwood, 2013).

Subjects' performance suggests that the information available in either static images alone (no-video condition) or very brief videos, while usable, constrains performance in our task. As shown by the main effect of preview duration, subjects generally performed better when they were provided with longer video segments (see Figure 2c). In other words, across the range of preview durations we tested, the increase in temporal-order prediction performance suggests that providing more information in the preview phase improves the fidelity of subjects' representations of the road environment. This is intimately linked to the increase in performance across frame separation; as separation increased, making the two frames more discriminable and the task easier, performance on the temporal-order prediction task improved. In particular, subjects achieved the greatest accuracy (88.9%) with a combination of long preview durations (4,000 ms) and large frame separations (4,000 ms).

Performance also depended on the type of road environment shown. Since driving environments can vary widely in their visual features, we included both highway and urban videos in the stimulus set for this experiment. One of the most striking results we observed is that subjects performed better in urban compared to highway environments. Although several factors could contribute to this difference, one possibility is that highway environments are often comparatively homogenous, with a wide, flat roadway separated from other constructed elements, while urban environments are less defined by the roadway itself and contain a greater diversity of built structures and other objects (e.g., cyclists, pedestrians, parked vehicles). In addition, the angular subtense of road users and other objects varies across urban and highway scenes as a consequence of travel speed and roadway design (see Supplementary Movie S1 for examples), although this is inherent to the road environments in question. While perceiving the gist of natural scenes is a very fast process and minimally affected by the contents of the scene (Greene & Oliva, 2009), it appears that making predictions of natural scenes in motion requires more time and may show a larger effect of scene content than gist perception.

Experiment 2: Separating driving environment from duration

The results of Experiment 1 indicated that there was a difference in subjects' ability to perform the temporal-order prediction task as a function of road type, and that both preview duration and frame separation were linked to performance. However, as Experiment 1 was an online study, the data were collected under variable viewing conditions that would be very different from everyday driving (i.e., in terms of stimulus size, overt time pressure). We therefore rebuilt the experiment for laboratory testing, rather than via Mechanical Turk, to exercise greater control over stimulus presentation and subjects' experience. Videos were scaled to approximate a 1:1 representation of the scene from the driver's viewpoint, and we instructed subjects to respond as quickly and as accurately as possible. By using a more immersive environment, we can gain a more complete understanding of the impact of scene content on subjects' representation of the road environment. In particular, if we observe a robust effect of road type, consistent across preview conditions and in this type of immersive setting, it may be a significant potential factor when considering handoffs between automated systems and drivers.

Furthermore, it is unclear from the results of Experiment 1 what factors contribute to the improvement in performance with increasing preview duration. Is this improvement driven by the motion information in the scene, or would subjects perform similarly when provided with a static image of the scene for an equivalent duration? We therefore directly compared performance across three preview conditions: 500 ms of a still frame, 500 ms of video, and 2,500 ms of video. In the 500-ms still-frame and video conditions, subjects previewed the scene for the same duration but were provided with motion in the latter condition only.

A final consideration in planning this follow-up experiment was to simplify the design, allowing for a more direct interpretation of the results. We modified our procedure from Experiment 1, limiting the preview conditions used and allowing the frame separations to fluctuate using an adaptive staircase procedure, while retaining the same fundamental task. Recall that increasing frame separation makes the prediction task easier, as the still frames differ from each other. Therefore, instead of quantifying performance based on accuracy in the task, we calculated the frame-separation threshold for each condition—the minimum separation necessary for 80% accuracy.

Participants

A total of six men and three women participated in this experiment (mean age: 28.4 years), all of whom were unaware of the purposes of the experiment. All subjects provided written informed consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Participants and the Common Rule prior to participating in the experiment. All subjects had normal or corrected-to-normal vision and were licensed drivers with at least 1 year of driving experience. The experiment took approximately 60 min to complete, and subjects were compensated $20 for their time.

Apparatus

For this laboratory experiment, we rebuilt the experiment with MATLAB (MathWorks, Natick, MA) and Psychtoolbox (Brainard, 1997; Pelli, 1997) and presented stimuli on an LG OLED TV (55-in. diagonal size, resolution: 1,920 × 1,080 pixels, panel size: 120 × 70 cm) at a viewing distance of 57 cm. The stimulus clips covered a large portion of the screen (78° × 44°) to provide an immersive experience, approximating the field of view of the in-vehicle camera used to record the original videos. This provided observers with an approximately 1:1 representation of the driving scene as the driver would have viewed it. Subjects' head position was maintained by a chin rest throughout the experiment.

Stimuli and procedure

Stimuli were shown on a gray background. Video clips and still images were taken from the same urban and highway videos as in Experiment 1, and videos for the preview portion of each trial were randomly selected from nonoverlapping segments within the longer 4-min road videos. Periods where the vehicle in the video was not in motion (e.g., waiting at lights, parked at the roadside, stopped at intersections) were removed to avoid requiring subjects to predict upcoming motion when the recording vehicle was static. Video clips and images were shown at the center of the display (78° wide and 44° high). To control exposure duration, noise masks were shown before and after each preview stimulus. Noise masks were the same size and location as the video clips and were generated by independently drawing random grayscale intensity values for each pixel between black and white. A green cross was centered on top of each noise mask (2° × 2°; line width: 0.4°). Given the large size of the display, the cross served to reorient subjects at the beginning and end of the clip to the location of the display center. Subjects were not given specific instructions about where to look, and could view the video clips freely.

On each trial, subjects were shown a mask for 250 ms, followed by a road stimulus from one of three preview conditions: a single still frame for 500 ms, a stimulus clip for 500 ms, or a stimulus clip for 2,500 ms. Immediately following the preview stimulus, subjects were shown a second mask for 250 ms. After a 500-ms interstimulus interval, subjects were shown a response screen consisting of two still frames, similar to Experiment 1, and performed the same temporal-order prediction task. The two still images on the response screen were shown side by side at the center of the display (each 39° wide and 22° high), separated horizontally by a 4° gap. The assignment of the two stills to the left and right sides of the display was balanced across trials. Subjects were asked to report which of the two still frames presented would have come next, based on the still or video they had just seen, by pressing one of two arrow keys on the keyboard. To prompt subjects to respond, the text "Which still comes next?" was shown on the response screen above the two stills, horizontally centered on the display. The labels "LEFT arrow key" and "RIGHT arrow key" were displayed underneath the left and right image stills, respectively. Subjects were instructed to respond as quickly and as accurately as possible (across subjects, median response time was 2.09 ± 0.73

s). Following the response, the program automatically advanced to the next trial. Trials were separated by a 500-ms intertrial interval consisting of a gray screen. Supplementary Movie S1 shows several exemplar trials from this experiment.

The still frames on the response screen of each trial were selected so that one of the two frames corresponded to a time point 500 ms after the end of the preview clip or still. The second still frame was randomly selected to be either before or after this time point (referred to here as early and late intervals, respectively). The separation between the two still frames was adaptively varied from trial to trial using a staircase procedure (see Staircase procedure for frame separation). For example, if the time corresponding to the end of the video on a given trial was 2,500 ms and the separation between the two frames was 800 ms (based on the staircase procedure), subjects were shown one still frame at 3,000 ms (the correct answer) and the other at either 2,200 ms (early-interval trial) or 3,800 ms (late-interval trial). Early and late intervals were introduced to balance the number of trials in which subjects were required to discriminate the correct frame from one that was too far in the past (typically before the end of the video clip) or too far in the future. Unlike in Experiment 1, subjects would not be able to answer correctly simply by reporting the earlier of the two still frames. If the earlier of the two still frames temporally overlapped with the video, the correct answer would be the other (later) frame. Because subjects would not able to deduce the correct answer from visual cues contained within the still frames alone, we expected them to perform at chance (50%) level if they did not use any information from the video. Therefore, we omitted a 0-m (no-video) preview condition in Experiment 2.

Subjects completed 12 practice trials in which they were given feedback on their accuracy (with a green frame surrounding the chosen response if correct and a red frame if the response was incorrect). Feedback was not provided for the remainder of the experiment. Following the practice, subjects completed a total of 480 trials, with 20 trials for each unique combination of road type (highway or urban), interval type (early or late), preview condition (500 or 2,500 ms video, or still frame), and response-screen configuration (correct response on left or right), shown in a random order.

Staircase procedure for frame separation

The separation between the two still frames varied on a trial-to-trial basis and was controlled by 12 randomly interleaved adaptive staircases, one for each unique combination of road type, preview condition, and interval type (early vs. late). All staircases were controlled by a three-down/one-up rule and terminated after 40 trials. The starting value of each staircase was a 2,000-ms frame separation, based on the results of Experiment 1, which indicated that this separation was comparatively easy for subjects. Separation values increased or decreased by an initial step size of 300 ms, which was reduced by 25% every three reversals. Separation values were constrained to be between 67 ms (two video frames) and 3,000 ms. Because we used an adaptive staircase design, performance (percentage of correct responses) on the temporal-order prediction task is no longer a measure of subjects' ability to use their representations of the driving environment (as it was in Experiment 1), because the staircase is designed to converge on ~80% accuracy. Rather, the frame-separation threshold was estimated by averaging the separations across all reversal points in each staircase. This provides an estimate of subjects' performance across duration and road type, with lower thresholds indicating that subjects can maintain that level of performance with smaller frame separations.

Analysis

To calculate the frame-separation threshold for each condition, we averaged the frame separations across all reversal points within each staircase. To estimate confidence intervals, we generated bootstrapped distributions of separation thresholds for each condition by resampling the reversal points with replacement and averaging them, and repeating this procedure for 1,000 iterations (Efron & Tibshirani, 1993). These bootstrapped distributions were generated individually for each subject and then averaged to estimate group confidence intervals. Nonparametric permutation tests were used to perform pairwise comparisons between the road types within each duration condition, as well as comparisons between duration conditions within each road type. For each comparison, we randomly shuffled the labels for the two conditions and then recalculated the threshold difference for the shuffled data. This procedure was repeated for 1,000 iterations to produce a null distribution of threshold differences between pairs of conditions. The two-tailed p value was calculated from the proportion of observations in the null distribution with an absolute value greater than or equal to the observed difference. The observed p values were compared against a Bonferroni-corrected alpha, α_B, for six comparisons.

Results

Figure 3 shows mean frame-separation thresholds for each combination of road type and preview condition. In the single-still-image preview condition,

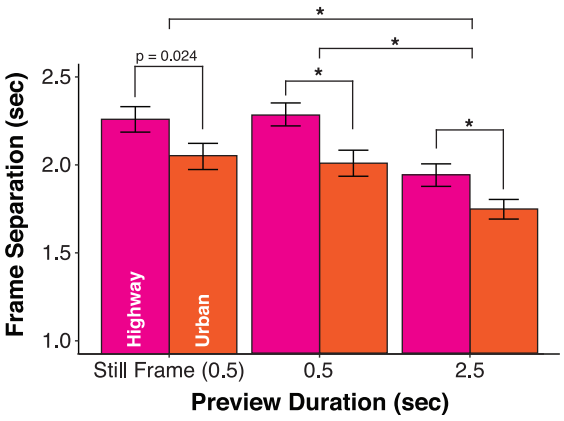

Figure 3. Results for Experiment 2. Mean separation threshold by preview condition and road type (magenta and orange indicate highway and urban stimuli, respectively). Note the difference in separation threshold between highway and urban driving environments, indicating that a greater separation between test stills in the highway condition was required for subjects to perform the task at the level enforced by the adaptive staircase. All significant differences marked with an asterisk (*) are $\alpha_B < 0.008$. Error bars represent bootstrapped 95% confidence intervals.

we found mean frame-separation thresholds of 2,247 ms for highway driving and 2,042 ms for urban driving ($p = 0.024$, permutation test; $\alpha_B = 0.008$, therefore a trending but not significant effect). In the 500-ms video condition, we found mean frame-separation thresholds of 2,271 ms for highway driving and 1,999 ms for urban driving ($p < 0.001$, permutation test; $\alpha_B = 0.008$). In the 2,500-ms video condition, we found mean frame-separation thresholds of 1,934 ms for highway driving and 1,741 ms for urban driving ($p = 0.001$, permutation test; $\alpha_B = 0.008$). Consistent with the results of Experiment 1, subjects' performance improved with longer preview durations. Comparisons between pairs of duration conditions, averaged across road type, showed significant differences between the single-still-frame and 2,500-ms conditions ($p < 0.001$, permutation test; $\alpha_B = 0.008$) and between the 500-ms and 2,500-ms conditions ($p < 0.001$, permutation test; $\alpha_B = 0.008$). Thresholds in the single-still-frame and 500-ms preview conditions were not significantly different ($p = 0.823$).

Discussion

By measuring separation threshold, rather than task performance on the temporal-order prediction task, Experiment 2 allowed us to focus more closely on the development of detailed representations over time and how they changed as a function of road type. In addition, the use of a large display in a laboratory setting allowed us to create an immersive environment and more closely approximate on-road visual information. We replicated the effect of duration in Experiment 1, showing that subjects' performance improved as duration increased. In addition, our findings clarify the results of Experiment 1, indicating that the fundamental difference between highway and urban driving environments persists, irrespective of motion or viewing duration. In particular, the differences between the 500-ms still-image condition and the 2,500-ms preview-duration condition, as well as between the 500- and 2,500-ms preview-duration conditions, suggest—consistent with Experiment 1—that additional video is remarkably useful to our subjects in predicting what comes next in these scenes. However, the lack of a difference between the 500-ms still-image and 500-ms preview-duration conditions suggests that any motion information in the videos may need to be integrated over more than 500 ms to provide a benefit over what subjects can ascertain from still images. While our results indicate that drivers are able to perceive and predict the driving environment with brief views, the critical result is that it is significantly more difficult for them to do so in a highway than an urban environment. Across both road types, increasing the preview duration yielded a mean improvement of 307 ms in frame-separation thresholds, but the overall difference in thresholds between road types was somewhat smaller, at 223 ms. Critically, this improvement in frame-separation thresholds suggests that when more information is available (from longer videos), the representation of the scene is more detailed and our subjects are able to perform a harder discrimination task. Together, they suggest that considering the environment drivers need to predict is critically important when estimating how long they will require to do so.

General discussion

We undertook this study as a step towards answering the question of how rapidly drivers can develop basic representations of their operating environment by assessing these representations based on subjects' ability to make predictions concerning the temporal sequence of natural scenes. While similar questions have long been the focus of extensive theory and research in driving, under the broad theoretical umbrella of situation awareness, we sought to better understand the time course on which these representations develop in the absence of a specific focus (e.g., hazard detection or avoidance), using a vision-science approach and drawing inspiration from work on scene gist.

Across two experiments, we found that subjects could develop adequately detailed representations of driving environments from brief clips of road video, and that they could accurately predict time-dependent aspects of the scene with less than a second of video. We found that longer viewing durations facilitated this prediction, and that making the prediction task easier (by increasing temporal separation between the test frames) had a significant effect on subjects' prediction performance. While our results indicate that longer durations facilitate subjects' representations of the scenes and improve performance on the task, our critical finding on duration is that subjects can acquire sufficient information to accurately predict the scene very quickly. That subjects can make better-than-chance predictions that improve as a function of preview duration suggests that they are quickly constructing a mental representation sufficient for the task, and their representation likely becomes more detailed with the availability of any additional information. We also found that the environment had a significant impact, and we performed a second experiment to examine this result in more detail. We found that, across all viewing durations, subjects required an easier task (larger temporal separation between still frames, hence greater discriminability) to maintain the same level of prediction performance for highway scenes than for urban scenes, suggesting that the viewing duration required would also be a function of environment and not a constant across all possible driving environments. One caveat to our work is that all of our subjects were young, and that perceptual processes and driving behaviors are known to degrade with age (Owsley, 2011). Future work examining performance in older populations would provide a more complete picture of how representation and prediction change across the life span.

Considering the difference between highway and urban environments, we would suggest that highway scenes are often comparatively visually impoverished, by virtue of their physical separation from other structures. Therefore, the built features of the environment must be taken into account when considering drivers' ability to represent their environment. Attempting to perceive the motion of other vehicles at speed in a comparatively sparse highway environment is significantly more difficult than doing so with a more densely populated scene, as in our urban driving stimuli. In urban settings, the greater availability of objects throughout the scene (e.g., parked vehicles, cyclists, pedestrians, buildings, signs) makes aspects of the task of understanding the scene itself—a necessary precondition for performing the temporal-order prediction task—much easier. The more information in the scene, the easier it is for the driver to understand which objects are moving, which are not, and whether they are moving towards or away from the driver. This idea echoes the work of Blättler and colleagues, who have found that representational momentum is affected by the content of the scene as well as the subjects' own expertise with the scene (Blättler et al., 2011, 2012; Blättler et al., 2010).

Our results suggest that drivers may acquire sufficient information to predict aspects of their proximate environment relatively quickly, suggesting that they may not need to attend to individual objects in the scene to be able to reasonably predict future frame-level changes (in contrast to the prevailing view of search and attention in driving; Ranney, 1994; Theeuwes, 1994). Although additional work is needed, our results may speak to the difference between the simple information drivers require for a split-second response and the more detailed information they require for situation awareness. While forthcoming changes in natural scenes can be predicted from still images, as shown in the representational-momentum literature (Freyd, 1983; Freyd & Finke, 1984), this requires longer viewing durations than acquiring the gist of a scene (which takes less than 100 ms; Greene & Oliva, 2009), although this work does not examine how well these changes can be predicted.

One might ask why driver perception still matters in an era when cars can, to varying extents, drive themselves. Automated vehicles are classified according to their capabilities into five levels of automation, moving from minimal to total autonomy (SAE International, 2018). A fully automated vehicle (often called a Level 5 vehicle) does not require a driver at all, and renders the focus of this article irrelevant. However, Level 5 vehicles are unlikely to be available any time soon. Lower levels of vehicular automation (Levels 1 through 3, and perhaps Level 4 in emergency conditions) assume that the driver still has a role in the safe operation of the vehicle and, critically, *assume the driver will be paying attention and be able to take control* under some circumstances—or at least is capable of doing so when required. However, there is no guarantee that the driver of an automated vehicle will treat a low-level, less capable, automated vehicle with the appropriate amount of caution; for example, a Level 2 system (e.g., Tesla Autopilot, Volvo Pilot Assist, GM Super Cruise, and other similar systems) assumes that the driver is always paying sufficient attention to be able to reassert control over the vehicle when needed. As a consequence, we must consider how the driver's visual system allows them to make the predictions necessary for action and, in particular, how quickly this can be accomplished.

Considering our results in the context of handoffs in automated-vehicle control between the vehicle and driver, our results on road type and minimum duration

for prediction have important implications. Some automated-vehicle systems are likely to be used on highways, where the environment is more comprehensible for the vehicles' automated perception and control systems. However, the same environment that helps the vehicle may also hinder the driver in an exchange of control. Our results suggest that the driver of an automated vehicle will take more time to adequately represent their environment, make predictions, and respond appropriately on a highway, where visual localization cues are reduced. For automated vehicles to be safe, they must be designed to account for the limitations of human perception relative to the capabilities and limitations of the vehicle.

To conclude, we believe that our results help to answer an outstanding question in visual perception and representation in driving: How quickly can drivers represent their environment in the absence of proximate hazards? Particularly in the context of automated vehicles and exchanges of control between the vehicle and the driver (Gold, Damböck, Lorenz, & Bengler, 2013; Samuel, Borowsky, Zilberstein, & Fisher, 2016; Samuel & Fisher, 2015; Zeeb, Buchner, & Schrauf, 2016), it is essential that we know how the driver's visual system constrains the process, and the influence of environment on the speed with which the driver can take control. While the capabilities of the driver limit what we should expect them to do in different types of handoff situations, our results contrast with the assumption that "sub-second viewing times are probably too short for processing dynamic traffic scenes" (Lu et al., 2017, p. 294). It is likely that the driver's peripheral vision (Wolfe, Dobres, Rosenholtz, & Reimer, 2017) is a major contributor to their ability to quickly represent their environment, and that serial search for objects in the scene, while occasionally necessary, may not be a component of general representation without a specific task. Automated systems may need to be designed with the awareness that the driver is remarkably capable of representing the immediate future with only a brief glimpse of the road; however, a single glimpse may not be enough to make them a safe driver, or to enable a safe takeover. However, since this work was done in the laboratory (and on Mechanical Turk), the degree to which our results extend to actual driving conditions and behavior is unknown. Future work—particularly on-road experiments and naturalistic observation studies focused on particular settings, tasks, and age groups—may shed more light on how drivers acquire the information they need and how it affects their behavior in handoff situations.

Keywords: driving, scene perception, prediction, visual attention

Acknowledgments

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

Stimuli, code, and data are available from the corresponding author on request. Experiment 1 was designed by BW, LF, AK, BS, BM, BR, and RR, programmed for Mechanical Turk by LF and analyzed by BW and AK. Experiment 2 was designed by BW, AK, and RR and programmed by BW and AK, and data were collected by BW and analyzed by BW and AK. The manuscript was written by BW, AK, BS, BM, BR, and RR.

Commercial relationships: The authors report no competing interests.
Corresponding author: Benjamin Wolfe.
Email: bwolfe@mit.edu.
Address: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.

References

Alberti, C. F., Shahar, A., & Crundall, D. (2014). Are experienced drivers more likely than novice drivers to benefit from driving simulations with a wide field of view? *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 124–132, https://doi.org/10.1016/j.trf.2014.09.011.

Blättler, C., Ferrari, V., Didierjean, A., & Marmèche, E. (2011). Representational momentum in aviation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1569–1577, https://doi.org/10.1037/a0023512.

Blättler, C., Ferrari, V., Didierjean, A., & Marmèche, E. (2012). Role of expertise and action in motion extrapolation from real road scenes. *Visual Cognition*, *20*(8), 988–1001, https://doi.org/10.1080/13506285.2012.716799.

Blättler, C., Ferrari, V., Didierjean, A., van Elslande,

P., & Marmèche, E. (2010). Can expertise modulate representational momentum? *Visual Cognition*, *18*(9), 1253–1273, https://doi.org/10.1080/13506281003737119.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Crundall, D. (2016). Hazard prediction discriminates between novice and experienced drivers. *Accident Analysis & Prevention*, *86*, 47–58, https://doi.org/10.1016/j.aap.2015.10.006.

Crundall, D. E., & Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics*, *41*(4), 448–458, https://doi.org/10.1080/001401398186937.

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, *7*(1), 16471, https://doi.org/10.1038/s41598-017-16739-x.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64, https://doi.org/10.1518/001872095779049543.

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception and Psychophysics*, *33*(6), 575–581, https://doi.org/10.3758/BF03202940.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 126–132, https://doi.org/10.1037//0278-7393.10.1.126.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1938–1942, https://doi.org/10.1177/1541931213571433.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472, https://doi.org/10.1111/j.1467-9280.2009.02316.x.

Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics*, *52*(2), 154–164, https://doi.org/10.1080/00140130802030714.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, *9*(10):6, 1–16, https://doi.org/10.1167/9.10.6. [PubMed] [Article]

Lu, Z., Coster, X., & de Winter, J. (2017). How much time do drivers need to obtain situation awareness? A laboratory-based study of automated driving. *Applied Ergonomics*, *60*, 293–304, https://doi.org/10.1016/j.apergo.2016.12.003.

Mackenzie, A. K., & Harris, J. M. (2015). Eye movements and hazard perception in active and passive driving. *Visual Cognition*, *23*(6), 736–757, https://doi.org/10.1080/13506285.2015.1079583.

McKenna, F. P., & Crick, J. L. (1994). *Hazard perception in drivers: A methodology for testing and training*. Final Report. Wokingham, UK: Transportation Research Laboratory, Department of Transport, UK.

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, pp. (251–256). Cambridge, MA: Academic Press.

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, https://doi.org/10.1016/S0079-6123(06)55002-2.

Owsley, C. (2011). Aging and vision. *Vision Research*, *51*(13), 1610–1622, https://doi.org/10.1016/j.visres.2010.10.020.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pelz, D. C., & Krupat, E. (1974). Caution profile and driving record of undergraduate males. *Accident Analysis & Prevention*, *6*(1), 45–58, https://doi.org/10.1016/0001-4575(74)90015-3.

Ranney, T. A. (1994). Models of driving behavior: A review of their evolution. *Accident Analysis & Prevention*, *26*(6), 733–750, https://doi.org/10.1016/0001-4575(94)90051-5.

SAE International. (2018). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Warrendale, PA: SAE International, https://doi.org/10.4271/j3016_201609.

Samuel, S., Borowsky, A., Zilberstein, S., & Fisher, D. L. (2016). Minimum time to situation awareness in scenarios involving transfer of control from an automated driving suite. *Transportation Research Record*, *2602*, 115–120, https://doi.org/10.3141/2602-14.

Samuel, S., & Fisher, D. L. (2015). Evaluation of the minimum forward roadway glance duration. *Transportation Research Record*, *2518*, 9–17, https://doi.org/10.3141/2518-02.

Scialfa, C. T., Borkenhagen, D., Lyon, J., Deschênes, M., Horswill, M., & Wetton, M. (2012). The effects of driving experience on responses to a static hazard perception test. *Accident Analysis & Prevention*, *45*, 547–553, https://doi.org/10.1016/j.aap.2011.09.005.

Theeuwes, J. (1994). Endogenous and exogenous control of visual selection. *Perception*, *23*(4), 429–440, https://doi.org/10.1068/p230429.

Underwood, G., Crundall, D., & Chapman, P. (2002). Selective searching while driving: The role of experience in hazard detection and general surveillance. *Ergonomics*, *45*(1), 1–12, https://doi.org/10.1080/00140130110110610.

Underwood, G., Ngai, A., & Underwood, J. (2013). Driving experience and situation awareness in hazard detection. *Safety Science*, *56*, 29–35, https://doi.org/10.1016/j.ssci.2012.05.025.

Underwood, G., Phelps, N., & Wright, C. (2005). Eye fixation scanpaths of younger and older drivers in a hazard perception task. *Ophthalmic and Physiological Optics*, *25*(4), 346–356, https://doi.org/10.1111/j.1475-1313.2005.00290.x.

Ventsislavova, P., Gugliotta, A., Peña-Suarez, E., Garcia-Fernandez, P., Eisman, E., Crundall, D., & Castro, C. (2016). What happens when drivers face hazards on the road? *Accident Analysis & Prevention*, *91*, 43–54, https://doi.org/10.1016/j.aap.2016.02.013.

Wolfe, B., Dobres, J., Rosenholtz, R., & Reimer, B. (2017). More than the Useful Field: Considering peripheral vision in driving. *Applied Ergonomics*, *65*, 316–325, https://doi.org/10.1016/j.apergo.2017.07.009.

Wolfe, B., Fridman, L., Kosovicheva, A., Seppelt, B., Mehler, B., Rosenholtz, R., & Reimer, B. (2017). Perceiving the roadway in the blink of an eye: Rapid perception of the road environment and prediction of events. *Proceedings of the 9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: Driving Assessment 2017* (pp. 207–213). Iowa City, IA: Public Policy Center, University of Iowa. https://doi.org/10.17077/drivingassessment.1637.

Zeeb, K., Buchner, A., & Schrauf, M. (2016). Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident Analysis & Prevention*, *92*, 230–239, https://doi.org/10.1016/j.aap.2016.04.002.