Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator

Oskar Palinko[1], Andrew L. Kun[1], Alexander Shyrokov[1], Peter Heeman[2]
[1] University of New Hampshire, [2] Oregon Health & Science University
{oskar.palinko, andrew.kun, shirokov}@unh.edu, heemanp@ohsu.edu

Abstract

We report on the results of a study in which pairs of subjects were involved in spoken dialogues and one of the subjects also operated a simulated vehicle. We estimated the driver's cognitive load based on pupil size measurements from a remote eye tracker. We compared the cognitive load estimates based on the physiological pupillometric data and driving performance data. The physiological and performance measures show high correspondence suggesting that remote eye tracking might provide reliable driver cognitive load estimation, especially in simulators. We also introduced a new pupillometric cognitive load measure that shows promise in tracking cognitive load changes on time scales of several seconds.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces

Keywords: eye tracking, pupillometry, cognitive load

1 Introduction

In-car electronic devices are becoming ubiquitous. While these devices provide entertainment and useful information to the driver, they may also be a distraction from the primary task in a car which is driving. In order to model user interactions with such devices, researchers often use the concept of cognitive load. Cognitive load (also referred to as mental workload) is commonly defined as the relationship between the cognitive demands placed on a user by a task and the user's cognitive resources [Wickens 2002]. The higher the user's cognitive load is, the higher the chance is that the user will not complete a given task without an error. Cognitive load can be estimated using performance, physiological and subjective measures. The focus of this paper is evaluating drivers' cognitive load in driving simulator studies using driving performance and physiological measures.

Researchers, including our team [Kun et al. 2009], have used eye trackers in driving simulators to gather visual attention data. However, eye trackers can also be a source of physiological measures of cognitive load. This is due to the fact that when people are faced with a challenging cognitive task, their pupils dilate. This simple phenomenon, called *task evoked pupillary response* [Beatty 1982] can be used in estimating cognitive load. Until recently only head-mounted eye tracking systems could provide pupillometric data. One problem with these trackers is that they can be cumbersome to use and could therefore affect the outcome of measurements [Marshall 2002]. Recently remote eye trackers appeared which are precise enough to provide useful pupil diameter measures. For example, Klinger et al. [2008] estimated cognitive load with a remote eye tracker in a desktop environment. We propose using remote trackers in driving simulators. We hypothesize that pupillometric measures will correspond with driving performance measures of cognitive load.

Two common pupillometric cognitive load measures are the index of cognitive activity (ICA) and the mean pupil diameter. ICA uses the frequency of minute dilations of the pupil [Marshall 2002]. This method is used almost exclusively with head-mounted eye trackers owing to their high precision. The mean pupil diameter can be easily calculated even with remote trackers. Because of the averaging process, this calculation is more resistant to measurement noise than the ICA. In section 3.2.3 of this paper we propose a new pupillometric measure, the mean pupil diameter change rate, for estimating rapid cognitive load changes.

The goals of this study are to evaluate the correspondence between driving performance and pupillometric measures and to provide a preliminary evaluation of the mean pupil diameter change rate as a measure of rapid changes in cognitive load.

2 Related research

2.1 Cognitive load

In the literature researchers dealt with three types of cognitive load measures: performance, physiological and subjective measures [O'Donnell and Eggmeier 1986]. Performance measures capture how well the user is performing a given task. For driving, this can include lane departures, steering wheel variance, visual attention to the outside world, etc. Physiological measures include pupil dilation [Bailey and Iqbal 2008], heart-rate variability, and galvanic skin response. Changes in these measures have been shown to correlate with varying levels of cognitive load [Reimer et al. 2009]. However, physiological measures depend on many factors, including other aspects of the user's cognitive state (such as stress [Healey and Picard 2000] and arousal), the user's physical activity and environmental variables (such as temperature). Subjective measures capture the user's subjective assessment of cognitive load. A commonly used assessment tool is the NASA-TLX questionnaire [Hart 1988]. While such a tool is relatively simple to administer, it cannot account for rapid changes in cognitive load that may be the result of changes in experimental conditions. In this work we evaluate the agreement between physiological measures based on pupil size measurements with a remote eye tracker and driving performance measures.

2.2 Pupillometry

Iqbal et al. [2004; 2005] conducted experiments with subjects performing manual-visual tasks in front of a computer screen. They measured task completion time, percent change of pupil size (PCPS), average percentage change of pupil size (APCPS) and subjective ratings. PCPS is calculated as the difference between the measured pupil size and a baseline pupil size divided by the baseline pupil size. APCPS is the average of this measure over a time period. The authors found that the PCPS correlated well with

Figure 1 Driver and dispatcher.

the mental difficulty of the task. More complex tasks resulted in higher values of PCPS compared to easier tasks. This study used a precision head-mounted eye tracking system (EyeLink 2). The same eye tracker was used by Schwalm et al. [2008] in a driving simulator study. Their subjects performed the standardized lane change task [Mattes 2003] and an additional visual search task. As dependent variables, they looked at driving performance, NASA TLX and the index of cognitive activity (ICA). As driving performance, the mean lane position deviation was considered which is the mean difference between the driven path and a so-called optimal path. The study found that the ICA correlates well with the driving performance measure: when the additional visual task was introduced, driving performance decreased and the ICA increased. The authors attributed this change in ICA to changes in workload, but the visual nature of the secondary task could also have affected the index of cognitive activity.

While head-mounted eye trackers are useful for precise eye measures, they are impractical in consumer driving environments and can also affect the results of the experiments [Marshall 2002]. Thus researchers have turned to remote eye tracking in cars to infer the cognitive load of the driver. Recarte and Nunes [2000] used monoscopic remote eye tracking in a naturalistic driving experiment. The eye tracker measured gaze information as well as pupil diameter. The diameter was measured by the number of image pixels. While driving on the road, subjects were given two kinds of secondary mental tasks: a verbal and a spatial-imagery task. Pupil diameter measures showed differences between secondary task and no secondary task conditions, but did not show significance for the different kinds of secondary tasks.

In a later study Recarte et al. [2008] built on their prior findings, while using a remote eye tracker in front of a screen instead of in a car. The authors compared three measures of mental workload: NASA-TLX, pupil size and blink rate. They found that NASA-TLX and pupil size cannot discriminate between mentally and visually challenging tasks. On the other hand blink rate was a very good measure for indicating these differences: high visual demand inhibited blinks while a high mental workload task without a visual component increased the blink rate. In our experiment, the secondary tasks are predominantly cognitive so we should expect to see changes in pupil size measurements.

Recently, Klingner et al. [2008] reported on a study that involved cognitive load estimation using remote eye tracking in front of a computer screen using the monoscopic Tobii 1750 device. The change in cognitive load was caused by the introduction of tasks such as mental multiplication, digit sequence repetition and aural vigilance. Lighting conditions were strictly controlled in this experiment. The authors concluded that remote eye tracking is a viable way of cognitive load estimation using pupil diameter measurement. Building partly on the results of this work we use a remote eye tracker in a driving simulator to estimate the cognitive load. While we do not explicitly control lighting conditions, the brightness of the simulated world (road surface, sky, vegetation, etc.) varied less than ±5% from the average brightness along the simulated roads.

3 Experimental setup

We hypothesize that human multi-threaded spoken interactions can provide insight into how to design interfaces for spoken human-computer interactions [Shyrokov 2010]. In multi-threaded dialogues conversants switch between individual dialogue threads and these threads may overlap in time. We are especially interested in multi-threaded dialogues when one of the conversants is involved in a manual-visual task, e.g. driving. Thus, in our experiment pairs of subjects are engaged in two spoken tasks and one of the subjects (the driver) also operates a simulated vehicle. One spoken task is the ongoing task and it is periodically interrupted by another spoken task. The interruptions force subjects to switch between different dialogue threads. We track the pupillometric and driving performance measures of the driver's cognitive load.

3.1 Equipment

Two subjects (driver and dispatcher) participated in each experiment (see Figure 1). They communicated using headphones and microphones. Their communication was supervised by the experimenter and synchronously recorded as a 44100 Hz mono signal.

The driver operated a high-fidelity driving simulator (DriveSafety DS-600c) with a 180º field of view, realistic sounds and vibrations, a full-width car cab and a tilting motion platform that simulates acceleration and braking effects. We recorded pupillometric data using a SeeingMachines faceLab 4.6 stereoscopic eye tracker mounted on the dashboard in front of the driver.

3.2 Method

3.2.1 Participants

The experiment was completed by 32 participants (16 pairs) between the ages of 18 and 38 (the average age was 24). Nine participants were female. Subjects were recruited through advertisements and received $20 in compensation.

3.2.2 Driving (primary) and spoken tasks

The primary task of the drivers was to follow a vehicle while driving responsibly. They drove on two-lane, 7.2 m wide roads in daylight. The lead vehicle traveled at 89 km/h (55mph) and it was positioned 20 meters in front of the subject. There was also a vehicle 20 meters behind the subject's car. No other traffic was present on the road. The roads consisted of six straight and six curvy road segments with straight and curvy segments alternating.

Our ongoing spoken task was a parallel version of *twenty questions* (TQ). In TQ, the questioner tries to guess a word the answerer has in mind. The questioner can only ask yes/no questions, until she is ready to guess the word. In our version, the two conversants switch roles after each question-answer pair is completed. Words to guess were limited to a list of household items (hair dryer, refrigerator, TV, etc.). We trained participants to use a question tree in order to guess the objects. The words to be guessed were presented to the subjects visually. We showed words to the driver just above the dashboard which minimizes interference with driving. We told subjects that there was a time limit to finish a game, and we enforced this time limit.

Our interrupting task was a version of the *last letter* word game (LL). In our version of this game a participant utters a word that starts with the last vowel or consonant of the word uttered by the other participant. For example, the first participant might say, "pag*e*" and the second says "*e*arn" or "g*e*ar." Subjects had 30

seconds to name three words each. After completing this task they resumed the TQ game. Subjects played one TQ game and were interrupted by one LL game per curvy and straight road segment.

3.2.3 Design

We conducted a within-subjects factorial design experiment with the part of the spoken interaction as our primary variable, *St_part*. We split the spoken interaction into three parts: 1) the beginning of the twenty questions game (*st_part* = TQ1) which goes on before the conversants switch from TQ to the interrupting task, 2) the interrupting last letter game (*st_part* = LL) and 3) the completion of the twenty questions game (*st_part* = TQ2). In this paper we only consider data from interactions on curvy road segments. We evaluate the following three dependent variables.

Standard driving performance measures. In this paper we look at the variances of lane position and steering wheel angle. We calculate both as the average of variances for the six curvy road segments. One average is found for each of the three parts of the spoken interaction. Lane position refers to the position of the center of the simulated car and is measured in meters. A large variance in lane position is a sign of poor driving performance, since it means that the participant cannot keep the vehicle on a steady path. Steering wheel angle is measured in degrees. Steering wheel angle variance can be used as a relative measure of driving performance when comparing the performance of multiple participants on road segments of the same type. A higher variance is an indication of increased effort to remain in lane.

Pupillometry. We analyze the commonly used mean pupil diameter change (MPDC), which is calculated as the average for the six curvy road segments of the difference between the mean pupil diameter in a given curvy segment and the overall mean pupil diameter for a given subject. The overall mean is subtracted from the segment mean in order to compare results between subjects with different pupil sizes.

Next we look at how the driver's cognitive load changes during LL. In LL the interaction between driver and dispatcher can be divided into driver turns and dispatcher turns. During the driver's turn the driver's attention is divided between driving and the LL game, as she has to think of a word to utter and then has to utter the word. During most of the dispatcher's turn the driver can concentrate on driving and only has to pay attention to the LL game when the dispatcher utters a word. We hypothesize that this interaction pattern will result in larger driver cognitive load during the driver's turn than during the dispatcher's turn. Figure 2 shows a sequence of pupil diameter measurements demonstrating this phenomenon. While the driver is thinking and then speaking her pupil is generally dilating, indicating a possible increase in cognitive load. Conversely, while the dispatcher is thinking and speaking the driver's pupil is generally contracting.

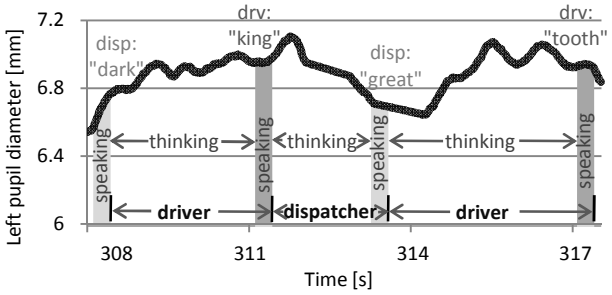

Figure 2 Example time diagram of pupil diameter change during a last letter game (LL).

To evaluate cognitive load changes in LL we conducted a within-subjects factorial design experiment with the turn as our primary variable, *Turn*. We evaluate the three dependent variables from above as well as the new mean pupil diameter change rate (MPDCR). MPDCR is the average for the six curvy segments of the mean value of the first difference (which is the discrete-time equivalent of the first derivative) of the pupil diameter for a participant's turns. A positive MPDCR indicates that on average the pupil dilated for that subject's turns (possibly due to increased driver cognitive load) while a negative value signifies pupil contraction. Note that, in evaluating cognitive load changes in LL, the other three dependent variables were also calculated using means for participant turns in curvy segments.

4 Results and Discussion

We performed a series of repeated measured ANOVAs for the first three dependent variables with *st_part* as the independent variable (Figures 3 and 4). We found statistically significant differences in steering wheel variance ($F(2,30)=25.0$, $p<.001$) while performing different parts of the spoken tasks (TQ1, LL and TQ2). Post hoc pair-wise comparisons also show differences between all levels ($p<.006$). The differences in lane position variance are also statistically significant ($F(2,30)=10.0$, $p<.001$). Post hoc comparisons show significance between TQ1 and LL ($p<.002$), TQ1 and TQ2 ($p<.007$) but not between LL and TQ2 ($p<.175$). Note that on average TQ1, LL and TQ2 took about 22, 30 and 45 seconds to complete respectively.

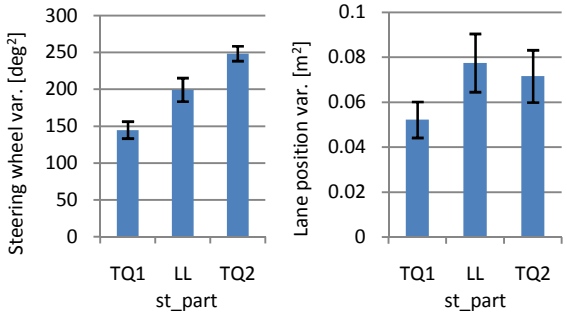

Figure 3 Driving performance (with standard error).

Similar to Schwalm et al. [2008] we found that performance and physiological data largely agree: in our experiment both the driving performance data and the physiological (pupillometric) data indicate that the driver's cognitive load is lowest during TQ1. Specifically, MPDC was significantly different between TQ1, LL and TQ2 ($F(2,30)=15.6$, $p<.001$). In post hoc comparisons MPDC was significantly different between TQ1 and LL ($p<.001$) as well as between TQ1 and TQ2 ($p<.001$), but not between LL and TQ2 ($p<.47$). Note that in contrast to Recarte and Nunes [2000] our cognitive load measures were significantly different for different secondary tasks (TQ and LL).

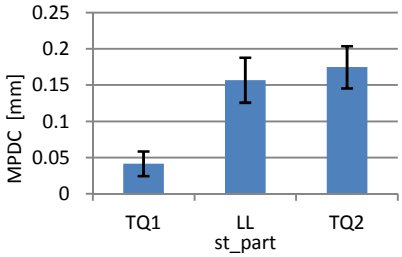

Figure 4 Mean pupil diameter change (with standard error).

Statistical analysis also confirmed that the pattern in Figure 2 is not an isolated incident. MPDC during LL, shown in Figure 5, was significantly larger during the driver's turn compared to the dispatcher's turn ($F(1,15)=59.69$, $p<.001$). Similarly, MPDCR shows that during the driver's turns, the driver's pupil diameter was increasing, and it was decreasing during the dispatcher's turns ($F(1,15)=14.37$, $p<.002$). Thus, both MPDC and our newly introduced MPDCR appear to be valuable tools in detecting rapid changes in cognitive load. This is in contrast to the two driving performance measures, which were not significantly different between driver and dispatcher turns during LL. Driving performance is too coarse of a measure and does not neatly follow rapid changes in cognitive load. Note that on average both driver and dispatcher turns took about 4.6 seconds to complete.

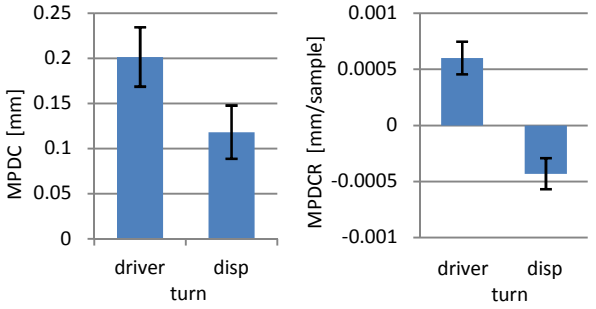

Figure 5 Pupillometric measures during LL (with standard error).

5 Conclusion

Our results show correspondence between two driving performance measures and the MPDC under our experimental conditions. We suggest that this correspondence is due to convergence of physiological and performance measures of cognitive load. Thus we expect that remote eye tracking is a viable way of cognitive load estimation in a simulated driving environment. Our results also indicate that the MPDCR shows promise as a pupillometric measure of cognitive load. We found it to be a sensitive measure of changes in cognitive load. We expect that this measure might be especially useful when observing rapid changes in cognitive load. For such changes the average pupil size might not change significantly between different tasks, but the first difference might. Finally, our results indicate that both MPDC and MPDCR are finer measures of cognitive load in a driving simulator than variances of lane position and steering wheel angle.

6 Acknowledgements

This work was funded by the US Department of Justice under grant 2006DDBXK099 and by the NSF under grant IIS-0326496.

References

BAILEY B.P., AND IQBAL S. T. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans. on Computer-Human Interaction*, 14(4):1–28.

BEATTY, J. 1982. Task Evoked Pupillary Responses, Processing Load and Structure of Processing Resources. *Psychological Bulletin*, 91(2): 276-292.

HART S.G., AND STAVELAND L.E. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoritcal research. In *Human Mental Workload*, by Meshkati N. Hancock P. A., 239–250. Amsterdam: North Holland Press.

HEALEY J., AND PICARD R. 2000. Smartcar: Detecting driver stress. *Proc. ICPR.* Washington DC: IEEE Computer Society.

IQBAL S.T., ADAMCZYK P.D., ZHENG X.S., AND BAILEY B.P. 2005. Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution. *Proc. CHI'05.* Portland, OR: ACM, 311-320.

IQBAL S.T., ZHENG X.S., AND BAILEY B.P. 2004. Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. *Proc. CHI'04.* Vienna: ACM, 1477-1480.

KUN A.L., PAEK T., MEDENICA Z., MEMAROVIC N., AND PALINKO O. 2009. Glancing at personal navigation devices can affect driving: experimental results and design implications. *Proc AutomotiveUI'09*, ACM.

KLINGNER J., KUMAR R., AND HANRAHAN P. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. *Proc. ETRA '08.* Savannah, GA: ACM, 69-72.

MARSHALL, S.P. 2002. The Index of Cognitive Activity: measuring cognitive workload. *Proc. IEEE Conf. on Human Factors and Power Plants.* 7-9.

MATTES, S. 2003. The lane-change-task as a tool for driver distraction evaluation. *Proc. ISOES.*

O'DONNELL R.D., AND EGGMEIER F.T. 1986. Workload assessment methodology. In *Handbook of perception and human performance: Vol. II*, by Thomas J.P. Kaufman L. New York: Wiley Interscience.

RECARTE M.A., AND NUNES L.M. 2000. Effects of verbal and spatial-imagery tasks on eye fixations while driving. *J. Experimental Psychology*, 6(1):31-43.

RECARTE M.A., PEREZ E., CONCHILLO A., AND NUNES L.M. 2008. Mental workload and visual impairment: Differences between pupil, blink and subjective rating. *Spanish Journal of Psychology*, 11(2):374-385.

REIMER B., MEHLER B., COUGHLIN J.F., GODFREY K.M., AND TAN C. 2009. An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. *Proc AutomotiveUI'09*, ACM, 115-118.

SCHWALM M., KEINATH A., AND ZIMMER H.D. 2008. Pupillometry as a method for measuring mental workload within a simulated driving task. In *Human Factors for assistance and automation*, by Flemisch F., Lorenz B., Oberheid H., Brookhuis K. De Waard D. Maastricht: Shaker Publishing.

SHYROKOV, A. 2010. Human-human multi-threaded spoken dialogs in the presence of driving. PhD dissertation (in preparation). University of New Hampshire.

WICKENS, C.D. 2002. Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Sci.*, 3(2):159-177.