Data Collection Report

Cheng Chen Liu

April 19, 2017

Abstract

Driver distraction from secondary in-vehicle activities is recognized as a significant source of people injuries and fatalities on the road. Cognitive workload as one main source of diver distraction is vital to understand the driver state in partially automated cars. eDREAM Project, conducted during May 2015 to November 2016, was initiated to develop an advanced driver monitoring system that utilizes advanced sensory and vision technologies to improve driving experience and safety. Vehicle-based measures, physiological measures and video-based measures data were collected in order to discover the various impacts of cognitive load. Those measures were collected from a total of 36 gender-balanced participants and a driving simulator under three incremental cognitive task-load conditions. The NASA-TLX questionnaire was used for rating various demands and efforts in order to collect participants' perceived cognitive workload level after each drive that contained different task-load. This document focused on the process of experiment design and implementation, future sections on resulted dataset and analysis results will be added.

Contents

Glossary

ANOVA Analysis of variance.. 7

Critical Period A period in each drive where different levels of task-load would be imposed (while all other conditions are controlled to be the same). 8, 22–25, 28

D-Lab Software for collecting signals from physiological sensors and webcams in this experiment. 16, 17, 19, 31, 32, 34, 35, 38

ECG Electrocardiogram. 7, 11, 25, 32, 35

EEG Electroencephalogram. 3, 7, 10, 11, 25, 32–34

EV External Vehicle, refers to the vehicle controlled by the experiment participant in the driving simulator. 21–23, 27

FaceLab The commercial eye-tracker used in the experiment. 15, 19, 25, 31, 36, 37

GoPro The colour camera used in the experiment. 16, 19, 31, 38, 39

GSR Galvanic Skin Response. 7, 11, 25, 32, 34–36

HRV Heart Rate Variation. 35

IVIS In-Vehicle Intelligent Systems.. 7

LV Lead Vehicle, which is the vehicle directly in front of the external driver in the driving simulator. It is programmable and controlled by the simulator. 21–24, 26, 27, 29

miniSim The driving simulator system. 15, 16, 19, 21, 23, 24, 27, 28, 30–33, 36, 37

RESP Respiration. 7, 11, 25

1 Exclusive Summary

With the popularization of the vehicle and development of technology, the public begin to pay attention to the issue of enhancing driving safety through some techniques. One of the way which is helpful to reduce the accident rate is enabling the vehicles or devices to detect driver inattention. So the "EDREAM" project is presented to reach the goal of mitigating negative effects of high secondary cognitive task-load on drivers through developing an advanced driver monitoring system with advanced sensory and vision techniques. In general, we know that distractions can be classified into three categories including manual distraction, visual distraction and cognitive distraction. In EDREAM project, it narrows the problem by focusing on cognitive tasks which is a more challenging area comparing to fatigue or visual distraction. In studies in this field, the impact of increased mental workload on various measures by conducting statistical analysis is always the important issue. It worth mentioning that fewer works research on cognitive distraction and mental workload assessment comparing to other topics in driver monitoring such as driver's vigilance or distraction, instead of "daydreaming" cases.

In previous studies, researchers mainly chose to employ statistical analysis tools to extract essential trends from the signals or build machine learning models and analyze selected features. However, this experiment does not have the limitations of the ensuing data analysis approach. The objectives of this experiments are collecting various typed of data from drivers and isolating the impacts of cognitive workload. For the purpose of finding different impacts of cognitive load and facilitate the study of possible information fusion for driver mental workload estimation, three approaches were employed to collect data which are vehicle-based measure, physiological measures and video-based measures. Because of the insensitivity on the onset of driver state changes and some other reasons, vehicle measures may not provide the most reliable and consistent indications for assessing driver's mental workload. Although physiological measures enjoy the advantages of being more sensitive and having faster response, it is hard or even not realistic to apply those intrusive physiological sensors in real life driving. The approach of video-based measures, which aims to detect blink, gaze and visual attention estimation, facial expression and so on, is easy to use and quick at detecting the onset of abnormal driver states. In order to isolate the impacts of cognitive workload, three levels of task-load were showed in separate drives in a counter-balanced order. In addition, the external conditions, such as vehicle behaviour, ambient traffic and driving route, are controlled to be the same. However, it is more difficult to elicit, isolate and identify changes happened purely in cognitive states comparing to other common driving problems. And it is also challenging to detect internal cognitive changes by observations during run time. As a consequence, the effectiveness of the tasks should be designed carefully and optimized during design process.

During the early phase of the project, we collect data through a driving simulator study from three incremental cognitive workload scenarios. We collect driving performance measures including driver's various physiological, visual and performance changes due to three incremental cognitive task-load conditions: no task (baseline), 1-back (lower external cognitive task-load), and 2-back (higher external cognitive task-load).

Beyond that, video recordings, a remote eye-tracker and a comprehensive set of physiological signals are also acquired including EEG, Electrocardiogram (ECG), Galvanic Skin Response (GSR) and Respiration (RESP). To minimize the variability in participant skills, controlling the age group and ensuring the function of eye-tracker system, we recruit people who satisfy the requirement of under 35-years old, driving at least several times per month and hold a full driver's license (G license or equivalent) for at least 3 years, and without glasses. Our data acquisition system contains driving simulator and vehicle-based measures, vision-based measures, physiological measures, subjective measures and enjoys the property of data synchronization. The implantation of the experiment is challenging mainly due to the goal of controlling all traffic conditions while not being completely monotonic and the instability of the miniSim system and its associated tools. We increase the complexity of map for miniSim to induce enough workload for participants so that we could create a scenario which is closer to real world. Besides, we specify LV's behaviours, redesign drives and task arrangements, ensure signal quality to get a better experiment result.

In this report, we first introduce the background and objective of the EDREAM project, briefly summarizing some previous works and analyzing their room for improvement and differences between our works and some previous studies. Then we present our experiment from the aspects of the requirement of choosing participants, the composition and characteristics of the Data Acquisition System, scenario design and implementation, challenges and discussion and so on. In the final section, we shown how the collected data were organized.

2 Introduction

In order to fulfill the modern needs of enhancing driving safety under cognitive distractions imposed by In-Vehicle Intelligent Systems (IVIS), the “EDREAM” project was initiated to develop an advanced driver monitoring system that utilizes advanced sensory and vision technologies to improve driving experience and safety. As an early phase of the EDREAM project, a driving simulator study was conducted to collect data from various sources under three incremental cognitive workload scenarios. In addition to driving performance measures, video recordings, and a remote eye-tracker, a comprehensive set of physiological signals were also collected, including EEG, Electrocardiogram (ECG), Galvanic Skin Response (GSR) and Respiration (RESP).

This report provides a comprehensive documentation of the EDREAM experiment. Experiment objectives and related works are introduced later in this section. The experiment design, including the scenario implementation, data acquisition system and participant recruitment methods, are reported in Section 2.0. Challenges and lessons learnt are discussed within each topics’ subsections. In the future, contents for data reduction process and analysis results should be added.

2.1 Research Questions

Cognitive distraction and mental workload assessment are less developed fields comparing to other topics in driver monitoring: most studies focused on driver’s vigilance or distraction would either exclude the cases of “daydreaming”, or treat cognitive distraction as part of general distraction induced by the secondary tasks. Eliciting, isolating and identifying changes that happened purely in cognitive states would be much more difficult comparing to other common driving problems.

However, recent crash statistics suggested the use of voice-command interface or hands-free mode of IVIS would not remove all the adverse effects, although they omit taking eyes off road or hands off wheel [1]. This revealed the necessity of studying driver’s cognitive workload, and build the foundation for developing next-generation smart vehicles that could monitor driver’s cognitive states, and take actions (either take control from the driver or send out appropriate alerts) to avoid detrimental consequences.

As there were less established conclusions, studies of driver’s cognitive states would consider searching for informative measures (“discovery”) on top of creating the most effective application system (“application”). There were two main approaches in previous studies on driver’s cognitive states: either employ statistical analysis tools (e.g. ANOVA) to extract essential trends/characteristics from the signals [2], or build machine learning models and analyse selected features [3]. This experiment does not impose limitations on the ensuing data analysis approach, and the resulted dataset could be used to answer the following research questions:

(a) What kind of changes would accompany increased cognitive workload?

   i. Would our findings be consistent with previous studies (if similar measures were employed)?

ii. How could we index cognitive workload?

iii. How to determine when cognitive workload would impair the driving performance?

(b) How to assess drivers' cognitive workload based on signals from vehicle-based measures, physiological sensors, and/or video-based measures?

i. Is it possible to determine cognitive workload in real time?

ii. Which modelling method would be most efficient in determining the drivers' cognitive workload?

(c) Do driving performance (i.e. speed keeping, lane keeping and reaction to hazardous events) degrade when cognitive task-load increased?

(d) How effective are the chosen secondary cognitive tasks in raising cognitive workload?

2.2 Objectives

With the research questions identified in the previous subsection, two principal objectives that guided the whole data collection process were identified.

The first objective of this experiment is to collect various types of data from drivers. In previous studies, there has been three approaches to measure driver's states objectively: (1) vehicle-based measures, (2) physiological measures and (3) video-based measures [4]. In order to discover the various impacts of cognitive load and facilitate the study of possible information fusion for driver mental workload estimation, this experiment collected data coming from all of these three approaches (see Table 1). To our knowledge, this is the first driving experiment that collects such a wide selection of signals. Thus, the end dataset could also provide a unique opportunity for studying the differences/correlations of different types of signals. As EDREAM project is a multi-disciplinary collaboration, this dataset would be analysed by researchers coming from different backgrounds, with diverse skills and interests.

The second objective of this experiment is to isolate the impacts of cognitive workload. Three levels of task-load (from no task to higher task-load) were presented in separate drives in a counter-balanced order. It is assumed that different task-load levels should significantly influence the participants' internal cognitive states. In order to isolate the impacts of cognitive workload, this simulator study designed several Critical Period where all external conditions (except the cognitive task-loads) are controlled to be the same. These conditions include the lead vehicle behaviour, ambient traffic, and driving route. The secondary task responsible for imposing the different task-load level was a modified version of the n-back task, which was inspired by its application in several driving studies [5]. Nonetheless, it is very difficult for experiment investigators to detect internal cognitive changes by observations during the run time. Thus, the effectiveness of the tasks should be carefully considered and optimized during the design process.

2.3 Background

Enabling vehicles/machines to detect driver inattention has been a long desired function, and will continue to be in the future of self-driving cars. Conditions that could negatively influence driving safety include drowsiness, extreme emotions, and distractions. Specifically, distractions can be classified into three categories: manual distraction (e.g. operating the car's air conditioner), visual distraction and cognitive distraction. Different types of distractions are not always individually separated or independent to each other, thus several of them can contribute to an accident at the same time.

In EDREAM project, we narrow the focus to driving problems associated with cognitive tasks. This is a more challenging area comparing to fatigue or visual distraction. Studies in this field often aimed to discover the impact of increased mental workload on various measures by conducting statistical analysis (e.g., [6, 7, 8, 2]). On the other hand, a handful of studies that applied machine learning to develop real-time classification or prediction systems also research for the most representative features of cognitive distraction [9, 10, 3, 11]. These studies proposed various theories. But due to the complexity and difficulty of this problem, a big gap remains for developing a robust detection system.

In the Background section, we attempt to provide some insight into the topics that forms the background of this experiment, including: 1) similarities and differences of cognitive distraction, workload and task-load, 2) possible approaches for driver monitoring, and 3) previous findings on the influence of high cognitive workload.

2.3.1 The Cognitive Aspect

Before moving further, we should clarify the terms that are used to describe driver's cognitive states in this study. Cognitive distraction and mental workload are two inter-related concepts that appears very similar in many studies, but they are different at root. In real world, the driving task itself can already be very demanding in terms of mental workload: for example, an experiment was dedicated to capture the changes of driver workload due to different road conditions [12]. On the other hand there are studies on cognitive distraction (e.g. [9]), where some secondary task will be imposed during driving.

The Yerkes-Dodson law [14] is a classical model to relate task performance with mental arousal, which was used in the experiment's proposal to motivate the problem. Its core idea is often represented as a bell-shape curve (Figure 1). On the left end of the curve, task performance degrades when the arousal level is low, which corresponds to the danger of driving under fatigue or low-vigilance situations. Secondary tasks that adds mental workload (e.g. listening to radio or having conversation with passengers) may have positive effects on driver when they help to bring the arousal level higher, moving to the center, optimal range of the bell shape. However, when the mental workload is overloaded, causing anxiety or stress, driving performance also degrades.

This model does not demonstrated the conflict of cognitive demands and reallocation of resources, where the total mental workload or arousal level might not have changed significantly. There are other mental

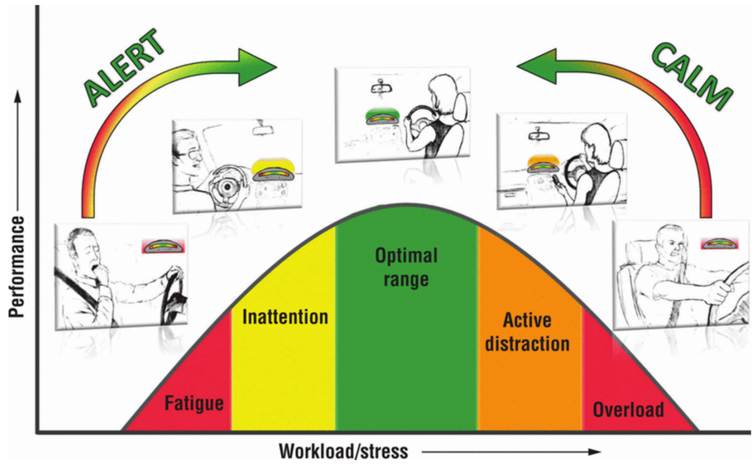


Figure 1: The Yerkes–Dodson law models a relationship of task performance vs. mental arousal. This version is adapted to MIT wellness concepts for driver monitoring, originally presented in [13].

workload models that could be employed to provide explanation for this. For example, Wickens' multi resources model [15] suggests that different tasks and different stages of the task may require similar cognitive resources and the high risk of driving with secondary task might be caused by the conflict of cognitive resources. This explains the potential danger of cognitive distraction.

The EDREAM project targets to mitigate the negative effects of high secondary cognitive task-load on the driver, which could involve both increase in mental workload and cognitive distraction. A variety of secondary tasks have been used in previous driving studies regarding distraction or mental workload. Some of these tasks are more realistic (e.g. conversation or following GPS navigation) while others are more abstract and artificial (e.g. arithmetic problems or n-back tasks [5]). Researchers often make assumptions that the task load manipulated the participant's cognitive states in their expected way. Otherwise, they rely on subjective measurements (e.g. NASA-TLX [16]) or expert ratings to obtain the "ground truth" of cognitive distraction level. It should be noted that obtaining the instantaneous ground truth of the driver's internal cognitive changes is very challenging, if not unrealistic.

2.3.2 Driver Monitoring Approaches

To study different driver states and advance driving safety, researchers have managed to accommodate a variety of sensors into vehicles and/or driving simulators. Overall, signals can be acquired from the following aspects to measure driver state:

- Vehicle-based measures: lane deviation, speed steering wheel angle, brake response time;
- Physiological measures: heart rate, skin conductance level, respiration rate, EEG band power;
- Video-based measures: blink, gaze and visual attention estimation, facial expressions, body gestures and movements.

Vehicle-based measures are the most convenient to acquire, it only requires the vehicle to be equipped with necessary sensors. It received higher user acceptance and has been incorporated in some commercial vehicles for fatigue detection already [17]. However it is insensitive to the onset of driver state changes since there may be no immediate effect reflected by driving performance changes. On the other hand, physiological measures extract signals from driver's body directly, so it is considered more sensitive and has faster response. This type of measures has been used widely in researches, but it is not realistic to apply those intrusive physiological sensors in real life driving. Currently, there are efforts to making these sensors less intrusive and/or "hide" them into car seats or steering wheels, but the signal quality may also degrade [18]. Finally, video-based measures observe drivers' behaviours and visual attention. The effect can be considered similar to having a co-pilot looking after the driver all the time (with supervision and special knowledge of "reading" driver's mental states). This approach is not only technically challenging because it requires more processing steps, but also because the necessary computer vision modules are computationally expensive. However, a working video-based driver monitoring system combines the advantage of vehicle-based and physiological approaches: it should be easy to use, and also quick at detecting the onset of abnormal driver states.

Comprehensive reviews for driver inattention monitoring systems covering all approaches can be found in [4] and [19], while [17] focuses solely on drowsiness detection. In another paper [20], systems that features face monitoring is surveyed.

Table 1: Example Impacts of Mental Workload to Various Measures

Measure	Trend with Increased Cognitive Task-load
Physiological	
EEG	Alpha band ↓ [21, 22]
ECG	Heart rate ↑ [7]
GSR	↑ [6]
RESP	Rate ↑ [7]
Vision-based	
Visual Attention	Peripheral Checking Times ↓ [2]
Horizontal Gaze	Standard deviation ↓ [6]
Blinking	↑ [10]
Performance-based	
Vehicle Speed	Average ↓ [6]
Steering Angle	Reversal rate ↑ [7]
Subjective	
NASA-TLX	Perceived workload ↑ [2]

2.3.3 Impacts of High Mental Workload

Degradation of driving performance was observed under high mental workload (e.g. intense braking [2], increased standard deviation of speed [6]), indicating the hazard of paying attention for non-driving tasks.

In some studies, reducing speed and increasing headway were considered as means for drivers to accommodate higher cognitive workload [23]. However, while a change in driving behaviour may be attributed to voluntary adjustments, it also depends heavily on the traffic conditions, task difficulties and individual capabilities. Therefore, vehicle measures may not provide the most reliable and consistent indications for assessing driver's mental workload. For example, [6] and [7] found contradictory trends in standard deviation of vehicle speed.

In physiological measures, increase of heart rate and galvanic skin response (GSR) were commonly observed when the mental workload increases [5, 6]. Electroencephalogram (EEG) sensors which provides direct measure of electrical activity of the brain were also favoured in this research area. A variety of changes in EEG band power were reported in driving or aerial studies concerning high mental workload (as reviewed in [21]), while several studies agreed on decreasing of power in alpha band when the workload increased [21].

The impacts of cognitive distraction on driver's visual attention was featured in a naturalistic on-road experiment [2]. Drivers spent more time looking centrally ahead and reduced glances to periphery area, instrument displays or mirrors. Similar symptom was referred as "visual tunneling" effect in other driving studies [8]: the visual attention tends to concentrate to the center and peripheral vision is reduced. This was measured by decrease of horizontal gaze dispersion, captured using remote eye-trackers in on-road studies [8, 7]. On the other hand, eye blinking has also been studied for cognitive workload estimation, however mixed results were reported. Seeking a possible explanation, Recarte et al. theorized that higher visual demand reduces blink frequency while higher mental workload increases it [24]. This suggests that eye blinking maybe heavily influenced by the visual demands of the driving task and can not always reflect the variation of mental workload very well. A wide range of eye-related measures for assessing mental workload were reviewed in [25].

3 Experiment Specification

This section provides a comprehensive description of the EDREAM experiment design and decision making process on topics including scenario design, apparatus and participant requirements.

To summarize, this experiment measures driver's various physiological, visual and performance changes due to three incremental cognitive task-load conditions: no task (baseline), 1-back (lower external cognitive task-load), and 2-back (higher external cognitive task-load). All task conditions were presented for approximately 2 minutes within three separate 10-minutes drives; the orders of the conditions were counterbalanced across participants. The task-load levels is designed to be the sole dependent variable; the other external conditions (such as driving route and traffic) are the controlled variables and are kept constant throughout the experiment. Participants only needed to follow a lead vehicle at a normal, constant speed and were not subjected to changing lanes or making turns. Besides keeping the speed, the only requirement embedded in the driving task was to respond to the lead vehicle braking events, which were inserted strategically to measure the participant's reaction time under different conditions. Extensive training and preparation that ensures the task completion quality were conducted before the actual data collection drives. Subjective questionnaires were completed immediately after each drive.

The finalized experiment design presented in this section is a collective result of numerous revisions from August 2015 to June 2016. Multiple difficulties were encountered and addressed during experiment implementation and pilot testing. These include inducing mental workload effectively, mitigating participant's anticipation or fatigue, and coping with various practicality issues. Design decisions are presented as a small subsection following the description of each design element; the effectiveness of our design and other challenges are discussed in the last subsection.

3.1 Participants

A total of 36 gender-balanced participants were recruited through campus and online posts to participate in this driving simulator study. Participants were recruited with the following 3 requirements:

- Drive at least several times per month and hold a full driver's license (G license or equivalent) for at least 3 years;
- Must be under 35 years old;
- Drive without glasses (contact lenses are allowed);
- The 36 participants are gender balanced (18 males and 18 females).

The first requirement ensured the participants are practised drivers; we excluded novice drivers while focused on drivers younger than 35 years old in order to minimize the variability in participant skills. The second requirement controlled the age group, since previous driving studies found age to be an influential factor for cognitive tasks, and the youngest age group showed best performance on a similar secondary task [7]. The third requirement was placed to ensure the function of the eye-tracker system.

Participants were compensated at CAD$12 per hour, and participants were told that they could receive a bonus of up to CAD$14 bonus amount based on their task performance as an incentive for engaging in the secondary task. Each experiment took approximately 2.5 hours, and we allocated an additional half hour contingency time in case of any problem raised from sensors or software; only one of the experiments needed more than 3 hours, unfortunately the participant chose to resign at that time so it was not completed. All participants were paid for three hours and full bonus amount (i.e. CAD$50) regardless of their actual time spent or task performance.

3.1.1 Notes on Recruitment

To summarize the recruitment procedure, individuals interested in participation would be directed to complete an online screening questionnaire on SurveyMonkeys.com. This questionnaire consists of a few questions evaluating their eligibility to attend the experiment. If all passed, then an email invitation would be sent for selecting a preferred experiment time on Google Doodles showing all available time slots. However, it was challenging to find participants because of the specificity of requirements, potential discomfort of wearing multiple physiological sensors, and the experiment length.

Advertising the experiment: While we prepared the screening questionnaire and could create a doodle for appointments easily, getting potential participants' interests was a lot more difficult. Due to the length and complexity of the experiment setup, and the strict requirements (experienced driver and no glasses) that ruled out huge portion of students in the university, lacking participants was a major reason for delays at several stages of the experiment. A variety of advertising methods were employed. At first, posters and recruitment emails targeting university students were sent, which received few responses. The first group of paid participants were all friends of the researchers, which ended up to be the pilots for "Version 1" because the experiment design were changed considerably after them. When recruiting for "Version 2" participants, more success came from listings on classified websites (e.g. Kijiji or Craiglist). Our summer student reported that advertising in any boards/categories with remote correlation to the targeting demography helped attracting more responses. For example, after posting in "Domestic/Community" sections, more females signed up for the screening questionnaire.

Scheduling and delays: Each appointment requires 3.5 hours, 3 hours for running the experiment and 0.5 hour for preparation and cleaning up. Exporting experimental data needs another 30 to 60 minutes, but it was not necessarily performed after each experiment due to time limitation. We could fit three time slots everyday for 3.5-hour appointments in morning, afternoon and evening (e.g.: 8:30am to 12:00pm, 1:00pm to 4:30pm, 5:30pm to 9:00pm); we also opened possibilities for setting up appointment on weekends. As recorded in the simulator room calendar, EDREAM experiment had a total of 53 bookings. There were more than half of the appointments booked in evenings or weekends (16 in weekday evenings and

12 on weekends), revealing the necessity of coping with participants' schedule and open for appointments after regular working hours. In the end, there were only 36 participants who successfully completed the experiment. The biggest loss came from participants' last-minute cancellations or no-shows. Secondly, three participants resigned from the experiment because of 1) ineligibility (only held a G2 licence), 2) severe motion sickness and 3) overtime. Finally, there was a cancellation due to malfunction of the simulator displays. In these cases, eligible participants were always compensated at the hourly rate for the time they spent on the experiment. In addition to the above conditions, one major cause of delay was the difficulty of recruiting female participants. This was the main reason that completion date was stretched to late-October, 2016.

3.2 The Data Acquisition System

This section describes the driving simulator, apparatus for collecting visual or physiological measures from participants, and the subjective measures applied.

3.2.1 Driving Simulator and Vehicle-based Measures

The study was conducted on a NADS miniSim^{TM} driving simulator (Figure 2). This fixed-based simulator has three 42-inch screens, creating a 130° horizontal and 24° vertical field at a 48-inch viewing distance. The center screen displays the left and center parts of the windshield; the right screen displays the rest of the windshield, the rear-view mirror, and the right-side window and mirror; while the left screen displays the left-side window and mirror. Driving data was recorded at 60 Hz.

Figure 2: NADS Driving Simulator and placement of eye-tracker.

3.2.2 Vision-based Measures

Eye-trackers: Eye-tracking information was collected at 60 Hz through FaceLab 5.0, a eye-tracker by Seeing Machines. It is mounted at the center of the platform above dashboard, providing an ideal frontal view of the driver (see Figure 2). The sampling frequency of FaceLab is 60 Hz. The eye-tracking data is stored locally in the standalone computer that runs the eye-tracking software, while part of the data is also forwarded to the miniSim computer and stored in the miniSim log. In addition, another application, EyeWorks, grabs image frames from miniSim and overlays the tracked gaze position on it, producing an intuitive visual record of where the eye-tracker estimates the person is looking at within the central screen.

Colour Cameras: A GoPro Hero4 camera is placed at front-right of the driver on a tripod. Also, there are two professional HD Logitech Webcams (model numbers C920 and C930) fixed on the driving simulator, facing the driver from front-left and upper-front angles (see Figure 3 for an illustration). The placement of this set of cameras recreates the possible camera positions in a real car, which would be most convenient to set up, interfere the least with the driver's sight field, and provide a non-obstructed view of the driver face. Note that due to differences in participants' heights and seat adjustments, the cameras' relative locations to the participants' faces could not be perfectly fixed.

The GoPro camera records videos into a stand-alone SD card, while the Logitech Webcams are connected to the D-Lab software. Unfortunately, signals from the Logitech Webcams were often lost during the experiments, possibly caused by hardware limitations of the D-Lab computer. Therefore, GoPro recordings should be considered as the main, reliable video source while the other sources are complementary.

Figure 3: Camera and eye-tracker placements in the driving simulator. The webcams (left and upper) are framed in red, while GoPro camera (right) is framed in green. Eye-tracker cameras (centre) are also marked (in yellow). The grey area at lower middle of the image represents the approximate location of participant's head.

3.2.3 Physiological Measures

EEG data was collected using the Muse headband developed by Interaxon (Figure 4), a wireless non-intrusive headband consisting of 2 dry sensors located at Fp1 and Fp2 positions and two gel foam electrodes at TP9 and TP10 positions. The EEG headband was worn around the forehead (Fp1 and Fp2) with two electrodes attached behind the ears (TP9 and TP10). The associated software, MuseLab, was used to record and analyze the EEG signals. The sampling frequency was 220 Hz and the software calculated the power of EEG bands at 10 Hz.

Figure 4: Research MUSE Headset used for experiment.

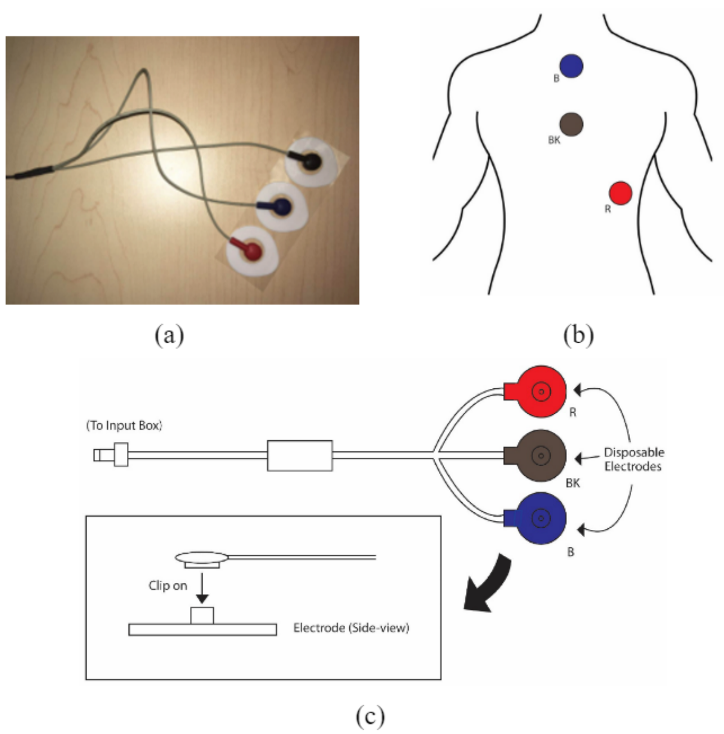
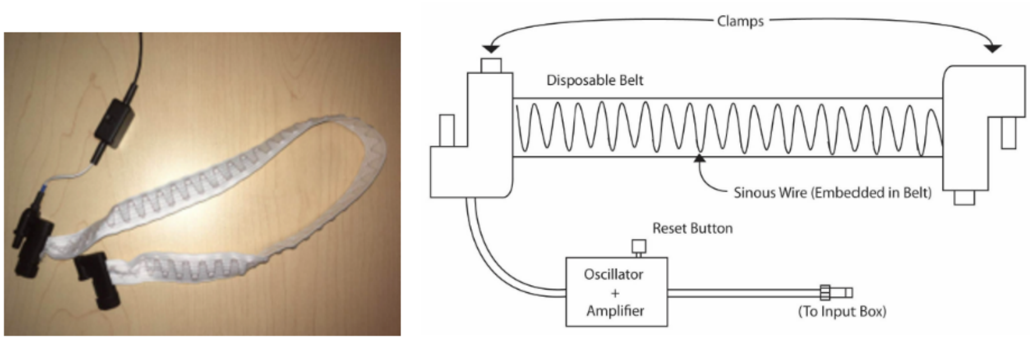


Figure 5: ECG device (a), placement (b), and wiring setup (c).

ECG, GSR, and respiration sensors by Becker Meditec collected data at 240 Hz using the D-Lab software developed by Ergoneers. Solid gel foam electrodes were used for ECG (Figure 5) and GSR sensors (Figure

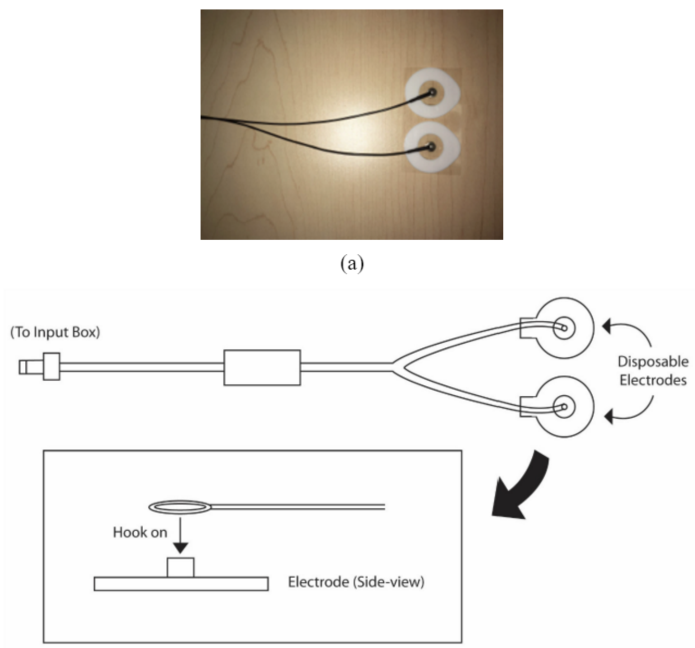
7).

ECG sensor (By Becker Meditec): For this experiment, heart electrical activity was recorded by the ECG electrodes placed on the participants' body. One electrode was placed on the neck over the vertebra, one placed on the left side of the rib cage over the second lowest rib, and one placed over the uppermost part of the centerline of the rib cage. These electrodes detect electrical changes on the skin caused by heart muscle depolarization during each heartbeat.



(a) (b)

Figure 6: RESP device (a), and wiring setup (b).



(a)

(b)

Figure 7: GSR device (a), and wiring setup (b).

Respiration Belt (by Becker Meditec): The respiration belt is an elastic belt containing wires that change resistance as it stretches and relaxes. The ends of the belt are attached to clamps that can be assembled together to form the circular loop around the thorax/abdomen area. The respiration belt was worn around the chest or abdomen, at the position that exhibited most heaving when the participants breathed. The changes in frequency caused through stretching the belt is used to characterize breathing patterns and measure changes in respiration during the driving sessions. For every application of the belt, the reset button should be pressed (hold for at least 2 seconds) to bring the signal into the middle of the measuring range and identify the "natural stretch length" of the belt. For better fit, the belt may be cut and adjusted to individual participant as needed.

GSR (by Becker Meditec): The GSR sensors were attached beneath the bare left foot with one sensor in the middle and the other under the heel. The GSR Amplifier measures the change in electrical properties of the skin to monitor the participants' sweat gland activity. Since sweating is controlled by the sympathetic nervous system, GSR measures psychological/physiological arousal. The amplifier can be used with the disposable electrodes by hooking the circular end of the wires to the center of the electrodes and applied to the suitable locations on the body (palm of hand, two fingertips, and sole of foot).

3.2.4 Subjective Measures

In order to collect participants' perceived cognitive workload level, we asked them to complete three subjective questionnaires after each drive that contained different task-load. The NASA-TLX questionnaire was used for rating various demands and efforts. For example, the participants were asked to rate how physically/mentally demanding an individual drive is. In addition, risk perception and mental effort were also rated by the participants. All questionnaires are stored in the final data copy.

3.2.5 Data Synchronization

There were mainly four data collection systems running on three different computers: miniSim, FaceLab, D-Lab and MuseLab (the later two runs on the same computer). Ethernet connections were used for communications between each system. It was decided that all systems would be synchronized to miniSim's frame number. Two summer undergraduate students worked on creating scripts that forwarded miniSim frame number to D-Lab and MuseLab, which were written in Python and Java correspondingly. On the other hand, there were existing function in miniSim that receives output from FaceLab. Once enabled, the correspondence of miniSim frame number and FaceLab frame number could be found in miniSim's log file.

The only part that was not synchronized with others was the GoPro camera, mainly because it was added after the formal experiments had started. We thought of recording a couple seconds of the FaceLab frame number, since this was the only accurate timing information displayed during the experiment. However, later processing work discovered that the FaceLab frame number were not increasing uniformly and thus this information gathered at the very beginning of each GoPro recording would not be useful for events several minutes later. As an alternative solution, it would be possible to identify the segments were tasks were presented from its recorded audio. Extra effort would be required for achieving more accurate synchronization.

An alternative way for data synchronization would be based on universal time standard (i.e. UTC). In a lot of the result logs or recordings, UTC timestamps were used. However, we were not able to use them because this would need the clocks on the computer to be accurately calibrated. This method might be considered in the future if each individual system's clock could be automatically calibrated (Internet connection may be required); this would save a lot of trouble for producing the relaying programs.

3.3 Scenario Design and Implementation

This section describes the primary and secondary tasks that the participants were required to perform. Details of design considerations and implementations were also covered, which includes task arrangements and specific rules for enforcing certain conditions. These aspects of considerations were incorporated as "triggers" in customized scenario files created for this driving simulator experiment.

3.3.1 The "n-back" Task

A modified version of the n-back task was employed to introduce external cognitive load to participants in previous driving studies [5]. In our n-back tasks, participants listened to a prerecorded series of 10 randomized letters, separated by approximately 2.5 second intervals (thus 25 seconds for one task). Two types of n-back tasks were presented (see examples in Figure 8):

- 1-back task (lower taskload): participants were asked to count the number of times two identical letters appeared in pairs in a sequence (e.g., YY), and answer the count at the end of the sequence;
- 2-back task (higher taskload): participants were asked to count the number of times two identical letters appeared in pairs with one letter in between (e.g., MOM) and answer the count at the end of the sequence.

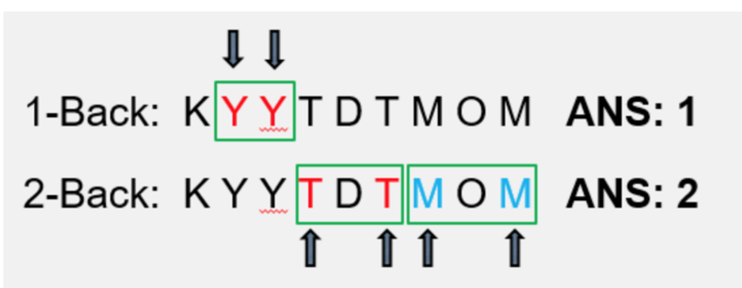


Figure 8: Two n-back task examples. Letters counted as n-back patterns are highlighted in colours.

The Meaning of "n-back": In general, the "n-back" task refers to a family of continuous performance tasks that tests participant's working memory. The core idea is that when presented with a sequence of stimuli (e.g. numbers, letters, picture of objects) drawn randomly from a finite pool, participants should try to identify if the current stimulus is identical the to one presented "n" steps ago. When the value of "n" increases, participants would need to store more items in their memory. In addition, parts of working memory must be constantly updated as new stimulus is received. Thus, the task become more difficult as "n" increases.

Modifications for EEG collection: Different than the "delayed digit recall (n-back) task" described in [5] where participants always answered the stimulus presented "n" steps ago, in our modified version, participants only answer the total count of n-back pattern appearances at the end of each series. This was to minimize the facial muscle movements that interfere with the EEG signals. Note that this modified version of n-back task requests more cognitive resources as participants need to keep track of the running total. To partially accommodate this added difficulty, letters instead of numbers were used as the stimuli to minimize the interference of remembering the stimuli and counting (e.g. avoid the confusion of the count "1" and the stimulus "1" that appeared). On the other hand, since participants only answer the count at the end of each sequence, there is no way to check their correctness at detecting each single n-back patterns.

Implementation Details: Audio recordings for the n-back tasks were prepared, and they were the same for all participants. The actual letter sequence were generated by randomly drawing letters from a pool of 6 letters ("A, B, C, F, G, H"). Initially we used the first 10 letters from the alphabet, then we removed ones that created confusion (e.g. "D" or "E") and reduced the pool size to increase the occurrence of n-back pattern. Participants were told that we would reward bonus money based on task performance; however, we did not expect them to get all answers correct because 2-back could be quite challenging during driving. Their answers were recorded in a spreadsheet during the experiments.

3.3.2 The Driving Task

In order to resemble the common driving situation realistically, the primary task was defined to be safely driving the External Vehicle (EV). The driving task was designed to involve only basic operational decisions, with little strategic or tactical decisions (such as navigating a route or passing a vehicle). Participants were instructed to follow the Lead Vehicle (LV) at a speed of 40 Mph on the a 4-lane urban route. Participants understood that they did not need to change lanes or take turns at intersections, but would need to respond the LV's abrupt braking.

Map and route. A new miniSim map in an urban environment was created for this experiment (see Figure 9). It consists of three straight segments and two curved segments in between; the curves introduces some variation to the simplistic driving task, otherwise the participants would only need to control the vehicle speed throughout the whole drive. There are two lanes for each traffic direction; part of the outer lane were be filled by pre-rendered parked cars. It was designed to allow placement of two groups of tasks on the later two straight segments, while the first straight segment functions as a start and curved segments provides small breaks within the drive.

Traffic and objects. By design, participants would not need to respond to behaviour of other objects in the simulator except the LV and traffic lights. However, ambient traffics (e.g. vehicles driving in other lanes or waiting at intersections) and pedestrians were added to make the scenario more realistic. These not only avoided the scenarios to look too empty that makes it unrealistic, but also encouraged participants to pay more attention to observe the peripheral environment (and hopefully check the mirrors). There were some variations in placements of the vehicles and objects across different drives, but we controlled the traffic to impose the same amount of demand during the Critical Period.

Braking events. The lead vehicle was programmed to brake (either intensively or slightly) throughout the drives. These braking events also allowed for collecting the participants' reaction times under different task conditions. It was hypothesized that when distracted by secondary cognitive task-load, the participants would respond slower to the braking events, and thus reveals the harm of overloading cognitive tasks. On the other hand, braking events also increased the difficult and danger of the simplistic driving task. To minimize the variation in braking events demanding level, the LV was programmed to always automatically adjust itself to create a 2-sec time gap between the EV. The time gaps achieved at the lead vehicle brake onset varied due to vehicle dynamics (mean=2.11s, SD=0.56s for the first 15 participants).

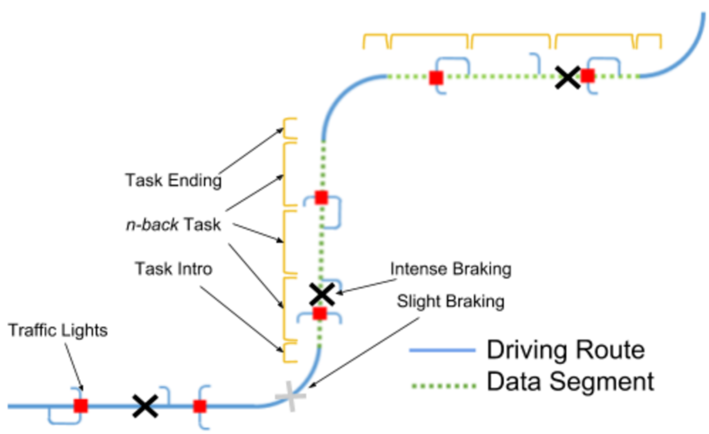


Figure 9: Illustration of a formal experiment scenario (map is not to scale). The driving route contained straight and curved segments. The green dotted portions of the map identified the data segments, within which all conditions are controlled such that changes of cognitive task-load could be isolated. The secondary task's audios (marked by yellow brackets on the side) lasted slightly longer than the data segment. Example placement of braking events were also marked in black crosses, note that this is changed in different drives. Traffic lights were always green during data segments.

3.3.3 Task Arrangements

Each participant went through three drives with 1-back tasks, 2-back tasks and no task; the order of these drives were counter balanced. There were two Critical Periods in each drive, covering most of the later two straight segments that lasted for 2 to 3 minutes each; a section of curved road in between acted as a break for approximately 45 seconds. The insertion of this break followed the design methodology presented in [5], which was a way to reset the participants cognitive states and avoid cognitive fatigue.

The task groups: The n-back tasks always appear in a group consisted of three tasks (a series of 10 letters each) of the same type, presented consecutively. In addition, a notification and a brief reminder of the task were provided before each task group to let the participant know that the n-back task was starting. Also, at the end of each task group, another notification was provided to let the participant know that the task had ended. To ensure the period where we played the audio recordings would fall inside the Critical Period, the control of the traffic conditions were elongated to start earlier and end later than the audio. There was one task group placed on each of the later two straight segments, thus totaling six n-back tasks in each drive. Figure 9 illustrates the above arrangement.

Traffic and LV braking events in Critical Periods: We controlled all other conditions to isolate the effect of the dependent variable (i.e. task-load level) during the Critical Periods. During these periods, traffic lights were always green and the ambient traffic consists of vehicles driving on the opposite directions only. There were no vehicles or pedestrians on the side of the road or at intersections, thus minimizing the variation of visual distractions during the whole Critical Period. The LV were programmed to brake once during each task group, either during the first task or the third task (always different within the same drive). Thus there were a total of 2 braking events from each drive, and only 4 out of 6 n-back tasks were not affected by this dramatic event. The baseline (or no-task) drive had the same traffic pattern and LV behaviour as the n-back drives. The only difference was that there would be no audio recording played during Critical Period.

3.3.4 Training and Warm-up

In addition to the three data-collection (or "formal") drives, participants needed to complete two additional drives and one training session for the n-back tasks before-hand to ensure proper preparation and mitigate any learning effects, which means that the participants get more proficient and feel a lot easier after the first few times of performing the task.

Training for driving in miniSim: Placed at the beginning of the whole experiment, the "training" drive was the first time participants were introduced to the driving simulator and the primary driving task of this experiment (i.e. following the LV at 40 Mph through an urban area). As a beginning, it was designed to be very simple such that participants could get accustomed to the rendered graphics and disappearance of the force of deceleration/acceleration. This was especially useful for people who were prone to motion sickness. Also during the training drive, guided by the investigators' explanations, participants experienced the LV mechanism of always trying to keep a "constant" time gap with them, therefore when their speed decreases the distance gap would seem to decrease, and vice versa. This helps to alleviate the fear of keeping a 2-second gap with the LV, which was closer than some people's preference and might be unsafe in a real-world driving situations. Participants were suggested to keep around the 40 Mph speed limit, under which condition the gap would feel more comfortable. In addition, the LV braking events would be introduced in the later half of the training drive. There were multiple braking events for them to practise the operation of the EV in miniSim. They were free to ask any questions during this drive; although investigators would not answer with exact implementation detail, they could suggest them to keep a reasonable speed and watch-out for braking events, which usually alleviate the insecurity. The training drive could last for as long as 15 minutes, but it could be ended earlier as long as the participants felt comfortable with all the driving conditions.

Training for n-back task: There would then be the training for n-back, which ensured the participants were fully capable of handling this task. The training script is included in complementary documents. We always made sure them complete three tests after providing explanation and examples of each task. However, it was still found to be inadequate as it was common for the pilot participants to decrease the vehicle speed during the first Critical Period since they did not experience with handling both the driving and secondary task together.

The warm-up drive: The "warm-up" drive was added after the above-mentioned trainings and setting up all sensors (see Table 2 for the experiment procedure). Participants were deceived to believe that this drive was the first out of four formal drives they needed to complete. In this drive, there were a group of three 1-back tasks and a group of three 2-back tasks, appeared at the same time of the Critical Periods in formal drives. We found this provided more practising and such could alleviate the learning effects, and participants would not feel extra nervous or stressed during the actual data-collection segments of the formal drives. In addition, the warm-up drive were also richer in both LV's and other traffic/object's behaviour. Since we try to control most of these elements to be very similar during the actual formal drives, the variations inserted in this drive was thought to encourage participants' observation of the external environment. More specifically, more braking events were presented in a group of n-back tasks to minimize participants' anticipation of the systematic nature of braking events that were going to happen in the formal drives (i.e. once in every task group, or once in every straight segments of the road). This drive was added after discovering the problems of learning effect and anticipation problems from our pilot tests. Finally, after this drive, participants were also introduced to the NASA-TLX questionnaire, Risk Perception questionnaire and Rate Scale Mental Effort questionnaire at the end of the warm-up drive. Data collected from this drive was also saved and might be used as testing data.

3.4 Experimental Flow

Table 2 presents the experiment steps and time allocations. Prior to the experiment, the investigators must arrive approximately 20 minutes earlier to ensure everything were prepared. It was necessary to start all computers and software. Sensors and cameras must be taken out from storage and placed at designated locations, if they were moved by other users of the simulator room. Connectivity, battery level and relay program should also be checked. If the scenario files were not produced, then new ones must be made by modifying previous files. Consent forms, cash and receipt were also prepared at this time.

Participant eligibility was verified and consent form was signed upon arrival. The participants first went through a practice drive in the simulator, on a route identical to the one used in the formal experimental drives. They practiced following the lead vehicle at a 2-second time gap and experienced lead vehicle braking as it would happen in the experimental drives. If the participant did not experience severe motion sickness and wish to continue, they were then given written and oral instructions on the modified n-back task and practiced it without driving. This ensured that they fully understood and were capable of doing the n-back task. Physiological sensors were then placed on the participants and the eye tracker was calibrated.

Next, participants were told they would complete four formal drives, while only the later three drives were formal data collection drives where Critical Period were embedded in. The first drive was designed to be an additional "warm-up" drive, where they could familiarize performing the n-back task while driving. The participants were told that this was an formal drive in order to minimize their anticipation of where and when lead vehicle braking events were to occur in the experimental drives, since they were placed at drastically different locations in the warm-up drive.

Finally, participants went on to complete the three formal drives. The task conditions followed the predetermined counterbalanced order. Subjective questionnaires were were given during the 5-minute break after each drive. At the end of the experiment, participants were debriefed and received their payment.

Table 2: Experiment Steps

Step and Description	**Time**
Prepare the data collection system: Start all data recording systems, assemble physiological sensors, battery check	10:00
Arrange other necessities: Ensure scenarios, forms and cash were ready	10:00
Meeting the participant: Greeting, introducing the experiment, and signing consent form	10:00
Training drive: Introduce the driving simulator, driving conditions, and braking events	15:00
n-back training: Introduce the secondary cognitive task and allow practise	20:00
Break: Allow participants to use washroom before attaching sensors	5:00
Sensors setup: Calibrate the FaceLab eye-tracker and attach EEG, ECG, GSR, and RESP sensors	30:00
Warm-up drive: With a group of 1-back and a group of 2-back	10:00
Break: Complete the subjective questionnaires, refresh	5:00
First formal drive: Task condition following a predetermined counterbalanced order	10:00
Break: Complete the subjective questionnaires, refresh	5:00
Second formal drive: Task condition following a predetermined counterbalanced order	10:00
Break: Complete the subjective questionnaires, refresh	5:00
Third formal drive: Task condition following a predetermined counterbalanced order	10:00
Break: Complete the subjective questionnaires, refresh	5:00
Debriefing: Remove sensors, concluding remarks, payment and collect receipt	10:00
Contingency: Extra time for unexpected problems (e.g. sensor calibration) or delays (e.g. longer break time)	30:00
Cleaning up: Dissemble physiological sensors, turn-off all data recording systems, battery charging	10:00
Total time	**3:30:00**

Table 3: Course of the EDREAM experiment

Experiment Phases	Tasks and Description
Phase 1: Planning (May - Dec 2015)	1.1 Experiment proposal and ethic review, determine research objectives 1.2 Design experiment methodology, task conditions 1.3 Build data collection system and adopt physiological sensors 1.4 Develop synchronization programs between data collection systems
Phase 2: Implementation (Sept 2015 - Feb 2016)	2.1 Training for operating multiple technical systems (e.g. eye-tracker) 2.2 Program task and driving conditions into simulator scenarios 2.3 Address problems with simulator scenarios and synchronization 2.4 Conduct pilot testing
Phase 3: Revision (Feb - Apr 2016)	3.1 Manage problems encountered during previous pilot testing 3.2 Incorporate more habituation time before formal data-collection drives 3.3 Adjust task and driving conditions, create new map and scenarios 3.4 Conduct pilot testing
Phase 4: Data collection (May - Oct 2016)	4.1 Recruit participants (advertisement, communication) 4.2 Conduct experiments and collect data following the designed experiment flow
Phase 5: Dataset organization (Oct 2016 - current)	5.1 Create organized data copies 5.2 Prepare dataset manuals and experiment documentation 5.3 Data synchronization and labeling

3.5 Process of the Data Collection Campaign

Significant amount of effort and time were put into this data collection experiment: starting from conceptual design, then testing and fixing practical problems during implementation, and finally complete the data collection process. Table 3 summarizes the entire process of this data collection campaign.

We have implemented two versions of experiment design. The first version (Version 1) was completed by September 2015 and the implementation already commenced before that; but two areas of problems were overlooked and unaddressed. First, there were still several functional problems with the scenario implementation and data acquisition system (such as undefined LV behaviors and incomplete synchronization methods). These were not fully addressed until February 2016, after which we conducted pilot testing to test this implementation. The second area of problems was that participants might not respond to the events in our expected way, which was suspected earlier but not confirmed until the first round of pilot testings. It was easier for participants to feel drowsy rather than cognitively overloaded during our experiments. Also, the driving task was found to be unnatural and/or unrealistic, leading to diverse driving behaviours for different participants. To address these problems rooted in experiment methodology and faithfully achieve the experiment objectives, we revised the map, scenario and procedural arrangement. Numerous new design considerations and extra steps were incorporated. This second version (Version 2) of design was implemented and tested by May 2016, and formal experiments were started soon after that.

3.6 Challenges and Discussions

This section discusses the challenges of designing and implementing this experiment, documents the efforts for trouble-shooting and shares the suggested solutions when possible. It also provides more insights in how the final experiment implementation was achieved.

3.6.1 Implementation of Experiment Scenario

Realizing our experiment design in miniSim required a lot of effort. The troubles were rooted in two reasons: 1) the goal of controlling all traffic conditions while not being completely monotonic, and 2) the instability of the miniSim system and its associated tools. It was not difficult to deal with this system if certain expertise was acquired, but it was a difficult start for novice users as the training and resources were quite limited.

Creating a map for miniSim: A new map could be built based on several kinds of map "tiles", where the road, buildings, side-walks and objects (e.g. traffic lights, signs, trees) were already pre-defined. This step was performed in miniSim's accessory software, Tile Mosaic Tool (TMT). Before the map could be used by the simulator, it needed to be compiled using a sequence of command line programs; the exact procedure could be found in the official manual. Successful compilation of a 3-D world was based on the the user's TMT "tiles" having types and placements that satisfy a set of internal rules, which were not clearly defined. We produced the final working result by trial and error, and did not encounter any problem with it during our experiments although there was a warning for mismatching types in TMT. This map was shared for future research experiments, and was already employed in another experiment. It is unique comparing to other existing maps (which are often in rural environment) because of its increased complexity. The motivation for creating this map is that we are concerned that driving in simple, uneventful rural roads would not induce enough workload and the participants might feel drowsy really easily.

Specifying LV's behaviours: In addition to the map files, another essential ingredient to building an experiment on miniSim were the scenario files, which allows us to configure and program dynamic or static elements in the simulator (e.g. EV, LV, other vehicles or pedestrians, road signs). They were created using Interactive Scenario Authoring Tool (ISAT). The basic building blocks of a scenario are various types of "triggers", which essentially tells an element to do something (e.g. change speed) under certain conditions. A major difficulty when implementing our design was to specify the LV to keep a constant time-gap with the participant's vehicle, which ensures the impact of the braking events would be the same. At the beginning, we used the "MaintainGap" type of trigger and specified the gap to be time-based. However in our testing, the LV still only kept a constant distance-gap regardless of our driving speed (to keep a constant time-gap, distance-gap should increase when the speed is higher). We consulted previous student who already implemented the same feature, but this implementation also failed our testing. We reached out to the miniSim technical support, and it was confirmed that setting the time-gap option was not functioning in miniSim, and the only way to achieve the effect was using a complicated work-around: directly inputting a line of code into the distance-gap field, such that the real distance-gap would be calculated based on EV's speed dynamically during run time (the idea is similar to those SQL injection attacks in computer security). The whole process of implementing this feature and debugging other small problems spanned approximately two months.

3.6.2 Redesign of Drives and Task Arrangements

In the original experiment design, there was only one baseline drive and one tasked drive. The tasked drive was a 30-minutes drive including three Critical Periods for 0-back (i.e. simply counting number of stimuli), 1-back and 2-back tasks accordingly. After working with it for several months and collected feedbacks from Stage 1 pilot testing, we decided to break the long tasked drives into several shorter drives. This benefited in several different ways:

- Allowed participants to take a break in between the drives and reset their cognitive states (i.e. help them feel refreshed)
- Reduced the cost of restarting a scenario if there was any performance problem, setup mistake or sensory failure
- Made the baseline a properer comparison: in Version 1, the baseline was not in the exact same traffic condition as other tasked Critical Periods
- Created the opportunity for collecting subjective questionnaires right after each tasked drives, which was also newly added in Version 2 design

We also increased the number of data samples collected under each task-load condition. In the Version 1 design, there was only one braking event and 4 tests under each task-load; the amount of information collected from each participant is really small, considering the length and overhead of completing an experiment. After the redesign, there were two braking events and 6 tests for each task-load. This was maximized under time constraints and requirements of intermittent breaks. Because the Version 2 design required considerably more time for each task-load condition, we decided to remove the 0-back task as the simple counting of letters was most unlikely to induce mental overload.

3.6.3 Ensuring Signal Quality

Just like the miniSim system, other parts of the data collection systems were not all user-friendly. Each system has a different set of parameters and several types of output values, which was difficult to digest for new users without prior knowledge or experience with similar technology. To successfully conduct an experiment, it not only requires proficiency at basic operations of all systems and their relay programs, but also the ability to tune the system when problems occurred. This knowledge of the systems was learnt mainly from the official manuals and sometimes experiences of previous students; however, there were still large areas left uncovered after exhausting these two resources, and external supports must be inquired. In many cases, the solutions could have been achieved a lot quicker if we proactively asked for assistance from all possible resources (especially from the developers of the systems).

Secondly, the sensors and eye-tracker are not the most robust systems, problems such as electrodes detaching, signal transmission interruptions, or signal quality degradation were all experienced at different stages of the experiment. The probability of sensory failure rises exponentially as the number of sensors increases; in this case, we have 9 sets of sensors running under 5 systems in total, which makes it a substantial problem. Although 30 to 50 minutes of the experiment time were dedicated to sensory setups, there were still cases that the experiment had to carry on without achieving good signal quality. It was difficult and demanding to detect the problems of physiological sensors based on human-reading of the signals displayed in the software at run time; most of these signals were not easily interpretable and only the most severe abnormality could be identified (such as loss of signal). The risk and associated stress were alleviated by having two investigators during each experiment.

The most severe reliability problem was experienced for the two D-Lab webcams (Logitech C920 and C930), which were originally designated for collecting videos. It randomly stopped recording in the middle of the experiments without warnings or errors (the method of checking this was found out a lot later). The problem was only discovered after the pilot testing was completed and the formal experiments already started. The unusual format of the videos exported from D-Lab made it difficult to check regularly and this problem was not spotted from earlier pilot testing files. After exhausting possible configurations of webcams and connection ports, the best result was still unreliable in the end. The cause might be the bandwidth limitation, or incompatibility issue of the drivers and the D-Lab software. At that time, unsure of the severity of the problem and under time pressure, another independent camera (GoPro) was added to ensure there would be one set of video data. It was placed at approximately the position symmetric to the webcam placed on the left-hand side of the dashboard.

3.6.4 Cognitive Task-load, Workload and Arousal

Although the experiment objective was to examine impacts of higher mental workload and focuses at the high-arousal side of the Yerkes-Dodson Curve, one persisting problem was that several participants (pilot or formal) experienced drowsiness. This could be inferred based on several pieces of evidence, such as adjusting to a more vigilant posture when the n-back task were starting, slower response to LV's braking events during baseline drives, and self-reporting. The fact that the simulator room was dark and stuffy also exacerbated this problem. The problem of drowsiness was especially prominent during the baseline drives, which was supposed to be the "optimal" operation condition corresponding to the middle part of the Yerkes-Dodson Curve (as stated in the experiment proposal and Version 1 design specifications). On the other hand, doing the n-back task might not harm the attentiveness to the driving task, because it raised the vigilant level as it added some external distraction to the task. As an analogy, in real-life, drivers often listen to radio or talk with passengers when driving under monotonic road conditions in order to keep vigilant.

Several above-mentioned design considerations were changes introduced to alleviate this problem, including 1) increasing the demand of the driving task by creating new miniSim urban map and scenarios, 2) breaking the formal drive into smaller drives with breaks in between, 3) reducing the length of non-eventful driving times, and 4) adding more diversity in warm-up drives to lead participants expecting more eventful drives. In addition, subjective questionnaires were added to collect participants' perceived workload level. While all these changes were effective, the most significant improvement was achieved by separating the longer drive into shorter ones, because the longer the participants stayed motionless the easier they would feel bored and sleepy. During the break, participants might chat with the investigators, drink some water or move around, which all helped to make them feel refreshed. Principally speaking, less actions would reduce noise for physiological sensors while less events would make it easier to minimize the variations in controlled variables. However, these would increase the chance of drowsiness.

It would be difficult to confirm the degree and direction of the introduced changes to the cognitive states because of the variations in individual sensitivities (and/or coping strategies) under the same conditions. For example, some of our participants used fingers to help counting the occurrence of n-back patterns, while some handled it purely cognitively. In practise, many previous studies made assumptions regarding the effects of varying task-load (e.g. conversation would induce cognitive distraction [3]), and usually labeled the data based on the tasks. However, most studies required a lot more interactions from the participants, which were purposely avoided in this experiment design. This made our investigation a lot less transparent and reduced participants' pressure from surveillance, thus the consistency of the modified n-back tasks' effect was of more concern. Instead of mental workload, it would be more robust to use the task-load levels as the dependent variable for this experiment. It might be also worth to re-examine the assumed “optimal” (baseline) and “degraded” (1-back or 2-back) conditions for driving performance.

3.7 Concluding Remarks

From summer 2015 to fall 2016, there were a huge amount of time and efforts put into overcoming practical difficulties, enhancing experiment mechanism and finally running this complex experiment with more than 36 participants. The objective of this experiment was quite ambitious and required considerable expertise from both the technical aspects and the experiment design aspects, which was acquired along the road.

This section of the report focused on reviewing the experiment design and implementation process comprehensively. Thus it contains large amount of details that may be only interesting in terms of the data collection process rather than the research based on the collected data. (Next section will be dedicated to the resulted dataset.)

4 Collected Data

As described in the previous section, three types of measures were used and multiple data files resulted from different data collection softwares run on four separate computers. In addition, results from subjective questionnaires had to be be exported from the survey websites. The following list summarizes different types of data and their associated signal collection methods:

- Vehicle-based data
  - miniSim log file(“DAQ” files)
- Physiological data
  - MuseLab log file (“*.muse” files)
  - D-Lab log file
- Video-based data
  - FaceLab log file
  - Video recordings from GoPro and D-Lab
- Subjective questionnaire answers:
  - NASA-TLX and Risk Perception
  - RSME

Originally, they were recorded in different format and naming convention. To ease future access and facilitate more efficient research, at the end of the data collection campaign, they were organized into folders based on data types (see above), and renamed with a consistent naming convention:

$$P[PID]\_[SID]\_[DT].*$$

The first parameter is the participant number PID = {“01”, “02”, . . . , “36”}. The second parameter denotes which scenario: SID = {“warmup”, “0”, “1”, “2”} (0 means baseline). Finally. the last parameter shows what data type it is: DT = {“miniSim”, “muse”, “dlab”, “facelab”, “gopro”}. The file extension could be “txt”, “csv” or “mat” depending on the data types. For example, the vehicle-based data (mat-file converted from MiniSim) of the 3rd participant, baseline drive would be stored in: ./Data/MiniSim/P03_0_MiniSim.mat. See Appendix A for the paths and file structure of the data storage.

Amongst all data types, only the MiniSim’s DAQ files are not readable without a special software; thus they were converted into Matlab files using the tool provided by NADS. There is no problem to read other data as text file or import into Matlab, thus the original exported data files were provided.

All data are synced using the frame number generated in NADS MiniSim, however the syncing cue differs depending on the device. This will be explained in details in following subsections.

4.1 Vehicle-based Data

NADS miniSim records the vehicle-related data. The data is stored in DAQ file format with extension .daq, which can be transformed to mat-file (.mat) using the "ndaqtool" provided by NADS. The DAQ file (and the transformed mat-file) contains all the variables that can be recorded in MiniSim, which means, even if some of the variables are not recorded, they will still be included in mat-file but the entries would be all 0 (empty). The sampling frequency of the signal is 60 Hz for continuous data; state signals will only be recorded when the state changes (CSSDS, Change State Signal Data Collection). After loading the mat-file in MATLAB, a struct variable named "daq_contents" can be found in workspace, which contains all the available variables from MiniSim. The detailed information and definition of the variables can be found in the official documents:"MiniSim Data Acquisition Cell Definitions" or "MiniSimUserGuide-v17". However, the following variables need to be paid special attention:

- The variable "frame" at the end is the frame number that will be relayed to all the other recording computers, which will be used to sync all the recorded data.
- The "SCC_LogStreams" is defined in scenarios we designed, which enables the experimenter to log certain events at certain conditions (e.g., at certain global time, location or at the start of or at the end of certain events). For our experiment, the SCC_LogStreams is defined as follows:
  - 1st column: values indicating the curve of the route, 1 means during curves and 0 means during straight lines.
  - 2nd column: values indicating if the audio (aural n-back test) is being played. Stepping from 0 to 1 means the start of the audio.
  - 3rd column: values indicating the braking behavior of the leading vehicle. Stepping from 0 to 1 means the leading vehicle starts to brake and stepping from 1 to 0 means the leading vehicle released braking pedal.

Usages.Vehicle-based measure is a direct and non-intrusive way of assessing driving performance, which has been a widely studied method to infer driver's mental workload states. For example, according to [7], vehicle speed will drop and the standard deviation of vehicle speed will increase during high mental workload periods; they also noticed higher steering wheel reversal rate. Harbluk et al. have known that there were more hard brakes (over 0.25g) when drivers were doing difficult mental tasks [2].

4.2 Physiological Data

The EEG signal was logged using the MuseLab software. The GSR, ECG and respiration were recorded using the D-Lab software. Usually, the recording devices were started before the start of each drive and stopped after each drive, which means, the length of the recordings will be longer than the drive and part of the physiological data (at the beginning and at the ending part) will not have miniSim frame number with it.

4.2.1 EEG Signals

EEG signals were relayed to MuseLab through Bluetooth transmitter on the headband and the receiver on a computer with MuseLab installed, which recorded the spontaneous electrical activity from four positions (shown in Figure 10): TP9, Fp1, Fp2 and TP10.

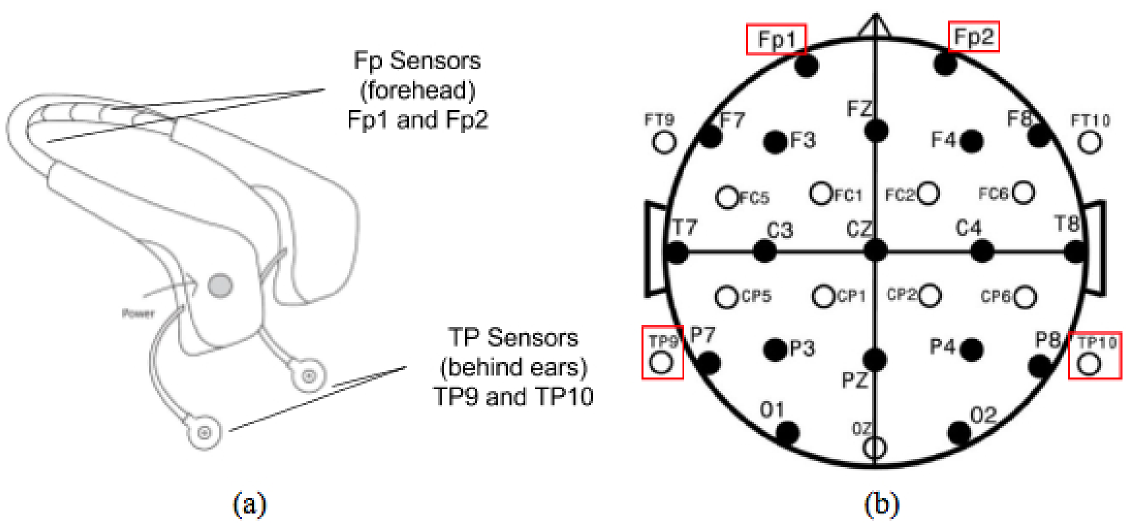


Figure 10: EEG data collection positions. To see how it is worn on a participant, refer to Figure 2. (a) The Muse EEG headband. (b) EEG positions. The red rectangles highlight the positions used in this experiment.

The Muselab is also able to do related calculations based on EEG signals and store them in the created data set, including raw FFT results, absolute and relative band power, blink counting, etc. The detailed description of the muse data can be found from `http://developer.choosemuse.com/research-tools/available-data`.To be specific, the sampling frequency of the raw EEG signals are 220Hz and the MuseLab software could provide the calculated power of low frequencies (2.5-6.1Hz), δ band (1-4Hz), θ band (4-8Hz), α band (7.5-13Hz), β band (13-30Hz) and γ band (30-44Hz) in its output. Still, if the power band with different frequency was needed, it can be calculate it from the raw EEG data. For the power calculated from MuseLab, the absolute band power is expressed as log-scale band power and the relative band powers are calculated by dividing the absolute linear-scale power in one band over the sum of the absolute linear-scale powers in all bands. The linear-scale band power can be calculated from the log-scale band power thusly:

$$linear - scalebandpower = 10^{log-scalebandpower}$$

Therefore, the relative band powers can be calculated as percentages of linearscale band powers in each band.

For example:

$$\alpha\_relative = \frac{10^{\|\alpha\|}}{10^{\|\alpha\|} + 10^{\|\beta\|} + 10^{\|\delta\|} + 10^{\|\gamma\|} + 10^{\|\theta\|}}$$

Besides EEG signals, MuseLab can also receive signals relayed from UDP or TCP ports. In this experiment, frame number from MiniSim was received from UDP ports and stored in MuseLab output data set. However, when we are dealing with the frame number in muse file or files converted from muse file, what should be kept in mind is that the frame number was created at 60 Hz, which is lower than the EEG recording frequency.

The original data file created by MuseLab is muse-file (.muse), which can be converted to CSV, mat-file or txt file or printed on screen directly using MusePlayer downloaded from `http://developer.choosemuse.com/research-tools/museplayer`. However, the converted mat-file will NOT contain frame number. If frame number is required for a specific purpose (e.g., syncing the signals with MiniSim data), the csv files (.csv) should be used, of which, each frame of data is chunked together in several continuous rows and each row represents a variable. If a variable is multidimensional, different dimensions of the variable will be stored in different columns: e.g., the raw EEG data consists of the signals from 4 positions and thus, will be shown in four columns after the second column named "/muse/eeg". The first column of each row is UTC (Coordinated Universal Time) and the slash in each variable name indicates the structure of the data set, with the variable after each slash nested in the variable before the slash. If you use the mat-file, the structure of the data set can be shown clearly from the structure of the variables.

Usages. EEG is a measurement of the electrical activity of the brain and can be employed to access the internal state of subject during cognitive tasks [21]. The power spectrum of EEG signal can provide information on mental activity, which is usually divided into five spectrums according to their frequency: delta (1-4Hz), theta (4-8Hz), alpha (7.5-13Hz), beta (13-30Hz) and gamma (30-44Hz). Many studies show that EEG is correlated with mental workload. For example, the suppression of alpha rhythm usually happens during complex and cognitively demanding tasks, concurrent tasks and when the working memory load increases [21], [22]. The change of other power spectrum waves also correspond to different mental states. For example, increased EEG Power spectra in theta wave is linked with the decrease in vigilance, the deterioration in performance or the emergence of fatigue; the increase of EEG power spectra in beta wave usually indicates increased alertness and arousal. In[26] , a standard scale of cognitive distraction was proposed based on ERP (P300), driving performance, NASA-TLX and response time and accuracy of the Detection Response Task (ISO 17488:2016). An increased P300 latency was observed with higher taskload in the laboratory. However, changing from laboratory environment to instrumented vehicle, the ERP became less sensitive because of added noise.

4.2.2 Other Physiological Data

The GSR, EEG and respiration data were recorded using the D-Lab software and exported into a single text file for each drive (baseline, 1-back and 2-back). The structure of these data is detailed in Table 4.

Table 4: D-Lab Variables and Description

Col. No.	Variable	Frequency	Description
1	MiniSim Frame Number	60Hz	Relayed from MiniSim
2	During Curve or Not	60Hz	Relayed from MiniSim, SCC_LogStream: 1: during curve; 0: during straight route
3	Audio States (n-back test)	60Hz	Relayed from MiniSim, SCC_LogStream: jumping from 0 to 1, start of the audio
4	Leading Vehicle Braking States	60Hz	Relayed from MiniSim, SCC_LogStream: jumping from 0 to 1, start of the leading vehicle brake
5	Headway	60Hz	Relayed from MiniSim, SCC_FollowInfo: unit: feet
6	ECG	240Hz	Electrical activity of the heart
7	GSR	240Hz	Electrical characteristics of the skin, unit: uSiemens
8	Respiration	240Hz	Shape change of torso while breathing

Because of the unequal sampling frequency of the variables, in the mat-file, there will be some empty values (NaN) in each variable. However, there is no time shift of different signals.

Usages.All three types of physiological measures collected in D-Lab have shown significant correlation with mental workload in previous driving studies:

- ECG

  ECG Signal can be converted into heart rate (HR) and heart rate variability (Heart Rate Variation (HRV)). Previous research indicates that the increase of mental workload is associated with the increase of HR [6][7]. Heightened levels of task difficulty result in an increase of HR in drivers, whereas fatigue/drowsiness decreases HR. However, HR is easily influenced by environment (e.g., noise, whether and etc.) and individual states (e.g., emotion).

  HRV is defined as the variation in the time interval between heartbeats, i.e., the beat-to-beat interval. In the context of driver monitoring, HRV is usually calculated by taking the Fourier transform of the HR signal, or by measuring the R peak from the ECG. According to its frequency range, there are: Very Low Frequency (VLF, $0.003 \leq 0.04Hz$), Low Frequency (LF, 0.04 Hz to 0.15 Hz) and High Frequency (HF, 0.15 Hz to 0.4 Hz) [27]. Studies in variety of areas have shown relationships between HRV and drivers' states, for example, LF and HF decreases with the increase of workload, and LF/HF increases during high mental workload periods[28].

- GSR

  GSR or Skin Conductance Level (SCL), which can provide information of changes in human sympathetic nervous system can also be used to estimate mental workload. Both on-road and simulator study have shown that GSR increases with the increase of workload when drivers are engaged in secondary mental tasks [6],[7].

- Respiration

  According to [7], increasing mental workload would result in increased respiration rate and decreased respiration depth. However, respiration is easily influenced by emotion and physical workload as well.

4.3 Video-based Data

4.3.1 Eye Tracking Data

Eye tracking data was collected using FaceLab 5.0. The pair of near-infrared cameras were placed on top of the simulator's dashboard (as shown in Figure 2). FaceLab can calculate the participant's' gaze intersection with objects defined in its "world model"; for this experiment, the world model is centered at the midpoint between two cameras, and includes planes representing left/right/centre display and dashboard. The sizes and placement were all measured in real life. For each participant, the camera model, head model and screen model were always re-calibrated according to the sitting position of the participant to ensure the data quality. However, for some participants, the eye tracker still failed to work well. These were noted in the "Experiment Log" spreadsheet.

A single FaceLab recording consists of five output files with different labels stored under the same folder (with naming convention described at the top of this section):

- The "time" file: experiment timing information.
- The "head" file: head position, rotation and other related information.
- The "eye" file: eye closure, blinking, eyeball center and raw gaze data (rotation angle).
- The "face" file: facial feature points in 3D world coordinate.
- The "world" file: gaze intersection with items defined in the world (i.e. centre screen, right/left screens or dashboard).

FaceLab has an official "Output Data Reference Guide" that explains the data that are stored in each columns of each files. Note that for most variables (e.g. blinking, gaze intersection), there are data for left eye, right eye and vergence, which could be different.

FaceLab was synced with the miniSim driving simulator, meaning that its frame number is forwarded to miniSim and stored in the DAQ files, in the field called ET_frame_num. However, it may not be very stable at the very beginning of the data recording (see an example in Figure 11), possibly caused by miniSim's initialization process: the log may have started before it started receiving eye tracking information. Thus when syncing, it is advised to look at the data after it is stabilized. It would not affect the data segments where n-back tasks played.

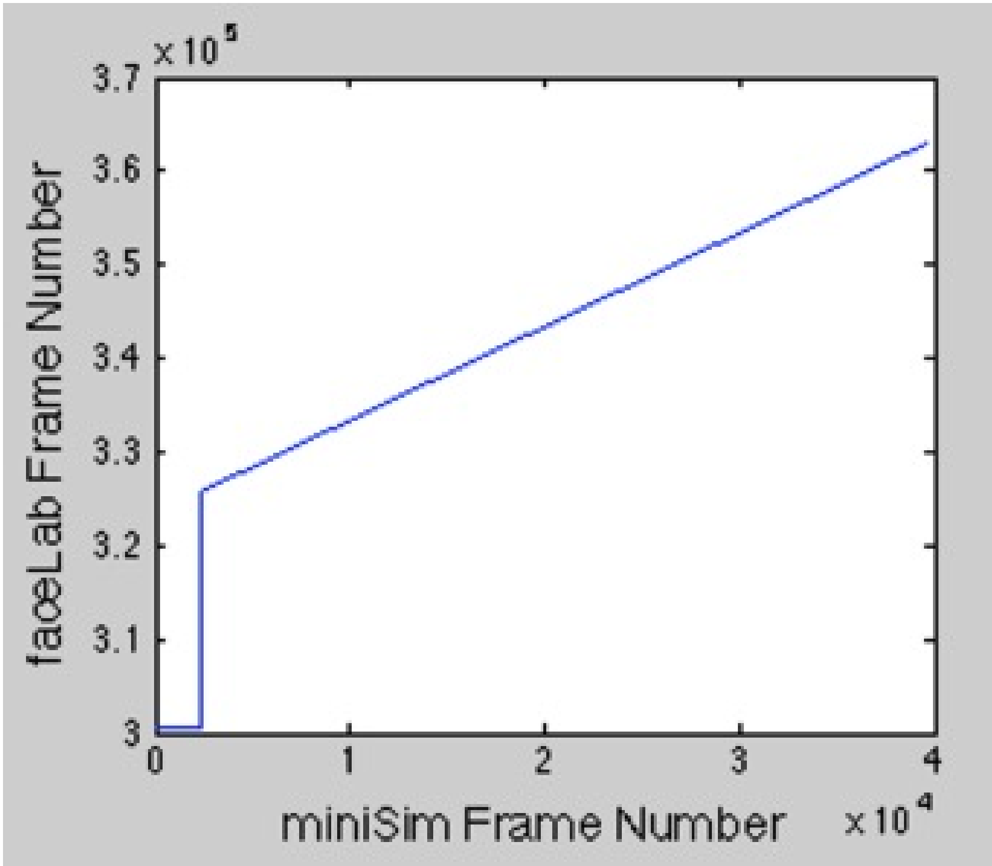


Figure 11: Example of the FaceLab frame number that was received by miniSim (this is the data of Participant 7's 2-back drive). Notice how the FaceLab frame number takes a jump near the beginning of the recording; it then becomes stabilized and increases steadily.

Usages.Eye-tracking has been a popular source of information in past driving studies. A wide range of eye-related measures for assessing mental workload were reviewed in [25]. Impact of cognitive distraction on drivers' visual behavior was studied in a naturalistic on-road experiment [2]. Drivers spent more time looking centrally ahead while visual attention to periphery area, monitoring instruments or mirrors were reduced [2]. More generally speaking, drivers experience a "visual tunneling" effect when the cognitive demand is too high [7], [8]: i.e. the visual field becomes smaller and peripheral vision is reduced. This was measured as decrease of horizontal gaze dispersion [8]. Eye blinking was also a useful cue for cognitive workload estimation, but mixed results were reported from different studies. Recarte et al. provided a possible explanation for the conflicting results: "blink inhibition for higher visual demand and increased blink rate for higher mental workload" [24]. This suggests that eye blinking may be heavily influenced by the visual demands and can not always reflect the variation of mental workload very well. Remote eye-tracking systems (such as FaceLab) were used to develop real-time cognitive distraction detection systems. These were realized by classifications based on gaze movement patterns [9] or implicit features extracted by semi-supervised machine learning algorithms [11].

4.3.2 Video Recordings

Three cameras were used in experiments: a GoPro camera that store video recording in its own SD card and two Logitech webcams (C920 and C930) that feeds video inputs into D-Lab recordings. They are approximately 27 degrees right-yaw, 12 degrees left-yaw and 12 degrees upper-pitch to the participant's face, respectively. Example frames from the camera are shown in Figure 12.

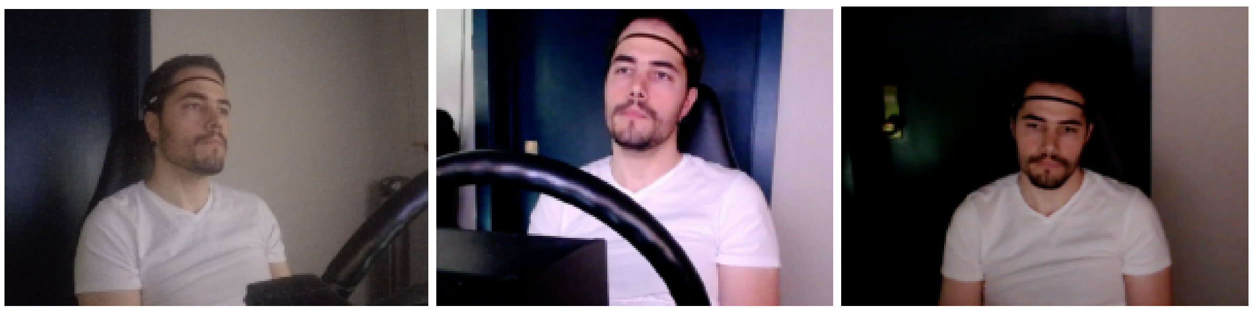


Figure 12: Video frames from three different cameras. From left to right, they are from GoPro, Logitech C930 and Logitech C920.

As mentioned above, there were some technical limitations that caused instability in D-Lab video recordings. Thus, the GoPro camera was added after the 6th participant to ensure there would be a reliable copy of video recording. The status of all recordings are reported in the file "EDREAM notes.xlsx".

The D-Lab exported videos have a peculiar encoding format that may not be recognized in many applications (e.g. Windows video players, MATLAB). An example script that uses ffmpeg[1] to convert these videos into a more readable format will be included alone with the D-Lab videos. The D-Lab videos have the following format:

- Video encoding standards = h264 (Constrained Baseline), yuv420p
- Resolution = 1920x1080, 30fps
- Audio encoding standards = N/A (no audio)

GoPro camera outputs three files for each recording with the same name and different extensions: ".mp4", ".lrv" and ".thm". The latter two files are auxiliary files generated automatically by GoPro, which are essentially a low-resolution video file and a thumbnail image file. The main MP4 videos have the following format:

- Video encoding standards = h264 (High), yuvj420p
- Resolution = 720x480, 29.97 fps
- Audio encoding standards = AAC-LC, 48000 Hz

[1]A popular cross-platform command line tool for video convertion: https://www.ffmpeg.org/

GoPro recorded at its own clock and thus was not synced with the overall system. In the later experiments (since participant 22), this problem is fixed by recording the faceLab frame number at the beginning of each recording and calculating a frame offset between these two data.

For earlier participants, the synchronization was based on identifying some audio cues exhibited by the driving simulator; script in the GoPro videos' folder will detail how this was performed. The offset values are also reported in "Video_Stats.xlsx".

Usages.Color video inputs can be analysed by computer vision algorithms to understand facial behaviours or expressions. These approaches have a lot of advantages compared to the traditional eye-tracking systems, as they require simpler hardware setup and calibration [29]. Many studies explored the usage of recent advances in face tracking and alignment to estimate drivers' gaze directions and visual attention [30, 31, 29, 32]. In an on-road study, facial action units were used as part of the multimodal signals for predicting both visual and cognitive distractions [3]; the authors obtained facial action units, coarse gaze and head movement using the CERT software [33].

4.4 Subjective Questionnaires

Subjective measures are considered to be the easiest and least expensive way of evaluating workload. But they may disrupt the continuity of the task. Three questionnaires including NASA-Task Load Index (NASA-TLX), Perceived Risk and Rating Scale Mental Effort (RSME) were conducted after warmup drive and after each formal drive (baseline, 1-back drive and 2-back drive). Participants need to fill these questionnaires according to what they feel during each drive. Instead of focusing on the secondary task, they need to report their general feeling about each drive. NASA-TLX and Perceived Risk questionnaire were conducted through online survey website and RSME was conducted individually on a PDF file, using Surface Pro3 and stylus. The files are stored as part of the data copy, please refer to Appendix A for the path.

4.4.1 NASA-TLX

The NASA-TLX is a frequently adopted subjective workload measurement in driving area. The NASA Task Load Index uses six dimensions to assess mental workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Twenty-step bipolar scales are used to obtain ratings for these dimensions. A score from 0 to 100 was calculated according to the weighting of each dimension. Pairwise comparison was performed by participants to decide the weights of each dimension: participants needed to choose the one they think were more relevant with the task they performed. The weight of each dimension was decided by how many times a dimension was regarded as more important. The detailed definition and description of each dimension were presented before the start of the questionnaire and before each time the dimension was used. Participants were told that they could ask questions regarding the questionnaire at any time.

4.4.2 Perceived Risk Questionnaire

Perceived Risk Questionnaire [8] was used to assess how risky participants feel about last drive. It requires participants to compare what they experienced in the experiment drive with a list of ten traffic events and choose the one they think that is as risky as the one (or the most closely risky one) they experienced in the last drive.

4.4.3 RSME

The RSME scale is a unidimensional scale. Ratings of invested effort are indicated by a cross on a continuous line. The line runs from 0 to 150 mm, and every 10 mm is indicated" [34]. This scale rates invested effort of the task, not explicitly mental effort. [34] found the RSME could distinguish between the "task-load situation and baseline." This does not seem to be a mainstream test because no other studies could be found that use the RSME to measure mental workload.

4.5 Complementary Materials

A comprehensive literature review on driver monitoring techniques was conducted during the earlier stage of this experiment. The background research materials and resulted paper are provided in "./Publications/IEEE SPM Literature Review"; the paper is to appear in the IEEE Signal Processing Magazine (SPM).

In order to obtain preliminary results on the acquired data, statistical analysis was performed on the first 15 participants. This work will be published during Transportation Research Board (TRB), January 2017. Please find the most updated version of the paper under "./Publications/TRB Paper". MatLab scripts for this analysis are stored under "./Data_Analysis".

Please see Appendix A for paths accessing the experiment data, and the above complementary materials.

References

[1] T. A. Dingus, S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta *et al.*, "The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment," Tech. Rep., 2006.

[2] J. L. Harbluk, Y. I. Noy, P. L. Trbovich, and M. Eizenman, "An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance," *Accident Analysis & Prevention*, vol. 39, no. 2, pp. 372–379, 2007.

[3] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, 2015.

[4] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 596–614, 2011.

[5] B. Mehler, B. Reimer, and J. Dusek, "Mit agelab delayed digit recall task (n-back)," *Cambridge, MA: Massachusetts Institute of Technology*, 2011.

[6] B. Mehler, B. Reimer, J. Coughlin, and J. Dusek, "Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2138, pp. 6–12, 2009.

[7] B. Mehler, B. Reimer, and J. F. Coughlin, "Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task an on-road study across three age groups," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 54, no. 3, pp. 396–412, 2012.

[8] B. Reimer, "Impact of cognitive task complexity on drivers' visual tunneling," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2138, pp. 13–19, 2009.

[9] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 340–350, 2007.

[10] Y. Liang, J. Lee, and M. Reyes, "Nonintrusive detection of driver cognitive distraction in real time using bayesian networks," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2018, pp. 1–8, 2007.

[11] T. Liu, Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1108–1120, 2016.

[12] S. Schneegass, B. Pfleging, N. Broy, F. Heinrich, and A. Schmidt, "A data set of real world driving to assess driver workload," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* ACM, 2013, pp. 150–157.

[13] J. F. Coughlin, B. Reimer, and B. Mehler, "Monitoring, managing, and motivating driver safety and well-being." *IEEE Pervasive Computing*, vol. 10, no. 3, 2011.

[14] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of comparative neurology and psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[15] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering psychology & human performance.* Psychology Press, 2015.

[16] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage Publications, 2006, pp. 904–908.

[17] A. Čolić, O. Marques, and B. Furht, *Driver Drowsiness Detection: Systems and Solutions.* Springer, 2014.

[18] D. Novak, "Engineering issues in physiological computing," in *Advances in Physiological Computing.* Springer, 2014, pp. 17–38.

[19] H.-B. Kang, "Various approaches for driver and driving behavior monitoring: a review," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 616–623.

[20] M.-H. Sigari, M.-R. Pourshahabi, M. Soryani, and M. Fathy, "A review on driver face monitoring systems for fatigue and distraction detection," *International Journal of Advanced Science and Technology*, 2014.

[21] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.

[22] K. Ryu and R. Myung, "Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic," *International Journal of Industrial Ergonomics*, vol. 35, no. 11, pp. 991–1009, 2005.

[23] K. Young, M. Regan, and M. Hammer, "Driver distraction: A review of the literature," *Distracted driving*, pp. 379–405, 2007.

[24] M. Á. Recarte, E. Pérez, Á. Conchillo, and L. M. Nunes, "Mental workload and visual impairment: Differences between pupil, blink, and subjective rating," *The Spanish journal of psychology*, vol. 11, no. 02, pp. 374–385, 2008.

[25] G. Marquart, C. Cabrall, and J. de Winter, “Review of eye-related measures of drivers’ mental workload,” *Procedia Manufacturing*, vol. 3, pp. 2854–2861, 2015.

[26] D. L. Strayer, J. Turrill, J. M. Cooper, J. R. Coleman, N. Medeiros-Ward, and F. Biondi, “Assessing cognitive distraction in the automobile,” *Human factors*, vol. 57, no. 8, pp. 1300–1324, 2015.

[27] T. F. of the European Society of Cardiology *et al.*, “Heart rate variability standards of measurement, physiological interpretation, and clinical use,” *Eur heart J*, vol. 17, pp. 354–381, 1996.

[28] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, “The effect of mental stress on heart rate variability and blood pressure during computer work,” *European journal of applied physiology*, vol. 92, no. 1-2, pp. 84–89, 2004.

[29] E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 11, no. 2, pp. 300–311, 2010.

[30] P. Jimenez, L. M. Bergasa, J. Nuevo, N. Hernandez, and I. G. Daza, “Gaze fixation system for the evaluation of driver distractions induced by ivis,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1167–1178, 2012.

[31] R. Oyini Mbouna, S. G. Kong, and M.-G. Chun, “Visual analysis of eye state and head pose for driver alertness monitoring,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 3, pp. 1462–1469, 2013.

[32] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi, “Driver gaze tracking and eyes off the road detection system,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2014–2027, 2015.

[33] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, “The computer expression recognition toolbox (cert),” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on.* IEEE, 2011, pp. 298–305.

[34] D. De Waard, *The measurement of drivers’ mental workload.* Groningen University, Traffic Research Center Netherlands, 1996.