Real-time Human Workload Estimation and Its Application in Adaptive Haptic Shared Control

by

Ruikun Luo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Robotics)
in The University of Michigan
2021

Doctoral Committee:

Assistant Professor Xi Jessie Yang, Chair
Associate Research Scientist Tulga Ersal
Assistant Professor Maani Ghaffari
Professor Nadine Sarter
Professor Jeffrey Stein

Ruikun Luo

ruikunl@umich.edu

ORCID iD: 0000-0003-3310-6381

DEDICATION

For my mom and dad, who always love me and support me.

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to the many people who have made it possible for me to complete this dissertation.

First and foremost, I would like to thank my advisor, Dr. X. Jessie Yang, for all the guidance, support, and encouragement throughout this dissertation. I am honored and extremely lucky to have been picked up by you in the middle of my Ph.D. journey. Thank you for leading me step by step to be an independent researcher. I am incredibly grateful for all your support, trust, and patience.

I would also like to thank the members of my dissertation committee: Dr. Jeffrey Stein, Dr. Tulga Ersal, Dr. Nadine Sarter, and Dr. Maani Ghaffari. I truly appreciate your valuable time and feedback throughout this dissertation.

I am also grateful for my previous advisor, Dr. Dmitry Berenson, who brought me to the University of Michigan. Thank you for your support and for understanding my decision to pursue my interest.

I am honored to have worked with the extraordinary faculty collaborators Dr. Tulga Ersal and Dr. Jeffrey Stein. I have always been inspired by your critical thinking, valuable suggestions, and kindness to students. I also appreciate my student collaborators Yifan Weng, Na Du, Yaohui Guo, Yifan Wang, Kevin Yiwei Huang, Christopher Schemanske, Sabrina Benge, and Jian Chu.

I am very grateful for the financial support I received from the University of Michigan Department of Industrial and Operations Engineering, the Automotive Research Center,

the U.S. Army CCDC Ground Vehicle Systems Center, the Rackham Graduate School, and the National Science Foundation. Their support helped me to finish my Ph.D. and this dissertation.

I am also thankful for the support I received from the staff members in the Department of Industrial and Operations Engineering and Robotics Institute. I would like to thank Christopher Konrad and Olof (Mint) Minto for their assistance with technical issues throughout my research. Thanks also go to Tina Picano Sroka, Teresa Maldonado, Wanda Dobberstein, Denise Edmund, and Damen Provost for all the contributions you have made to making my Ph.D. journey smooth and easy.

I would like to thank my labmates in the ICRL lab – Na, Taa, Yaohui, Qiaoning – and in the ARM lab - Rafi, Yu-chi, Calder, Dale, and Brad. Thank you for your support and fantastic memories. Special thanks to my friends Changshuo, Xiaofan, Jing, Wenzhen, Fanyi, Yu-chi, Qiwen, and Kaiwen for helping me get through the struggling moments.

Finally, I would like to thank my parents for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Automated vehicles (AVs) are promising to have the potential to reduce driving-related injuries and deaths. However, autonomous driving technology is currently limited in its scope and reliability, giving rise to the semi-autonomous driving model, where the autonomy and the human share the control of the vehicle. Workload, despite being an important human factor, has not yet been considered when designing adaptive shared control.

Workload is a critical human factor in human-automation interaction. Researchers have shown an inverted-U relationship between workload and human performance. When a person's workload is too low, s/he may experience vigilance decrement, leading to suboptimal task performance. On the contrary, when a person is overloaded, s/he may not have enough resources to complete a task successfully. When a person experiences moderate workload, s/he achieves optimal performance. However, the effects of adapting to the human workload in the haptic shared control of ground vehicles remain unclear.

Human workload can be measured offline or online. Offline measures are assessed after a human operator finishes a task, typically by using a questionnaire or post-hoc analyses. However, offline measures are not applicable for designing real-time adaptive systems. There are different ways to measure human workload online, including using behavioral measures, secondary task performance, and physiological measures. Using behavioral measures and secondary tasks to measure workload online usually relies on real-time behavioral data and task performance, which is sometimes unavailable. Therefore, researchers have developed methods to assess workload using physiological measurements, including heart rate variability and pupil size. Among these measures, some are intrusive,

such as electroencephalography (EEG). Eye-tracking devices, as a non-intrusive (or less-intrusive) technology, have been increasingly used to assess operators' workloads. Studies unitizing eye trackers to examine workload can be broadly categorized into two groups. Previous studies largely adopted statistical methods to show the relationships between certain eye-related measurements and workload. For example, greater pupil dilation indicates higher mental workload.

Recently, researchers have started to apply machine learning techniques to classify mental workload into different levels. However, most of these studies have adopted either a single-model-single-feature approach or a single-model-all-features approach. In the single-model-single-feature approach, researchers developed a single machine learning model for a single feature (i.e., the Hidden Markov Model for gaze trajectory). As the single-model-single-feature approach only uses one type of measurement, useful information from other measurements is missed. In the single-model-all-features approach, researchers attempt to utilize information from different measurements and apply a single machine learning model for different features by concatenating different features into one feature vector. However, different machine learning models are suitable for different features (e.g., the Hidden Markov Model works for gaze trajectory, whereas support-vector machines work for changes in pupil size). Therefore, combining these features with a single-model-all-features approach can lead to sub-optimal performance for workload estimation.

To address these shortcomings and research gaps, the goals of this dissertation were to (1) examine whether and to what extent haptic shared control performance can be improved by incorporating operators' workload; (2) develop a computational model for workload estimation, and the model should be able to leverage different machine learning models that work best for different features; and (3) investigate the generalizability of the

workload estimation model. To address these research goals, this dissertation was composed of four research phases with two pilot studies and four human subject experiments. The four phases were as follows:

(1) Collaborating with Yifan Weng, Dr. Tulga Ersal, and Prof. Jeffrey Stein from the Department of Mechanical Engineering at the University of Michigan, we developed a teleoperated dual-task shared control simulation platform where the human shared control of a ground vehicle with autonomy while performing a surveillance task simultaneously. In addition, we developed a real-time eye-tracking system based on Tobii Pro Glasses 2 to measure the human gaze points in a world frame and pupil sizes. We conducted two pilot studies with 16 participants. Pilot Study 1 selected tracks with similar difficulties to be used in the simulation platform. We determined the surveillance task urgency to impose different human workload levels in Pilot Study 2. We used this dual-task platform to represent a human-automation interaction system.

(2) We proposed a workload-adaptive haptic shared control scheme together with our collaborators. We conducted two human subject experiments during this phase. We collected the human eye-related data to a build workload estimation model in Experiment 1, which was conducted with 12 participants. In Experiment 2, we examined the effects of the workload-adaptive haptic shared control scheme. The results indicated that the proposed workload-adaptive haptic shared control scheme can reduce human workload, increase human trust in the system, increase driving performance, and reduce human effort without sacrificing surveillance task performance.

(3) We proposed a Bayesian inference model for workload estimation that can leverage the different machine learning models that work best for different features. Specifically, we used support-vector machines (SVMs) for pupil size change, the Hidden Markov Model (HMM) for gaze trajectory, SVMs for fixation feature, and Gaussian Mixture Mod-

els (GMMs) for fixation trajectory. Using the data from Experiment 1 and an additional 12 participants, the empirical results indicated that our proposed model achieved a 0.82 F_1 score for workload imposed by varying surveillance task urgency.

(4) We investigated the generalizability of our proposed Bayesian inference model for workload estimation by conducting two human subject experiments with 24 participants and using different factors to impose human workload. In Experiment 3, we introduced obstacles to the driving task and manipulated the obstacle headway to impose human workload. The results indicated that our proposed model achieved a 0.68 F_1 score for the workload imposed by obstacle avoidance. In Experiment 4, we manipulated driving speed to impose human workload. The results showed that the personalized version of our proposed model can distinguish the workload imposed by different driving speeds under high surveillance task urgency.

CHAPTER I

Introduction

1.1 Shared Control for Semi-autonomous Vehicles

Automated vehicles (AVs) are promising to provide fuel-efficient driving (Chen et al., 2019) and have the potential to reduce driving-related injuries and deaths (Eby et al., 2016). However, autonomous driving technology is currently limited in its scope and reliability, giving rise to the semi-autonomous driving mode. In this mode, the autonomy and a human share control of the driving task. Therefore, properly allocating the control authority between these two agents is critical for safety and team performance.

In cooperative shared control, both agents can affect the final control input. One type of co-operative shared control directly blends the steering angle inputs from both the human and the autonomy through a designed arbitrator (Anderson et al., 2011). This scheme closes the loop between the human and the autonomy after the steering wheel (i.e., the human will be able to feel the impact of the autonomy input only after the resultant steering command takes effect and through the response of the vehicle). The other type of cooperative shared control is haptic shared control, in which the human and autonomy can negotiate the steering angle through the torques they apply to the steering wheel (Griffiths and Gillespie, 2005; Petermeijer et al., 2015; Nguyen et al., 2018). In this scheme, the human operator can directly feel the torque from the autonomy and can choose to yield

to or fight it by exerting extra torque on the steering wheel. Researchers have developed and tested a haptic shared control framework, and the results showed that haptic control improved driving performance while reducing visual demand or shortening the reaction time of the secondary task (Griffiths and Gillespie, 2005). Others have used the haptic control framework with a bandwidth guidance version and a continuous guidance version, and the results showed that both helped reduce driver errors (Petermeijer et al., 2015).

The impedance of autonomy in a haptic shared control scheme can be considered a natural tuning parameter through which adaptability can be introduced. Indeed, even though earlier haptic shared control schemes used a fixed impedance (Griffiths and Gillespie, 2005; Mulder et al., 2008), later works started investigating adaptive impedance schemes. Some schemes adopt vehicle-performance-based switching rules as adaptation mechanisms, such as turning shared control on when the lateral error of the vehicle exceeds a designed threshold (Petermeijer et al., 2015). Others consider human-performance-based metrics, such as the human input torque and attention, as guidelines for designing control authority allocation to adapt to impedance continuously (Nguyen et al., 2018). However, workload, an important human factor, has not yet been considered for adaptation purposes.

1.2 Workload in Human-automation Interaction

The concept of workload can be most intuitively understood in terms of the ratio of the time required to do tasks to the time available to do them (Hendy et al., 1997; Parks and Boucek, 1989). More generally, mental workload can be defined as the ratio of the mental resources required to the resources available (Lee et al., 2017). Similarly, Parasuraman et al. (2008) defined mental workload as "the relation between the function relating the mental resources demanded by a task and those resources available to be supplied by the human operator."

Workload is a critical human factor in human-automation interaction. Automation is often introduced to reduce a human operator's workload. For example, an automated lane-keeping system reduces the driver's workload (Hancock et al., 1996). However, improperly using automation may increase human workload, which is against its aim. Recently, researchers found that a driver's self-reported workload increased when using Tesla's assistance driving system, which was designed to reduce the driver's workload (Stapel et al., 2019). Human workload may also influence the joint performance of the human and automation. For example, a high workload led to lower takeover readiness and worse performance when drivers were operating a conditionally automated vehicle (Du et al., 2020b).

Researchers have shown an inverted-U relationship between human performance and workload based on Yerkes Dodson's law, as shown in Figure 1.1 (Samms and Mitchell, 2010; Seong et al., 2013; Schutte, 2015; Zenati et al., 2020). When a person's workload is too low, s/he may experience vigilance decrement, leading to sub-optimal task performance (Lee et al., 2017). When a person is overloaded, s/he may not have enough resources to achieve the task, resulting in sub-optimal performance and even task failures (Lu et al., 2019). For example, driving under a high workload can lead to more driver errors (Hancock et al., 1990; Briggs et al., 2011), which are the main cause of 45% to 75% of all crashes (Wierwille et al., 2002). When a person experiences moderate workload, s/he achieves optimal performance (De Waard, 1996).

Specifically, in the context of transportation, drivers' workload can be affected by different factors, such as driver's behaviors and surrounding environment. Ma and Kaber (2005) found that different activities affected driver workload, i.e., usage of adaptive cruise control reduced driver workload, however, cell phone conversation increased driver workload. In addition, driver workload increased while performing turn maneuvers (Hancock et al., 1990), driving on road with small curve radius (Tsimhoni and Green, 1999), and ex-

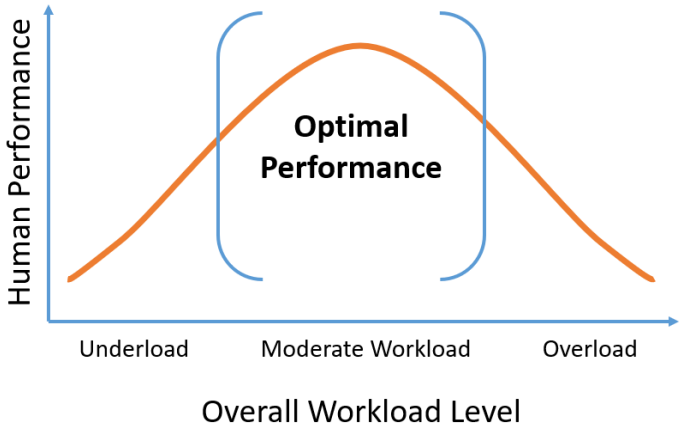


Figure 1.1: Relationship between human performance and workload (Samms and Mitchell, 2010; Seong et al., 2013; Schutte, 2015; Zenati et al., 2020).

periencing demanding surrounding traffic (Jahn et al., 2005; Patten et al., 2006; Teh et al., 2014). Researchers also found relationships between weather and driver workload, i.e., wind gust (Hicks and Wierwille, 1979) at the front of the vehicle and fog (Hoogendoorn et al., 2011) increased driver workload.

1.3 Workload Measurement

There are different ways to measure human workload, such as behavioral measures, secondary tasks, subjective measures, and physiological measures.

1.3.1 Behavioral Measures

In many manual driving tasks, researchers have used control behaviors and driving performance to estimate human workload. For example, Zhang et al. (2008) used the means and standard deviations of vehicle velocity, vehicle lane position, steering angle, and vehicle acceleration to estimate human workload via decision trees in a simulation study. Similarly, other researchers have used longitudinal performance measures (e.g., standard deviations of speed and acceleration) and lateral performance measures (e.g., standard deviations of lateral position and steering wheel angle) to measure drivers' workload (Hicks

and Wierwille, 1979; Lansdown et al., 2004; Liu, 2019). In a field study, Xing et al. (2018) used the signals (e.g. latitude, longitude, altitude, and speed) from a global positioning system (GPS) to estimate human workload.

Theoretically, behavioral measures can be online but require real-time behavioral data, which is sometimes unavailable. Therefore, previous studies have mainly used behavioral measures as offline measures by post-hoc analyses (Xing et al., 2018). Due to their offline nature, offline measures are not applicable to designing real-time adaptive automation systems (Wickens et al., 2015).

1.3.2 Secondary Tasks

Another way to measure human workload is by introducing a secondary task and measuring its performance. The logic behind this is that if a human performs the primary task at an adequate level, how well s/he performs the secondary task indicates how much residual capacity is available. Various secondary tasks have been used in previous literature, such as the n-back memory task (Owen et al., 2005; Herff et al., 2014; Lu et al., 2019) and the Detection Response Task (DRT) (ISO/TC 22/SC 39, 2016; Chang et al., 2017; Miller et al., 2018).

In the n-back memory task, a human operator is presented with a series of stimuli and must identify if the current stimuli are the same as the stimuli presented n trials ago. The stimuli can be either auditory or visual. Different types of stimuli have been used, including the sounds of letters (Lu et al., 2019), shapes (Cohen et al., 1994), pictures (Kim et al., 2002), and the locations of pictures (Du et al., 2020d,c).

The DRT is a standard workload assessment method in ground transportation (ISO/TC 22/SC 39, 2016). A human operator receives a sequence of stimuli (visual or tactile) and responds (i.e., clicks a micro switch) immediately after receiving the stimuli. The hit rate

and response time can indicate human workload (Chang et al., 2017).

If designed well, secondary tasks can have high fidelity (Lee et al., 2017). However, secondary tasks are designed to probe the residual capacity for the human and are not used for a primary task (Wickens, 2008). In addition, people can pay different levels of attention to secondary tasks (Verwey, 2000) and, hence, influence primary task performance. Therefore, secondary tasks are not applicable for adaptive automation systems.

1.3.3 Subjective Measures

Subjective measures are usually offline and are widely used for workload assessment to evaluate systems. The measurement is commonly done via questionnaires after a human operator finishes a task. Subjective measures can be multidimensional (e.g., NASA Task Load Index (NASA TLX) (Hart and Staveland, 1988) and Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988)) and unidimensional (e.g., Rating Scale Mental Effort (RSME) (Zijlstra and Van Doorn, 1985), activation scale (De Waard, 1996), and anchored rating (Schweitzer and Green, 2006)). The SWAT contains three dimensions: time load, mental effort load, and physiological stress load (Reid and Nygren, 1988). The NASA TLX asks human operators to provide separate subjective ratings on six subscales of mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart and Staveland, 1988). The NASA TLX has been widely used in previous studies. For example, researchers used the NASA TLX to measure a driver's workload during left-turn maneuvers (Hancock et al., 1990) and when answering mobile phone calls (Alm and Nilsson, 1995). Due to their offline nature, subjective measures are usually not applicable to adaptive automation systems.

Table 1.1: Summary of physiological sensors

Physiological Measures	Physiological Sensors	Intrusiveness
Brain signals	Electroencephalogram (EEG)	Intrusive
	Functional near-infrared spectroscopy (fNIRS)	Intrusive
Galvanic skin responses (GSR)	GSR	Intrusive
Heart rate indices	Electrocardiogram (ECG or EKG)	Intrusive
	Optical sensors (PPG) - finger/earlobe	Intrusive
	Optical sensors (PPG) - wrist	Nonintrusive
Eye-related measures	Eye-trackers	Nonintrusive

1.3.4 Physiological Measures

Physiological measures rely on changes in human physiological signals (Lee et al., 2017). Researchers have found that various types of physiological signals indicate human workload changes, such as brain signals, galvanic skin response (GSR), heart rate indices, and eye-related measures (Heard et al., 2018). Table 1.1 summarizes the physiological sensors for different physiological measurements and their intrusiveness.

Brain Signals

Researchers have used different brain signals to measure human workload, such as electroencephalogram (EEG) (Sterman and Mann, 1995; Hankins and Wilson, 1998; Borghini et al., 2014; Diaz-Piedra et al., 2020) and functional near-infrared spectroscopy (fNIRS) (Hirshfield et al., 2009; Ayaz et al., 2012). EEG measures electrical activity in the human brain by attaching electrodes to the scalp. Diaz-Piedra et al. (2020) measured army combat drivers' mental workload on different terrains. Borghini et al. (2014) reviewed previous studies of using EEG to assess aircraft pilots' and vehicle drivers' mental workload and showed that the increment of EEG power in the theta band and decrement in the alpha band indicated high mental workload. Ayaz et al. (2012) used fNIRS, which measures brain

activity, to obtain functional neuroimaging to assess human operators' mental workload while performing air traffic control tasks. Both EEG and fNIRS are considered intrusive measurements, as they require connecting many wires to the head. Therefore, they are sometimes not acceptable for adaptive automation systems.

Galvanic Skin Response

Galvanic skin response (GSR), also known as electrodermal activity (EDA), reflects variations in the electrical characteristics of the skin (e.g., skin conductance). Raw GSR signals contain two types of skin conductance: tonic and phasic. Tonic skin conductance, also referred to as skin conductance level (SCL), is the baseline level of skin conductance. Phasic skin conductance, referred to as skin conductance response (SCR), changes as humans perform tasks. GSR has been used to assess human workload (Shi et al., 2007; Reimer et al., 2009; Nourbakhsh et al., 2012; Schneegass et al., 2013). For example, Reimer et al. (2009) manipulated drivers' workload by different levels of auditory n-back memory tasks and showed that skin conductance increased dramatically when a 0-back memory task was first performed. Schneegass et al. (2013) measured drivers' SCR in real-world driving under different traffic conditions and measured subjective workload using a post-hoc video rating. They found that both the workload and SCR were high in the 30 km/h zone. Nourbakhsh et al. (2012) measured human cognitive load in arithmetic and reading tasks using both the temporal and spectral features of GSR. As GSR is usually measured by skin surface electrodes attached to the fingers, hands, and feet, it is also considered an intrusive measurement. Hence, GSR not acceptable for adaptive automation systems.

Heart Rate Indices

Heart rate and heart rate variability (HRV) are sensitive measures of human workload (Aasman et al., 1987; Vicente et al., 1987; Jorna, 1993; Hankins and Wilson, 1998; Backs et al., 2003; Reimer et al., 2009). For instance, Jorna (1993) assessed pilots' workload using heart rate and heart rate variability. In addition, Backs et al. (2003) showed that heart rate had a consistently significant association with visual demand, an indicator of visual mental workload (Tsimhoni et al., 1999), when driving a simulated vehicle. They found that heart rate increased when visual mental workload increased. These studies showed the pattern that heart rate increased and heart rate variability decreased when human workload increased (Heard et al., 2018). However, this pattern did not hold for all people. Reimer et al. (2011) studied the impact of a naturalistic hands-free mobile phone call task on heart rate and driving performance in two age groups: young adults (19–23) and late middle-aged adults (51–66). They found that for young adults, heart rate accelerated when answering the phone. However, the late middle-aged adults did not illustrate this pattern.

Heart rate can be measured using different types of sensors, such as electrocardiogram (ECG or EKG) and optical sensors (PPG). ECG records the electrical signals of the heart and is intrusive, as the electrodes are normally attached to the torso. PPG records blood volume changes to track heart rate. PPG can be both intrusive (attached to fingers and earlobes) and non-intrusive (attached to wrist). However, the PPG on the wrist can be easily affected by hand movement and, therefore, is not suitable for the driving task (Chen et al., 2015).

Eye-related Measures

Eye-tracking devices, as a non-intrusive (or less-intrusive) technology, have been increasingly used to assess operators' workload (Moacdieh et al., 2020). Previous stud-

ies have largely focused on statistical analysis to show that certain physiological signals significantly change under different workload conditions (Palinko et al., 2010; Demberg, 2013; Kun et al., 2013). Different types of eye-related measurements have been investigated in the previous literature, as shown in Table 1.2. These eye-related measurements can be categorized into three groups: 1) pupil-related measures, 2) blink-related measures, and 3) gaze-related measures.

1) Pupil-related measures. Pupil diameters are widely used to assess human workload (Palinko et al., 2010; Demberg, 2013; van der Wel and van Steenbergen, 2018). Researchers have found that pupil diameter, pupil diameter change, and pupil diameter change rate increased under high workload in different scenarios (van der Wel and van Steenbergen, 2018; Palinko et al., 2010). Pupil diameter change is the difference between people's pupil diameter and the baseline pupil diameter, and pupil diameter change rate is the first order derivative of pupil diameter over time. In an air traffic control task, Ahlstrom and Friedman-Berg (2006) found that the operators' mean pupil diameter was significantly larger when using a static storm forecast tool than when using a dynamic storm forecast tool, which indicated a higher workload with static storm forecast tools. Furthermore, Kun et al. (2013) found that a driver's pupil diameter change increased when first preparing to answer a question when driving on both straight and curvy roads. Similarly, Klingner et al. (2008) found that pupil diameter change increased under high workload during three different standard tasks: mental arithmetic tasks, short-term memory tasks (memorizing and repeating a sequence of digits), and aural vigilance task (identifying the misspoken digit in a sequence of numbers). These results were consistent with Ahern and Beatty (1979) and Kahneman and Beatty (1966). Palinko et al. (2010) found that the mean pupil diameter change rate was sensitive to cognitive load changes due to engaging in a dialogue during the driving task.

Table 1.2: Eye-related measures indicate workload

Metric	Reference
Pupil diameter	Recarte and Nunes (2000, 2003); Ahlstrom and Friedman-Berg (2006); Recarte et al. (2008); Vogels et al. (2018); Kahneman and Beatty (1966)
Pupil diameter change	Ahern and Beatty (1979); Backs and Walrath (1992); Klingner et al. (2008); Palinko et al. (2010); Benedetto et al. (2011); Palinko and Kun (2011); Kun et al. (2013)
Pupil diameter change rate	Palinko et al. (2010)
ICA (frequency of rapid pupil dilation)	Marshall (2000, 2002); Demberg (2013); Rerhaye et al. (2018); Vogels et al. (2018)
Blink duration	De Waard (1996); Van Orden et al. (2001); Ahlstrom and Friedman-Berg (2006); Benedetto et al. (2011)
Blink rate	De Waard (1996); Van Orden et al. (2001); Tsai et al. (2007); Recarte et al. (2008); Benedetto et al. (2011)
Blink latency	Eggemeier et al. (1990); Carmody (1994)
Fixation frequency	Backs and Walrath (1992); Van Orden et al. (2001)
Fixation duration	Rayner and Morris (1990); Backs and Walrath (1992); Recarte and Nunes (2000); Li et al. (2012); Marquart et al. (2015)
Variability of fixation duration	Recarte and Nunes (2000)
Variability of fixation position	Recarte and Nunes (2000); Reimer (2009)
Percentage of fixations in area of interest (AOI)	Recarte and Nunes (2000)
Saccadic extent	May et al. (1990); Recarte and Nunes (2000); Van Orden et al. (2001)
Saccadic amplitude	Moacdieh et al. (2020)
NNI (Nearest Neighbor Index)	Di Nocera et al. (2007)
Spatial density	Moacdieh et al. (2020)
Stationary entropy	Moacdieh et al. (2020)
Scanpath length	Moacdieh et al. (2020)
Transition rate	Moacdieh et al. (2020)

Instead of directly using pupil diameter, pupil diameter change, and pupil diameter change rate, researchers have developed the Index of Cognitive Activity (ICA) by applying a wavelet decomposition to the pupil diameter signal to calculate the frequency of

rapid pupil dilations (i.e., average number of abrupt discontinuities in pupil diameter per second) (Marshall, 2000, 2002). The ICA has been used as a general index for human workload, and higher ICA indicates higher cognitive workload (Demberg, 2013; Rerhaye et al., 2018; Vogels et al., 2018). Vogels et al. (2018) compared the ICA and pupil diameter of 32 participants during two dual-task scenarios: a language comprehension task with a memory task and a driving task with a memory task. In the language comprehension task, participants responded to comprehension questions after listening to some sentences. In the driving task, participants controlled the steering wheel to trace a moving target. In the memory task, participants were asked to recall the speed limit signs they had just passed. The results showed that people have higher ICA when performing more difficult language comprehension tasks. However, the results for the memory task were contradictory; with the more difficult memory task, ICA significantly decreased, where as pupil size significantly increased. This indicated that pupil diameter and ICA reflected different cognitive and neuronal processing in dual-task scenarios.

2) Blink-related measures. Different blink-related measures have been investigated in the previous literature, such as blink duration, blink rate, and blink latency (De Waard, 1996; Marquart et al., 2015; Heard et al., 2018). Blink duration is the length of a blink, and it decreases under high workload (Ahlstrom and Friedman-Berg, 2006). Blink rate, also called blink frequency, is the number of blinks per minute. Recarte et al. (2008) investigated human blink duration and blink rate under different cognitive tasks (listening, talking, and calculating) and visual demand (with visual search or without visual search). The results indicated that blink duration decreased as cognitive workload increased or visual demand increased. However, blink rate decreased for higher visual workload and increased for higher mental workload. Benedetto et al. (2011) found that blink duration is more sensitive and reliable than blink rate for measuring a driver's visual workload in a

simulated driving experiment. Blink latency is the time between consecutive blinks, which increases as cognitive and visual workload increases (Eggemeier et al., 1990; Carmody, 1994).

3) Gaze-related measures. Gaze-related measures are mainly based on fixation and saccades, the two phases of human eye movement. Fixations are the phases when humans maintain their gaze points at a location for a time period and gather new information from the area they are examining (Jacob, 1995; Rayner, 1995, 2009). Saccades are the rapid eye movements between fixations (Jacob, 1995; Salvucci and Goldberg, 2000; Jacob and Karn, 2003). The metrics computed from fixation and saccades can be roughly categorized into two groups: temporal information and spatial information (Marquart et al., 2015). The temporal information group includes fixation duration and fixation frequency (i.e., number of fixations in one minute). Both fixation duration and fixation frequency increase when a person experiences a high workload (Rayner and Morris, 1990; Backs and Walrath, 1992; Recarte and Nunes, 2000; Van Orden et al., 2001; Marquart et al., 2015). For instance, Backs and Walrath (1992) found that when searching for information on a symbolic display, people had longer fixation duration and a higher number of fixations when using a color-coded display instead of a monochrome display. The spatial information group includes different measures to describe gaze distribution. For example, Recarte and Nunes (2000) investigated different fixation-related measurements when drivers perform mental tasks (verbal or spatial imagery) while driving on highways or roads. They found that gaze distribution decreased when mental tasks were performed, and they used metrics like variability of fixation position, percentage of fixations in an area of interest (AOI), and saccadic size (i.e., range of saccadic extent). Similarly, Moacdieh et al. (2020) also found gaze distribution decreased under high workload, and they used metrics like spatial density, stationary entropy, saccadic amplitude, scanpath length per second, and transition

rate. Di Nocera et al. (2007) proposed the Nearest Neighbor Index (NNI) to measure the spatial dispersion of eye fixations, which is the ratio between the average of the minimum distances between fixation points and the mean random distance, if one would expect the distribution to be random. In a simulated flight, Di Nocera et al. (2007) showed pilots had larger NNI when landing and departure than when cruising.

Researchers recently started to use machine learning techniques to estimate human workload by modeling the workload estimation problem as a supervised classification problem (Heard et al., 2018). Given a sequence of physiological signals from a human operator, they first extract a feature vector from the signals and then classify this feature vector into different classifiers representing different workload levels. Previous studies have investigated different machine learning models for different eye-related features, as shown in Table 1.3. The last column of Table 1.3 shows the different evaluation methods in the previous studies. "Within-participants" means that the results are evaluated using the training data and testing data from the same participant. "Cross-participants" means that the training data and testing data are from different participants.

Kosch et al. (2018a) used SVMs with a linear kernel for pupil dilation to classify workload while performing a math task into two levels. Instead of using pupil diameters in a time domain, Yokoyama et al. (2018) used high- and low-frequency power of pupil size variations with linear SVMs to estimate human workload while driving. In addition to pupil-related measures, researchers have investigated other eye-related features. For instance, Halverson et al. (2012) used SVMs with different kernels (i.e., linear, quadratic, polynomial, multilayer perceptron [mlp], and Gaussian radial basis function [RBF]) to estimate human workload with features extracted from different time windows (1, 5, 10, and 30 seconds). Among the different features (i.e., blink duration, blink frequency, closure, fixation duration, NNI, percentage of eye closure [PERCLOS], pupil diameter, saccade

Table 1.3: Machine learning for workload estimation using eye-related features

Reference	Model	Feature	Evaluation method
Chen and Epps (2013)	Gaussian Mixture Models (GMMs)	Pupil diameter, saccadic amplitude, fixation duration	Within-participants
Liang et al. (2007)	SVM(RBF kernel), Logistic Regression	Fixation duration, mean and standard deviation of fixation positions, mean of blink frequency, other driving-related feature	Within-participants
Halverson et al. (2012)	SVM (linear, RBF, quadratic, polynomial, mlp kernel)	Pupil diameter, fixation duration, saccade duration, blink duration, blink frequency, NNI, saccade frequency, saccade velocity, percentage eye closure	Within-participants
Yokoyama et al. (2018)	SVM (linear kernel)	High and low Frequency power of pupil size variation	Within-participants
Kosch et al. (2018a)	SVM (linear kernel)	Pupil dilation	Within-participants
Kosch et al. (2018b)	SVM	Gaze deviation from reference track	Within- and cross-participants
Zhang et al. (2008)	Decision Tree	Mean and standard deviation of pupil size, number of gazes in AOI, portion of time in AOI, mean visit time of AOI, other driving related features	Within- and cross-participants
Fridman et al. (2018)	HMM, Convolutional neural network (CNN)	Gaze trajectory, eye image	Cross-participants
Hogervorst et al. (2014)	SVM (linear kernel), Elastic net	Pupil size, blink rate, blink duration, other EEG and ECG features	Within-participants

duration, saccade frequency, and saccade velocity), they found that pupil diameter from five-second time window with a linear kernel achieved the best performance.

Unlike the above previous studies, which focused on a single machine learning model

for a single feature, researchers have also used a single machine learning model for multiple features by concatenating different features into one feature vector (Liang et al., 2007; Zhang et al., 2008; Chen and Epps, 2013). For example, Liang et al. (2007) combined eye-related measurements (i.e., fixation duration, mean and standard deviation of fixation positions, and mean of blink frequency) and driving-related measurements into one feature vector for SVMs with an RBF kernel. Similarly, Zhang et al. (2008) used decision trees to combine gaze-related measurements (i.e., number of gazes in AOI, portion of time in AOI, and mean visit time of AOI), pupil-related measurements (i.e., mean and standard deviation of pupil size), and driving-related measurements. Instead of concatenating all measurements together, Chen and Epps (2013) selected top candidate measurements based on multiple regression analysis and used GMMs to classify human workload into different levels. Recently, Fridman et al. (2018) used a novel convolutional neural network (CNN) with raw eye images and the HMM with gaze trajectories to estimate a driver's workload.

The majority of previous studies have focused on a single machine learning model for a single feature or a single machine learning model for multiple features by concatenating them into one feature vector; however, there have been a few exceptions. For example, Hogervorst et al. (2014) trained multiple elastic nets using different types of features (i.e., eye-related features, EEG features, and ECG features) and combined these three models by averaging their probability outputs.

1.4 Research Aim

According to the above literature review on shared control for semi-autonomous vehicles and human workload measurement, the following research gaps exist:

First, existing studies have developed haptic shared control schemes for semi-autonomous vehicles that adapt to different factors. However, workload, as an important human fac-

tor for human-automation interaction (Lee et al., 2017), has not yet been considered for adaptation purposes.

Second, different physiological measurements have been used to estimate human workload. Among them, we are interested in non-intrusive physiological measures, such as eye-related measures using eye trackers. However, the existing studies mainly focused on single machine learning models for single features and showed that different machine learning models work best for different features. Therefore, it is unclear how to leverage different machine learning models that work best for different features to improve overall performance.

Third, the generalizability of the workload estimation is also critical for workload-adaptive human-automation interaction systems. However, the existing studies evaluated their workload estimation models using data collected from a user study, in which different workload levels were controlled by a single factor. Thus, it is unclear whether those models can be generalized to other scenarios.

To fill these research gaps, collaborating closely with Yifan Weng, Dr. Tulga Ersal, and Prof. Jeffrey Stein from the Department of Mechanical Engineering at the University of Michigan, we address the following specific research aims in this dissertation:

(1) We proposed, together with our collaborators from the Department of Mechanical Engineering, a workload-adaptive haptic shared control scheme for semi-autonomous vehicles and examined the effects of workload adaptation on haptic shared control performance.

(2) We explored different eye-related features to estimate human workload. We proposed a computational model to leverage the different machine learning models that work best for different features to improve workload estimation performance.

(3) We investigated the generalizability of our proposed workload estimation model

through different human subject experiments where the workload levels are manipulated by different factors.

1.5 Dissertation Structure

This dissertation is presented in six chapters. Chapter I provides an introduction to the dissertation's research problem, related work, and research aims. Chapter II introduces our developed teleoperated dual-task shared control simulation platform and two pilot studies to determine the design parameters for the platform. Chapter III shows the two human subject experiments used to collect data for building the workload estimation model and investigating the effects of the workload-adaptive haptic shared control scheme on human performance. Chapter IV introduces our proposed Bayesian inference model for workload estimation that leverages different machine learning models for different features. Chapter V illustrates the generalizability of our proposed Bayesian inference model in two human subject experiments. Chapter VI summarizes this dissertation and presents its intellectual merit and broad impact before making recommendations for future work.

CHAPTER II

Teleoperated Dual-task Shared Control Platform

2.1 Introduction

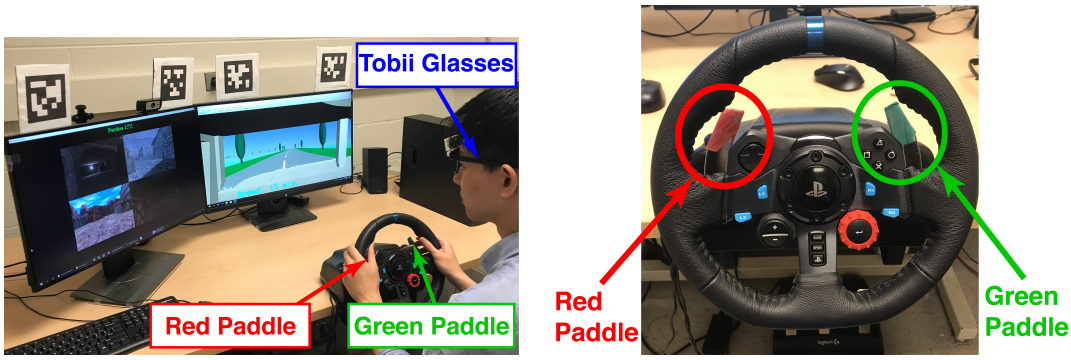
Together with our collaborators - Yifan Weng, Dr. Tulga Ersal, and Prof. Jeffrey Stein from the Department of Mechanical Engineering at the University of Michigan - we developed a teleoperated dual-task shared control simulation platform where human operators perform a driving task and a surveillance task simultaneously.

This chapter describes the design of the teleoperated dual-task shared control simulation platform and the selection of design parameters using two pilot studies with 16 participants. The simulation platform consisted of two tasks: a driving task and a surveillance task. In the driving task, participants shared the control of a ground vehicle together with autonomy. In the surveillance task, participants identified potential threats in the image feedings.

Pilot Study 1 was meant to develop the tracks used in the simulation platform. Pilot Study 2 was meant to determine the parameters in the surveillance task.

2.2 Dual Task Shared Control Simulation Platform

Collaborating with researchers from the Department of Mechanical Engineering at the University of Michigan, we developed a dual-task shared control simulation plat-



(a) Driving task (right screen) and surveillance task (left screen).

(b) Red and green paddles used to perform the surveillance task.

Figure 2.1: Dual-task shared control simulation platform.

form for teleoperation of a simulated notional high mobility multipurpose wheeled vehicle (HMMWV). The platform has been used throughout this dissertation.

In this testbed, participants performed two tasks simultaneously, a driving task and a surveillance task, as shown in Fig. 2.1. In the driving task, a participant and the autonomy shared the control of the HMMWV at a fixed speed of 15 m/s (around 34 mph) to drive as close to the centerline as possible. In the surveillance task, the participant received image feeds and needed to identify potential threats (Figure 2.2). If the participant identified a threat, s/he pressed the red paddle at the steering wheel to report "danger." Otherwise, the participant pressed the green paddle to report "clear" (Figure 2.1b). As the steering wheel can only rotate from -90 degrees to 90 degrees, participants would not need to cross their hands and could always keep the hands on the steering wheel. The potential threat will appear in only one of the four images. The screenshots are selected with the same difficulty as in (Yang et al., 2017; Du et al., 2018).

Figure 2.3 shows the pipeline for the surveillance task. There was a transition period with a white screen between two sets of image feeds. Participants needed to identify the potential threats within a certain time limit. The fixed time limit was varied to manipulate the workload level. In Pilot Study 2, we investigated the effects of different fixed time

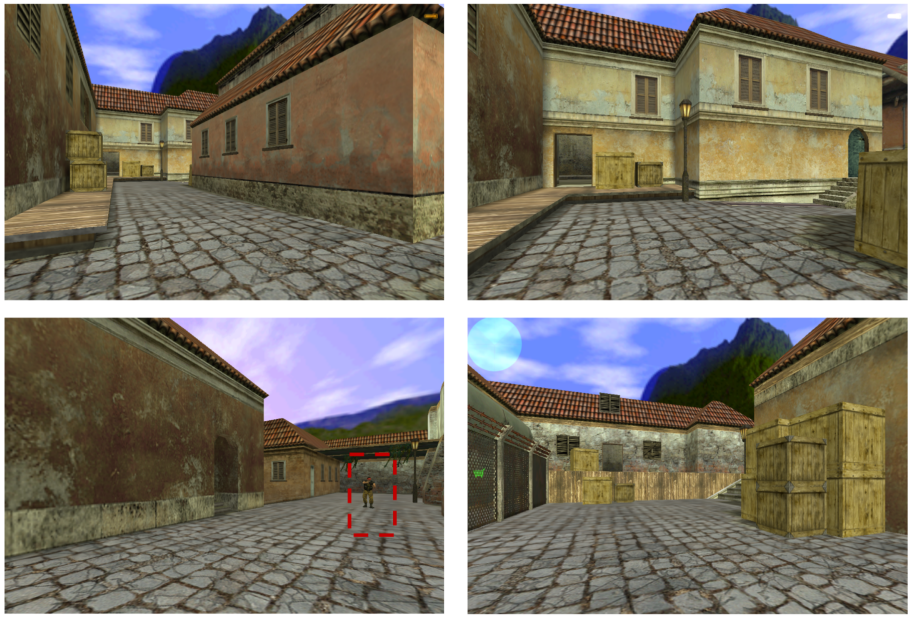


Figure 2.2: Illustration of the surveillance task. Lower left: threat.

limits on human operators' workloads.

2.3 Real-time Gaze Points in World Frame

We used a novel eye tracker, Tobii Pro Glasses 2 (Tobii Pro AB, 2014), to measure human pupil sizes and gaze points in real time. The Tobii Pro Glasses 2 only provided real-time gaze points in the Tobii front camera frame. However, we required human gaze points in the world frame to estimate human workload (see Chapter IV). Let O^F represent the Tobii front camera frame, as shown in Figure 2.4a. We captured a photo of the entire workspace, as shown in Figure 2.4b, and used its coordinate system as the world frame. Let O^W represent the coordinate system of this world image. Let p^F represent the gaze point in the Tobii front camera O^F, which could be directly obtained from the Tobii Pro Glasses 2. Therefore, our goal is to compute p^W, which is the gaze point mapped to the fixed world image O^W.

To transform p^F to p^W in real time, we need to compute the homography matrix be-

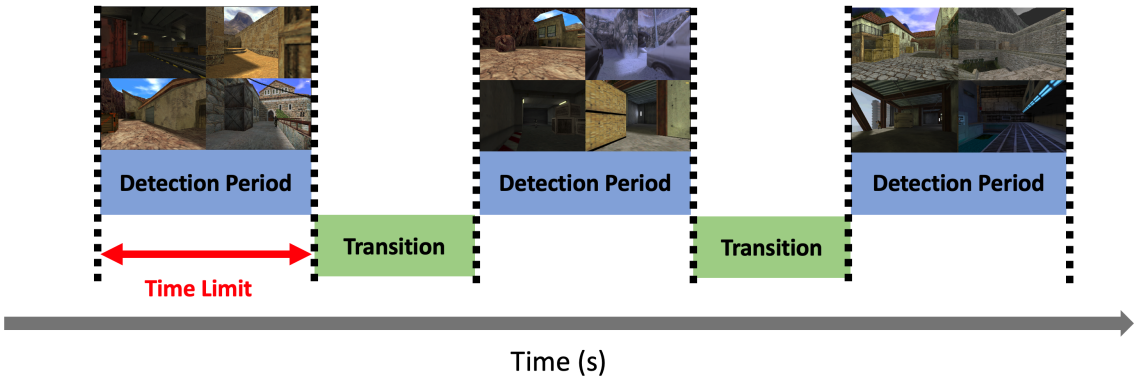


Figure 2.3: Pipeline for surveillance task. Participants received image feeds and needed to identify potential threats within the time limit. There was a transition period with a white screen between two sets of image feeds. The transition period lasted for one second.

tween O^F and O^W in real time. Thus we attached four AprilTags (Wang and Olson, 2016) on top of the monitors, as shown in Figure 2.4b. As we could not obtain real-time image feeds from the Tobii front camera, we introduced an additional camera on top of the Tobii Pro Glasses 2 that was mounted with a 3D printed frame, as shown in Figure 2.4a. Let O^C represent the coordinate system of the additional camera. We used O^C as a bridge between O^F and O^W.

First, let $p^F = [x_p^F, y_p^F, z_p^F, 1]^T$ represent the 3D gaze point in the Tobii front camera. Let R_F^C represent the rigid body transformation matrix between O^F and O^C (i.e., the homogenous transformation matrix). Thus, we have $p^C = R_F^C p^F$, where p^C represents the 3D gaze point in O^C.

Second, we transformed 3D point p^C to the pixel point $q^C = [u^C, v^C, 1]^T$ in the additional camera, where u^C, v^C represents the pixel location of the gaze point in the additional camera plane. Let $K^C \in \mathbb{R}^{3\times4}$ represent the intrinsic camera parameter. We have $q^C \sim K^C p^C$, where we normalized the third dimension of q^C to 1. Note that both the R_F^C and K^C are fixed; therefore, we can obtain these two matrices beforehand by a calibration procedure. Specifically, we used the multiple camera calibration tool with a pinhole cam-

era model from the open-source package Kalibr (Kannala and Brandt, 2006; Maye et al., 2013).

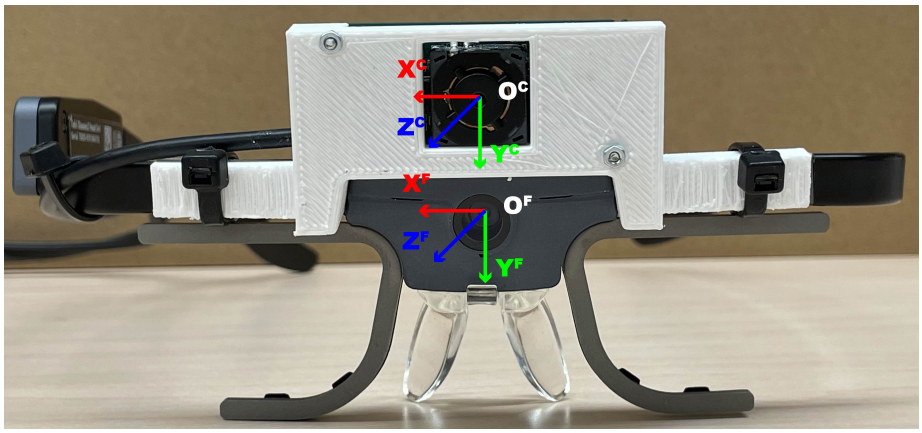
Finally, we transformed the pixel location of the gaze point q^C from O^C to O^W. Let H_C^W represent the homography matrix between O_C and O_W. We have $q^W \sim H_C^W q^C$, where q^W represents the pixel location of the gaze point in the world image (Figure 2.4b). We normalized the third dimension of q^W to 1.

Note that H_C^W is not fixed; it will change when human operators move their heads during the experiment. First, we obtained the corner points' locations of the AprilTags in both the world image and the real-time image feed from the additional camera. Then, we used Random Sample Consensus (RANSAC) algorithm to compute the homography matrix O_C and O_W (Fischler and Bolles, 1981). Specifically, we used the OpenCV package to compute the homography matrix (Bradski, 2000).

The Tobii Pro Glasses 2 provide gaze points and pupil sizes at 50 Hz. However, the additional camera captures and processes images at 30 Hz. Therefore, we down-sampled the Tobii Pro Glasses 2 to 30 Hz.

2.4 Pilot Study 1: Track Selection

In Pilot Study 1, we developed and selected six driving tracks with two considerations. First, the driving tracks should have the same difficulty. Second, along each track, the difficulty at every point should be roughly the same. The two considerations ensured that the difficulty of the dual-task mission can be easily manipulated by varying the surveillance task urgency because the difficulty of the driving task is fairly constant.



(a) Tobii front camera (O^F) and additional camera (O^C) frames

(b) World image frame (O^W)

Figure 2.4: Coordinate systems for the Tobii front camera (O^F), additional camera (O^C), and world image (O^W).

2.4.1 Method

Participants

Ten participants (age: mean = 21.8 years, *SD* = 2.7 years; two females, eight males) took part in Pilot Study 1. All participants had normal or corrected-to-normal vision and hearing, with an average of 4.1 years of driving experience (*SD* = 1.7 years).

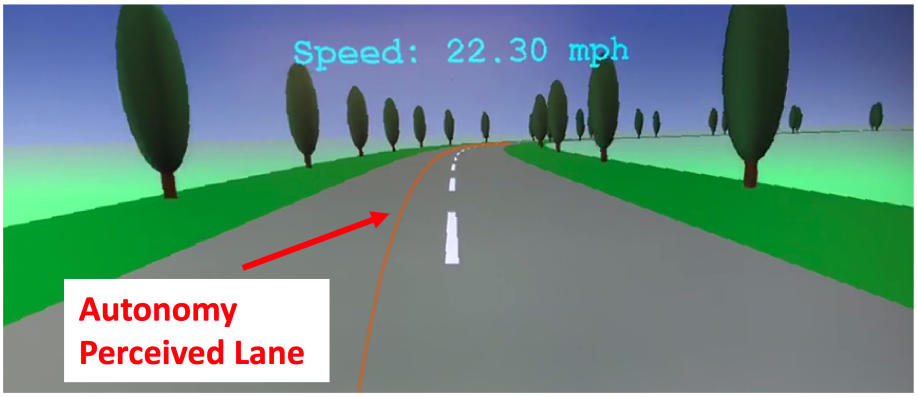


Figure 2.5: Illustration of offset between autonomy perceived lane (orange solid line) and centerline (white dashed line). The offset is 1 m and could be on either the left side or the right side of the centerline.

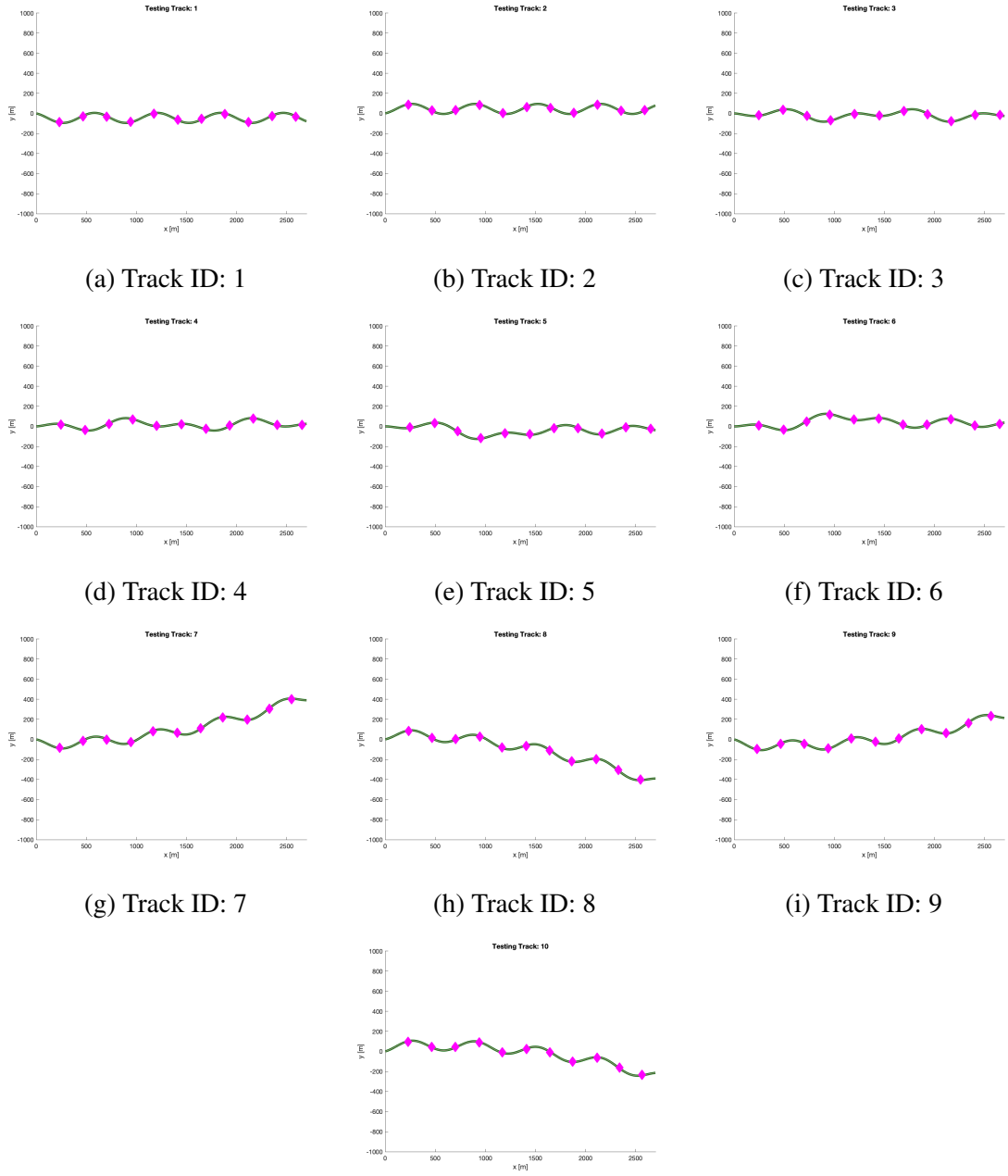
Experimental Apparatus and Stimuli

Pilot Study 1 used the simulation platform shown in Section 2.2. To emulate degraded localization due to sensor uncertainty, we introduced an offset such that the autonomy tracked a line that deviated from the centerline by 1 m in the two pilot studies and two experiments in this chapter. Figure 2.5 illustrates this offset. The orange solid line indicates the autonomy perceived lane, whereas the white dashed line is the centerline. Note that the offset could be on either the left side or the right side of the centerline.

In Pilot Study 1, participants only performed the driving task with the non-adaptive haptic shared control scheme (see Section 3.3.1 in Chapter III for details). We did not present the surveillance task to the participants, as we only wanted to evaluate the difficulty of the driving task.

Experimental Design

Pilot Study 1 used a within-subjects design with 10 different candidate tracks (Figure 2.6). The presentation of tracks followed a 10×10 Latin square design to eliminate potential order effects.



(a) Track ID: 1 (b) Track ID: 2 (c) Track ID: 3

(d) Track ID: 4 (e) Track ID: 5 (f) Track ID: 6

(g) Track ID: 7 (h) Track ID: 8 (i) Track ID: 9

(j) Track ID: 10

Figure 2.6: Candidate tracks. Magenta dots indicate the locations where the participants reported the difficulty of driving.

Measures

Along each track, participants reported the difficulty of driving at 11 locations using a 7-point Likert scale (1: easiest; 7: most difficult). The magenta dots in Figure 2.6 indicate

the locations where the participants reported the difficulty of driving. After completing each track, participants also evaluated to what extent the track had the same difficulty anywhere along it using another 7-point Likert scale (1: the same; 7: significantly different). We named it the "uniformity score." For each track, we calculated the average of the 11 reported difficulty scores as the "overall difficulty score" of the track.

Experimental Procedure

Participants provided signed informed consent and filled in a demographic survey. During the training session, the participants performed two trials on the training tracks, and each trial took approximately 1.5 minutes. In the first trial, the participants only drove on the track and did not report difficulty. However, in the second trial, the participants drove on the track and reported difficulties at the four designed locations, indicated by a sign on the side of the road in the driving simulator.

In the official pilot study, the participants drove on 10 different tracks and reported difficulties at the 11 designed locations. After each track, the participants were asked to evaluate whether driving was the same or significantly different at any location of the track using a 7-point Likert scale.

After finishing all 10 trials, the subjects were required to fill out a debriefing survey about any outstanding questions and their opinions of or suggestions for the experiment they had just completed.

2.4.2 Results

One-way repeated measures analysis of variance (ANOVA) was conducted for the driving tracks as the within-subjects variable. The results showed a non-significant difference between the 10 tracks in their overall difficulty scores ($F(9, 81) = 1.161$, $p = 0.331$) and in their uniformity score ($F(9, 81) = 0.557$, $p = 0.828$). Based on the results, we selected

tracks 2, 3, 5, 6, 8, and 9 to be used in Pilot Study 2 and Experiment 1, and tracks 2, 3, 6, and 9 to be used in Experiment 2.

2.5 Pilot Study 2: Surveillance Task Parameter Selection

In this chapter, we aimed to manipulate the difficulty of the dual-task mission and, hence, the human operators' workload by varying the surveillance task urgency. In Pilot Study 2, we selected a fixed time limit for the detection period of the surveillance task so that the difficulty and workload of the dual-task mission could be manipulated.

2.5.1 Method

Participants

Seven participants took part in Pilot Study 2. The data from one participant were discarded due to an equipment malfunction. The remaining six participants were on average 25.3 years old (*SD* = 1.6 years) and had an average of 2.7 years of driving experience (*SD* = 1.6 years). There were two females and four males in the remaining six participants. All participants had normal or corrected-to-normal vision.

Experimental Apparatus and Stimuli

Pilot Study 2 used the testbed mentioned in Section 2.2 with the tracks selected in Pilot Study 1. The dual-task mission was presented to the participants. A non-adaptive haptic shared control scheme was applied. The offset of the autonomy perceived lane was applied as well. We also used the non-adaptive shared control scheme (see Section 3.3.1 in Chapter III for details) in Pilot Study 2.

Experimental Design

The pilot study used a within-subjects design with six different time limits for the detection period of the surveillance task: 1.5, 2.5, 3.5, 4.5, 5.5, and 6.5 seconds (i.e., par-

ticipants had to complete the detection task within the given time limit). The six time limits were selected based on the results from our previous study (Luo et al., 2019). Figure 2.7 shows the histogram of the response time of the participants during the surveillance task. Participants performed both the driving task and the surveillance task on six different tracks, each with a constant different time limit for the detection period. The presentation of surveillance task conditions followed a 6×6 Latin square design to eliminate potential order effects.

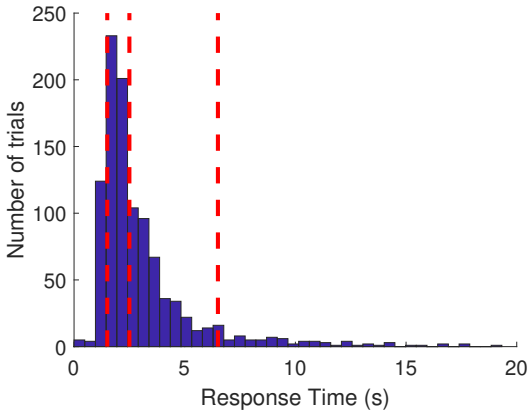


Figure 2.7: Histogram of response time for surveillance task from previous study. Red dash lines indicate 1.5, 2.5, and 6.5 seconds respectively.

Measures

Participants reported their workload of the dual-task mission using the NASA TLX survey (Hart and Staveland, 1988), and their perceived difficulty of the dual-task mission.

Experimental Procedure

Participants provided signed informed consent and filled out a demographic survey. After that, they were provided with instructions and training. Participants were first trained on the driving task alone, followed by the surveillance task alone. After that, they performed both the driving and surveillance tasks on three different tracks. Each track had a

different time limit for the surveillance task: 5.5, 3.5, and 1.5 seconds.

During the official pilot study, participants performed the driving task and surveillance task on six different tracks with six different surveillance task fixed time limits. Each track took approximately three minutes. After each trial, the participants were asked to fill out a survey regarding their workload and difficulty during each track.

After finishing all six trials, the subjects were required to fill out a debriefing survey regarding any outstanding questions and their opinions of or suggestions for the experiment they had just completed.

2.5.2 Results

One-way repeated measures ANOVA were conducted with the detection time limits for the surveillance task as the within-subjects variable. The results showed a significant difference of time limit on workload ($F(5, 25) = 10.458$, $p < 0.001$) and difficulty ($F(5, 25) = 13.423$, $p < 0.001$). We then performed a series of t tests between different pairs of time limits. The results revealed significant differences in workload and difficulty between 1.5 s and 2.5 s (workload: $p < .001$, difficulty: $p = .006$), between 1.5 s and 3.5 s (workload: $p = .005$, difficulty: $p = .012$), between 1.5 s and 4.5 s (workload: $p = .004$, difficulty: $p = .006$), between 1.5 s and 5.5 s (workload: $p = .001$, difficulty: $p < .001$), and between 1.5 s and 6.5 s (workload: $p = .004$, difficulty: $p < .001$). The differences between any other pairs of time limits were non-significant.

Based on the results, we selected 1.5 s and 6.5 s time limits to be used in Experiment 1 and Experiment 2 in Chapter III to induce varying levels of workload. Note that in Experiment 1, we also included the 2.5 s time limit, as we were interested in exploring participants' performance with a slightly larger time limit compared to the 1.5 s time limit.

2.6 Conclusion

In this chapter, we developed a teleoperated dual-task shared control simulation platform as an example of a human-automation interaction system. The human operators shared the control of an HMMWV together with autonomy while completing a surveillance task simultaneously. We introduced how to obtain real-time pupil sizes and gaze points in a world frame using Tobii Pro Glasses 2. We conducted two pilot studies to determine the tracks used in the driving task and the time limits of the detection period in the surveillance task. The selected tracks have similar difficulty, and the difficulty along each track is consistent. The selected time limits of the detection period in the surveillance task can manipulate human workload into different levels.

The findings should be viewed in light of the following limitations. First, due to the limitation of image processing speed, we had to down-sample the data from Tobii Pro Glasses 2 to 30 Hz. This may introduce some errors due to the imperfect synchronization between the eye tracker and the additional camera, particularly when the humans moved their heads quickly.

Second, the number of participants in the pilot studies is limited. More participants were required for Pilot Study 1 to receive non-significant results between the tracks. However, when we designed the candidate tracks, we tried to keep the track curvatures consistent and similar to each other. In addition, some tracks were mirrored with other tracks (i.e., Track 1 and Track 2).

Third, we tried to manipulate the human operators' workload by varying the time limit of the detection period in the surveillance task, and we tried to keep the difficulty of the driving task the same. However, human workload could also be imposed by the driving task, for example, by the driving speed and headway required to avoid an obstacle. We

addressed these issues in Chapter V.

Finally, the visualization system for the driving task in the proposed simulation platform was not realistic. We updated the visualization system to a high-fidelity visualization system and investigated human performance under the high-fidelity visualization system in Chapter V.

CHAPTER III

Workload-adaptive Haptic Shared Control

3.1 Introduction

To investigate the effects of real-time workload estimation in the human-automation interaction system, we utilized semi-autonomous vehicles as an example. Together with our collaborators - Yifan Weng, Dr. Tulga Ersal, and Prof. Jeffrey Stein from the Department of Mechanical Engineering at the University of Michigan, we proposed a heuristic design for the control consolidation that adapts to the human workload, which builds the adaptive haptic shared control scheme.

In this chapter, we conducted two human subject experiments. Experiment 1 recorded the eye-related measurements of human operators under different surveillance task conditions to collect a data set to build the workload estimation model. We used the Hidden Markov Model to analyze the human operators' gaze trajectory, based on which their workload was estimated. In Experiment 2, we evaluated the performance of the proposed workload-adaptive haptic shared control scheme.

3.2 Experiment 1: Data Collection for Workload Estimation

In Experiment 1, we collected a data set of human eye-related measurements (i.e., pupil sizes and gaze points) under different surveillance task urgency. We used the Hidden

Markov Model (HMM) to estimate the human workload by analyzing 4 s gaze trajectory data.

3.2.1 Workload Estimation with HMM

An HMM is a probabilistic model of the joint probability of a collection of random variables $\{O_1, O_2, ..., O_T, S_1, S_2, ..., S_T\}$. S_t is a discrete variable that represents the hidden state at time step t. S_t can take values from $\{1, 2, ..., N\}$, where N is the number of hidden states. O_t represents the observations at time step t. T represents the termination time step. An HMM also contains a tuple of parameters as $\Theta = (\pi, A, B)$. $\pi \in \mathbb{R}^N$ is the prior distribution of $P(S_1)$. $A \in \mathbb{R}^{N \times N}$ is the stochastic transition matrix, where $A = \{a_{i,j}\} = P(S_t = j | S_{t-1} = i)$. $B = \{b_j(\cdot)\}$ is a set of observation model for every hidden state $j \in \{1, 2, ..., N\}$, where $b_j(\mathbf{o}_t) = P(O_t = \mathbf{o}_t | S_t = j)$ and $\mathbf{o}_t$ is a given observation at time step t.

In the present study, the observations $\mathbf{o}_t$ are the gaze points, i.e., locations of where the human is looking at relative to the external world coordinate shown as the magenta dots in Figure 3.1. The observation models are a set of multivariate distributions over the gaze points, i.e., $b_j(\mathbf{o}_t) = P(O_t = \mathbf{o}_t | S_t = j) \sim \mathcal{N}(\mu_j, \Sigma_j)$, shown as the ellipsoids in Figure 3.1. Thus $B = \{\mu_j, \Sigma_j\}$.

We trained two HMMs, one for the high workload and one for the moderate workload. For each workload level w, we collected a set of L gaze trajectories $D_w = \{\mathcal{O}_l | \mathcal{O}_l = \{\mathbf{o}_1^l, \mathbf{o}_2^l, ..., \mathbf{o}_T^l\}\}$, where $l = \{1, 2, ..., L\}$. Thus, the learning process learns two sets of HMM parameters $\Theta_w = (\pi, A, B)$, one for each workload level using data D_w. The parameters of the HMMs were learned by the Expectation Maximization(EM) algorithm using the open source implementations from Rozo et al. (2016) and Calinon (2016). To learn the parameters, we defined four important probabilities:

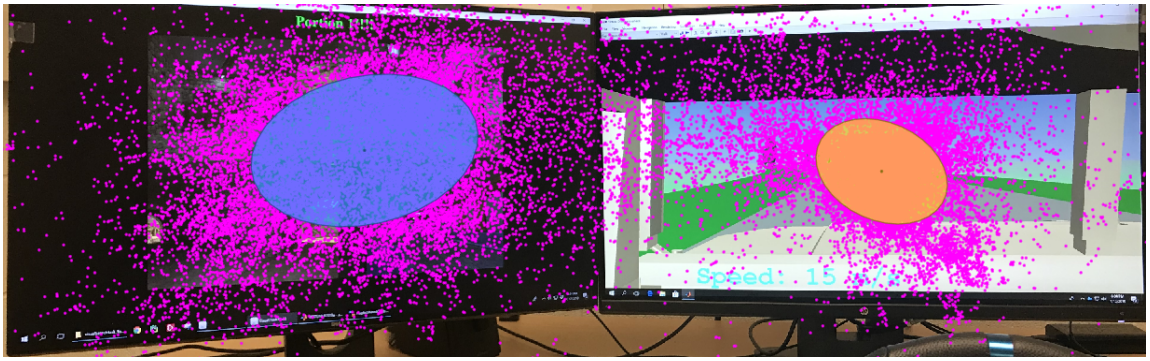


Figure 3.1: Example of using the Hidden Markov Model to model gaze trajectory to estimate workload. Magenta dots: gaze points. Ellipsoids: Multivariate normal distributions.

$$
\begin{aligned}
\alpha_i^l(t)^k &= P(O_1 = \mathbf{o}_1^l, ..., O_t = \mathbf{o}_t^l, S_t = i|\Theta^k) \\
\beta_i^l(t)^k &= P(O_{t+1} = \mathbf{o}_{t+1}^l, ..., O_T = \mathbf{o}_T^l|S_t = i, \Theta^k) \\
\gamma_i^l(t)^k &= P(S_t = i|\mathcal{O}_l, \Theta^k) \\
\xi_{i,j}^l(t)^k &= P(S_t = i, S_{t+1} = j|\mathcal{O}_l, \Theta^k)
\end{aligned}
\tag{3.1}
$$

where k represents the k^{th} iteration in the EM algorithm. The EM algorithm is then:

E-step:

Recursively update α:

$$\begin{aligned}
\alpha_i^l(1)^{k+1} &= \pi_i^k \mathcal{N}(\mathbf{o}_1^l; \mu_i^k, \Sigma_i^k) \\
\alpha_j^l(t+1)^{k+1} &= [\textstyle\sum_{i=1}^N \alpha_i^l(t)^{k+1} a_{i,j}^k] \mathcal{N}(\mathbf{o}_{t+1}^l; \mu_j^k, \Sigma_j^k)
\end{aligned}$$

Recursively update β:

$$\begin{aligned}
\beta_i^l(T)^{k+1} &= 1 \\
\beta_i^l(t)^{k+1} &= \textstyle\sum_{j=1}^N a_{i,j}^k \beta_j^l(t+1)^{k+1} \mathcal{N}(\mathbf{o}_{t+1}^l; \mu_j^k, \Sigma_j^k)
\end{aligned}$$

Update γ:

$$\gamma_i^l(t)^{k+1} = \frac{\alpha_i^l(t)^{k+1} \beta_i^l(t)^{k+1}}{\sum_{j=1}^N \alpha_j^l(t)^{k+1} \beta_j^l(t)^{k+1}}$$

Update ξ:

$$\xi_{i,j}^l(t)^{k+1} = \frac{\gamma_i^l(t)^{k+1} a_{i,j}^k \beta_j^l(t+1)^{k+1} \mathcal{N}(\mathbf{o}_{t+1}^l; \mu_j^k, \Sigma_j^k)}{\beta_i^l(t)^{k+1}}$$

M-step:

$$\begin{aligned}
\mu_i^{k+1} &= \frac{\sum_{l=1}^L \sum_{t=1}^T \gamma_i^l(t)^{k+1} \mathbf{o}_t^l}{\sum_{l=1}^L \sum_{t=1}^T \gamma_i^l(t)^{k+1}} \\
\Sigma_i^{k+1} &= \frac{\sum_{l=1}^L \sum_{t=1}^T \gamma_i^l(t)^{k+1} (\mathbf{o}_t^l - \mu_i^{k+1})(\mathbf{o}_t^l - \mu_i^{k+1})^T}{\sum_{l=1}^L \sum_{t=1}^T \gamma_i^l(t)^{k+1}} \\
\pi_i^{k+1} &= \frac{\sum_{l=1}^L \gamma_i^l(1)}{L}^{k+1} \\
a_{i,j}^{k+1} &= \frac{\sum_{l=1}^L \sum_{t=1}^T \xi_{i,j}^l(t)^{k+1}}{\sum_{l=1}^L \sum_{t=1}^T \gamma_i^l(t)^{k+1}}
\end{aligned}$$

The two steps iterate until convergence. The number of hidden states was determined by the Bayesian Information Criterion (BIC) (Calinon and Billard, 2005; Schwarz et al., 1978).

Given a gaze trajectory $\mathcal{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$, we computed the likelihood of $P(\mathcal{O}|\tilde{\Theta}_w)$ via the forward algorithm, where $\tilde{\Theta}_w$ represents parameters for different learned HMMs for the high workload and moderate workload. The forward algorithm is similar to the recursive update of α in the E-step of the EM algorithm. We have $P(\mathcal{O}|\tilde{\Theta}_w) = \sum_{i=1}^N \tilde{\alpha}_i(T)$. To estimate the workload of $\mathcal{O}$, we found the HMM with the maximum likelihood, i.e.,

$\arg\max_w P(\mathcal{O}|\tilde{\Theta}_w)$.

Our adaptive shared control scheme is based on the human operator's real-time workload, eyes on road, and input torque (see Section 3.3 for details). We used the gaze point data from a 4 s time window captured by the Tobii eye tracker (30 Hz sampling rate as mentioned in Section 2.3) to estimate participants' workload and eyes on road. Thus, $T = 120$. Let w_t represent a human operator's workload at time t, $w_t = c_1 \arg\max_w p(\mathcal{O}_t|\tilde{\Theta}_w) + c_2$, where c_1, c_2 are scaling and offset factors such that $w_t = 50$ represents moderate workload, and $w_t = 100$ represents high workload. A human operator's eye on road is defined as the percentage of time that s/he is looking at the driving task. Let e_t denote the human operator's eyes on road. e_t is calculated as the average number of times that a participant's gaze points fall on the driving screen within the time window T.

Due to the large mass and high center of gravity of the simulated HMMWV, a rapid change of control commands resulting from a rapid change of w_t and e_t could trigger a rollover. Therefore, we applied a moving average filter with a 1 s time window and downsampled w_t and e_t to 10 Hz.

3.2.2 Method

Participants

A total of 13 university students participated in the experiment. Data from one participant were discarded due to equipment malfunction. The remaining 12 participants were on average 26.7 years old (*SD* = 3.0 years) and had an average of 8.3 years of driving experience (*SD* = 4.4 years). There were 6 females and 6 males in the remaining 12 participants. All participants had a normal or corrected-to-normal vision.

Experimental Apparatus and Stimuli

Experiment 1 used the same testbed as mentioned in Section 2.2 with the tracks selected in Pilot Study 1. The dual-task mission was presented to the participants. Similar to Pilot Study 1 and Pilot Study 2, we applied the autonomy's perception offset for the centerline to emulate the sensor uncertainty as shown in Figure 2.5. The non-adaptive haptic shared control scheme was applied (see Section 3.3.1 for details).

Experiment Design

We manipulated the workload of the experimental tasks (the driving and the surveillance task) by varying the time limits for the detection period of the surveillance task. During the experiment, the participants drove on six different tracks, each lasting for approximately 3 min. Every track was equally segmented into three portions, and each portion had a different time limit for the detection period for the surveillance task, 1.5, or 2.5, or 6.5 s. The order of presentation for the time limits on each track is balanced by two 3×3 Latin squares.

Measures

Participants wore a pair of the Tobii Pro Glasses 2, and their gaze points and pupil sizes were recorded at 30Hz as mentioned in Section 2.3.

Experiment Procedure

Participants provided a signed informed consent and filled in a demographic survey. After that, they received a training session. In the training session, participants first performed a driving only task to get familiar with driving with the non-adaptive haptic shared control autonomy, which takes approximately 1.5 minutes and then performed three trials of surveillance task with $6.5, 2.5, 1.5$ second fixed time limit for the detection period

where each trial takes approximately 60 seconds. After that, the participants performed driving and surveillance task together on 3 different tracks with different surveillance task fixed time limits - $6.5, 2.5, 1.5$ second. Each track takes approximately 1.5 minutes. The order of the surveillance task fixed time limits ($6.5, 2.5, 1.5$ seconds) is designed to help the participants to build capability to perform surveillance task with different difficulties gradually.

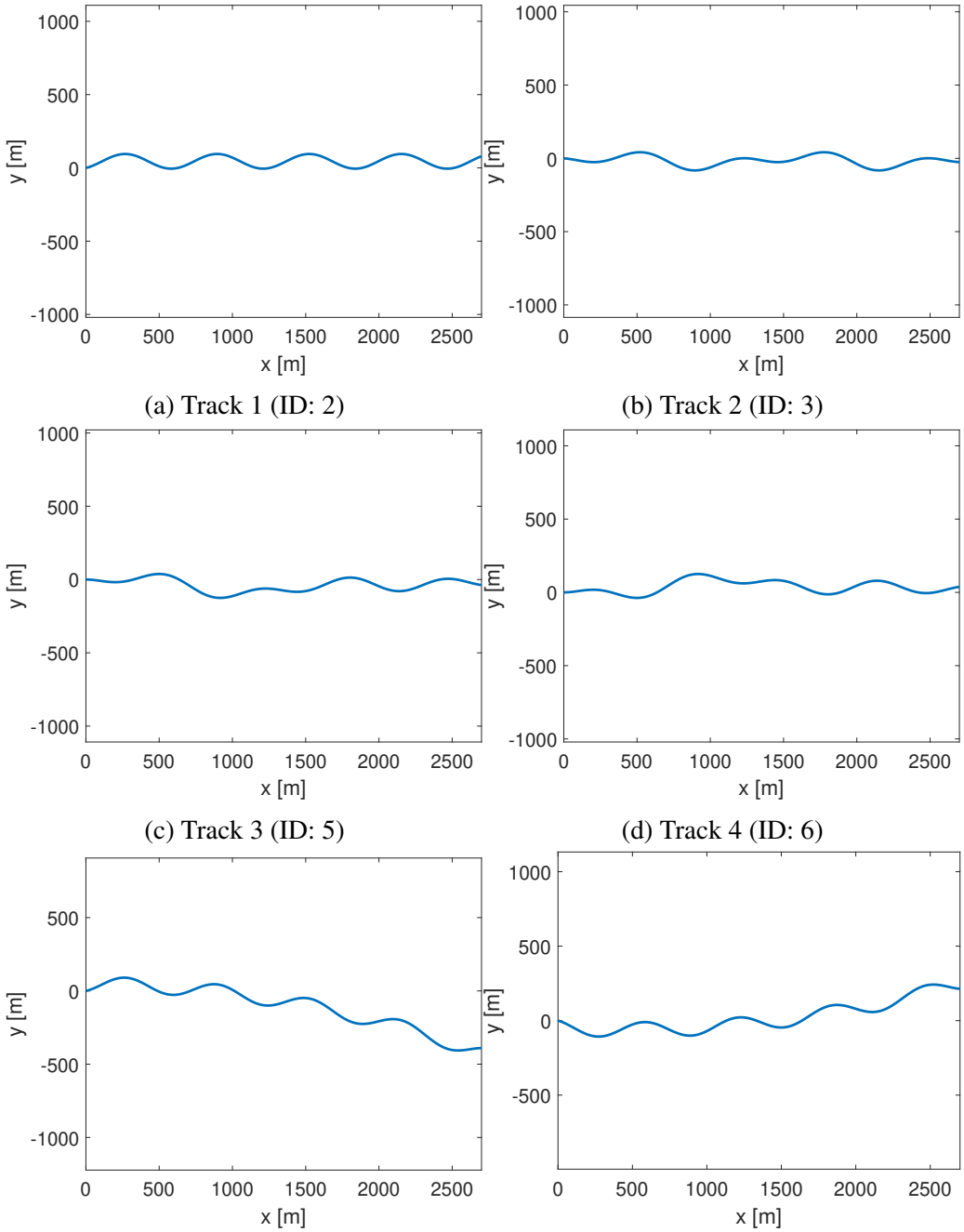
After the training session, participants were assisted to wear the eye tracker and underwent the calibration. With the normal room light and without any specific tasks, the experimenter measured each participant's baseline pupil diameters twice, each about 30 s. Participants were asked to sit down, look at the white wall, relax, and clear their minds during the measurement of the baseline pupil diameters. During the formal experiment, participants performed the driving task and the surveillance task on six different tracks, each lasting approximately 3 min.

3.2.3 Results

In Experiment 1, we used Hidden Markov Models to model gaze trajectories as a benchmark for workload estimation for the collected dataset.

Data Processing

Participants drove on six different tracks in this experiment as shown in Figure 3.2. As mentioned above, each track was segmented into three portions, and each portion had a different time limit for the detection period of the surveillance task as shown in Figure 3.3. We treated the portion with 1.5 s time limit as the high workload portion and the portion with 6.5 s time limit as the moderate workload portion. The ground truth labels were determined in Pilot Study 2. For each track, we randomly selected five sequences of data in each portion, and each sequence lasted 4 s, shown as the blue/yellow boxes in Figure 3.3.



(a) Track 1 (ID: 2)

(b) Track 2 (ID: 3)

(c) Track 3 (ID: 5)

(d) Track 4 (ID: 6)

(e) Track 5 (ID: 8)

(f) Track 6 (ID: 9)

Figure 3.2: Six selected tracks in Experiment 1.

Evaluation of Workload Estimation Performance

Due to the small dataset of 12 participants, we used the holdout method (Kim, 2009) for cross-validation and tested the performance of our proposed method. In each run of the

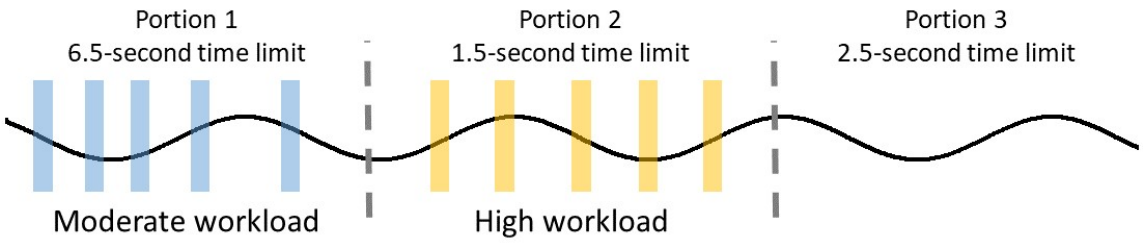


Figure 3.3: Illustration of track segmentation. Black curve indicates the track. Blue boxes indicate 5 randomly selected sequences of data in moderate workload portion. Yellow boxes indicate 5 randomly selected sequences of data in high workload portion. Each sequence lasted for 4 s.

Table 3.1: Performance of the HMM

	F_1	Precision	Recall
HMM	0.664 ± 0.005	0.668 ± 0.005	0.660 ± 0.005

holdout, we randomly selected data of 3 participants as the testing dataset and data of the remaining 9 participants as the training dataset. To find the best number of hidden states, we varied the number of hidden states from 2 to 10 for the HMM and ran 100 holdouts for each number of hidden states. We used the Bayesian Information Criterion (BIC) (Calinon and Billard, 2005; Schwarz et al., 1978) to determine the best number of hidden states. The results indicated that 2 was the best number of hidden states.

We then ran another 100 holdouts to evaluate the performance of the HMM for workload estimation. Precision, recall and F_1 score were used as performance metrics, where precision $= \frac{\#\text{true positives}}{\#\text{true positives}+\#\text{false positives}}$ and recall $= \frac{\#\text{true positives}}{\#\text{true positives}+\#\text{false negatives}}$. For our multi-classification problem, the precision is the mean precision of all classes and the recall is the mean recall of all classes. $F_1 = 2\frac{\text{precision}\cdot\text{recall}}{(\text{precision}+\text{recall})}$. Table 3.1 shows the mean and standard error of each performance metric. The results show that the HMM model achieved a 0.66 F_1 score, 0.67 precision, and 0.66 recall.

3.2.4 Discussion

In this experiment, we recorded human gaze trajectories and pupil sizes while completing a teleoperated dual-task mission under different workload levels. The different workload levels were imposed by varying different surveillance task urgency (i.e., different time limits for the detection period in the surveillance task). Using a baseline method, Hidden Markov Models for gaze trajectory, we achieved 0.66 F_1 score to classify human workload into high workload level and moderate workload level.

The findings should be viewed in light of the following limitations. First, we only used gaze trajectory to estimate human workload. Other eye-related features could be extracted and other machine learning models could be used to estimate human workload. For example, researchers have used support-vector machine for human pupil size change to estimate human workload (Hogervorst et al., 2014; Halverson et al., 2012; Kosch et al., 2018a). It is unclear how to leverage these different features to estimate human workload. In Chapter IV, we discussed our proposed Bayesian inference model that can leverage different machine learning models that work best for different features.

Second, we introduced different workload levels by varying the surveillance task urgency. However, it is unclear if we can distinguish workload induced by other factors, such as driving speed. In Chapter V, we investigateed the generalizability for our proposed model.

Third, we only recruited 12 participants to evaluate the workload estimation performance in this experiment. We recruited more participants to evaluate the workload estimation performance in Chapter IV.

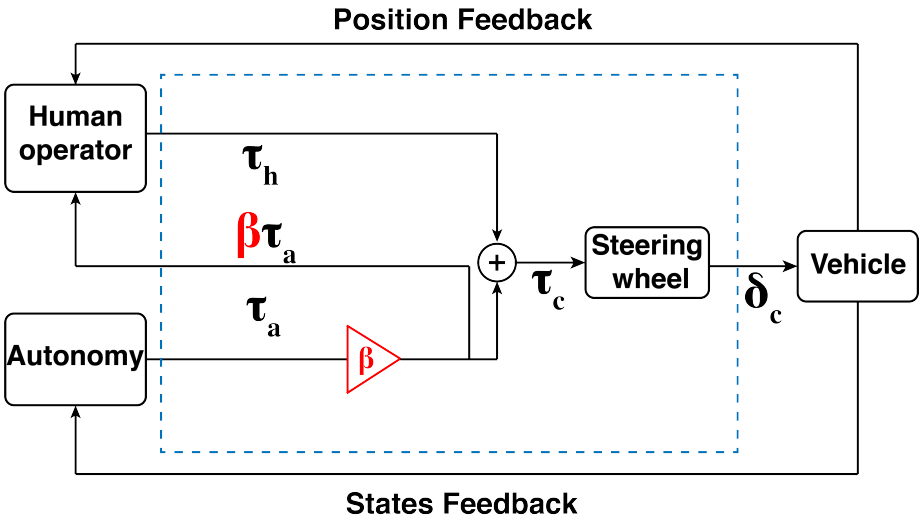


Figure 3.4: Block diagram for haptic shared control. τ_h and τ_a represent the torque from human and autonomy, respectively. τ_c and δ_c are the actual control torque and actual control steering angle. β is the assistance level, which is always 1 in the baseline non-adaptive scheme, whereas it varies in the proposed adaptive scheme.

3.3 Experiment 2: Workload-adaptive Shared Control Scheme

In Experiment 2, we tested whether by considering the drivers' workload, haptic shared control performance could be improved. Two haptic shared control schemes were used: the adaptive haptic shared control and the non-adaptive haptic shared control schemes. The adaptive haptic shared control scheme adapted to the estimated real-time workload, and the participant's eyes on road and torque input. We used the HMM learned with the data from all the 12 participants to estimate the participant's workload in real time.

3.3.1 Non-adaptive Haptic Shared Control

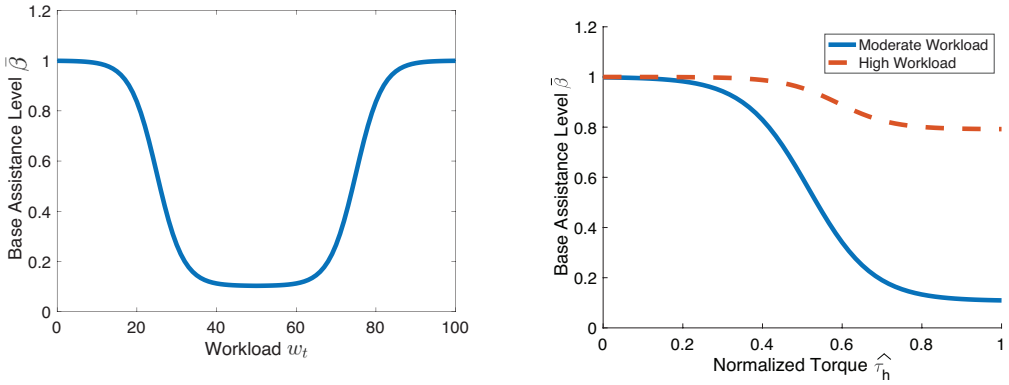
Haptic shared control combines the torques applied by the autonomy and human operator. It creates a smooth control authority transfer between the human operator and autonomy. The implementation is visualized in Figure 3.4, where $\beta = 1$ for the baseline non-adaptive case.

When there is no input from the human operator, the autonomy follows the reference centerline it perceives. The perceived reference centerline may be different from the actual centerline. When there is an input from the human operator that deviates the vehicle from the centerline autonomy perceives, the autonomy applies extra torque to bring the vehicle back to the perceived centerline. Hence, the human operator can feel the intention of the autonomy and decide whether s/he would agree with it and let autonomy have more control authority (yield) or claim more control authority (fight). The resultant torque applied on the steering wheel, which is the summation of the torques from the human operator and autonomy, determines the final steering angle applied to the vehicle.

3.3.2 Adaptive Haptic Shared Control

The adaptive haptic shared control scheme was designed by the interdisciplinary team. The adaptive was designed based on three different factors: human workload, human torque on the steering wheel, and human eyes-on-road. The eyes-on-road is the percentage of time that a human is looking at the driving screen, i.e., the human is focusing on driving. The team introduced this factor since human is performing a dual-task. The resultant torque τ_c in the adaptive scheme is $\tau_c = \tau_h + \beta(w_t, e_t, \hat{\tau}_h)\tau_a$, where the term β is referred to as assistance level and it determines the strength of assistance torque from autonomy. $\hat{\tau}_h$ is the normalized human torque calculated by dividing the input torque from the human operator by the maximum torque a human operator can apply. Figure 3.4 shows the implementation of the adaptive scheme. This scheme contrasts with the direct blending of the input torques from both the human operator and autonomy as in the non-adaptive haptic shared control scheme. Specifically, β is always 1 in the baseline non-adaptive haptic shared control scheme, whereas it varies in the proposed adaptive scheme.

In the heuristic design for the assistance level, β was separated into two parts: base



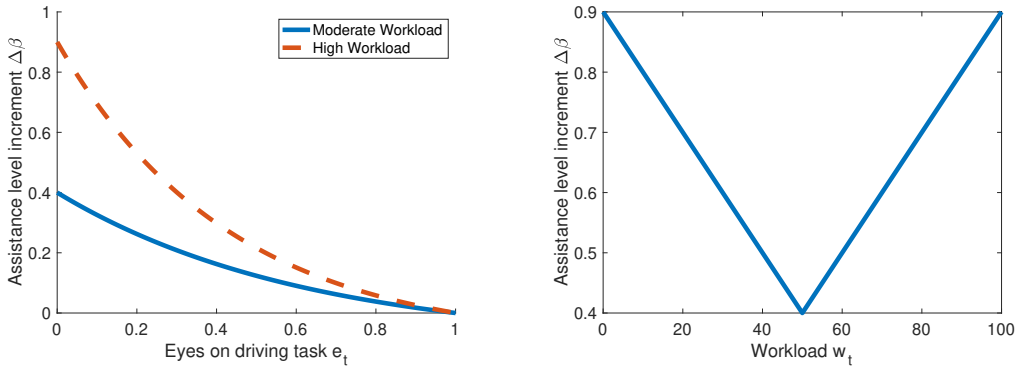
(a) Relationship between base assistance level $\bar{\beta}$ and workload w_t

(b) Relationship between base assistance level $\bar{\beta}$ and normalized human torque $\hat{\tau}_h$

Figure 3.5: Illustration of base assistance level $\bar{\beta}$ design principles

assistance level $\bar{\beta}$ and assistance level increment $\Delta\beta$; i.e., $\beta = \bar{\beta}(w_t, \hat{\tau}_h) + \Delta\beta(w_t, e_t)$. The base assistance level $\bar{\beta}$ considers the impact from workload and input torque from the human operator, while the assistance level increment $\Delta\beta$ considers the combined effect of eyes on road and workload due to the dual task nature of our experiment setup.

The base assistance level $\bar{\beta}$ was designed according to the principles illustrated in Figure 3.5. On the one hand, when the torque from the human operator is constant, the relationship between the base assistance level $\bar{\beta}$ and workload w_t is shown in Figure 3.5a. When a human experiences a moderate workload, the human is more capable of performing the task. Thus the base assistance level should be lower (i.e., when w_t is around 50 in Figure 3.5a). However, when a human is underloaded or overloaded, the human may either have vigilance decrement or lack of mental resources to achieve the task. In either cases, human control of driving may not be reliable, thus the base assistance level should be larger (i.e., when $w_t < 20$ or $w_t > 80$ in Figure 3.5a). On the other hand, the human input torque may indicate the human's intention to control the vehicle. Therefore, when the normalized human torque $\hat{\tau}_h$ increases, the base assistance level $\bar{\beta}$ should decrease as human are more willing to take control as shown in Figure 3.5b. Moreover, when the hu-



(a) Relationship between assistance level increment $\Delta\beta$ and eyes on road for different workloads

(b) Relationship between assistance level increment $\Delta\beta$ and workload w_t when $e_t = 0$

Figure 3.6: Illustration of assistance level increment design principles

man experiences moderate workload, the $\bar{\beta}$ should decrease more, shown as the blue curve in Figure 3.5b.

The assistance level increment $\Delta\beta$ was designed according to the principles illustrated in Figure 3.6. On the one hand, when the human eye-on-road decreases (i.e., moving from right to left in Figure 3.6a), the human is less focusing on the driving task. Therefore, the assistance level increment $\Delta\beta$ should increase. Moreover, when the human is overloaded, $\Delta\beta$ should increase more, shown as the orange dotted curve in Figure 3.6a. On the other hand, keeping the eyes-on-road constant, when the workload is high, the increment $\Delta\beta$ is large, while when the workload is moderate, the increment $\Delta\beta$ is small, which is shown in Figure 3.6b.

By line fitting with some designed points in above curves, we have:

$$\bar{\beta}(w_t, \hat{\tau}_h) = 1 - \left[1 - \left(\frac{0.9e^{0.3(|w_t-50|-25)}}{e^{0.3(|w_t-50|-25)}+1} + 0.1\right)\right]\left[\frac{e^{\frac{72\hat{\tau}_h-36.6-15(\frac{w_t}{50}-1)^2}{5.9-2.5(\frac{w_t}{50}-1)^2}}}{e^{\frac{72\hat{\tau}_h-36.6-15(\frac{w_t}{50}-1)^2}{5.9-2.5(\frac{w_t}{50}-1)^2}}+1}\right] \tag{3.2}$$

$$\Delta\beta(w_t, e_t) = 0.1(0.1|w_t - 50| + 5)^{1-e_t} - 0.1 \tag{3.3}$$

Figure 3.7 shows the 3D plot for the base assistance level $\bar{\beta}$, workload w_t and normal-

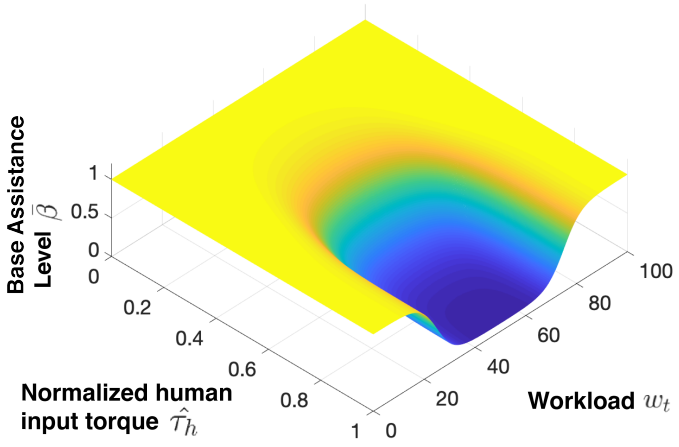


Figure 3.7: Relationship between base assistance level $\bar{\beta}$, workload w_t, and normalized human input torque $\hat{\tau}_h$.

ized human input torque $\hat{\tau}_h$. Figure 3.8 shows the 3D plot for the assistance level increment $\Delta\beta$, workload w_t, and eyes on road e_t.

3.3.3 Method

Participants

A total of 13 students participated in the experiment. Data of 1 participant were discarded due to the wrong experiment setup. The remaining 12 participants were on average 22.3 years old (*SD* = 3.7 years) and had an average of 5.7 years of driving experience (*SD* = 3.9 years). There were 5 females and 7 males in the remaining 12 participants. All participants had a normal or corrected-to-normal vision.

Experimental Apparatus and Stimuli

The same teleoperated dual-task shared control simulation platform was used in this experiment as in Experiment 1. We also applied the autonomy's perception offset for the centerline to emulate the sensor uncertainty as shown in Figure 2.5. Both the adaptive haptic shared control and the non-adaptive haptic shared control were used in this experiment.

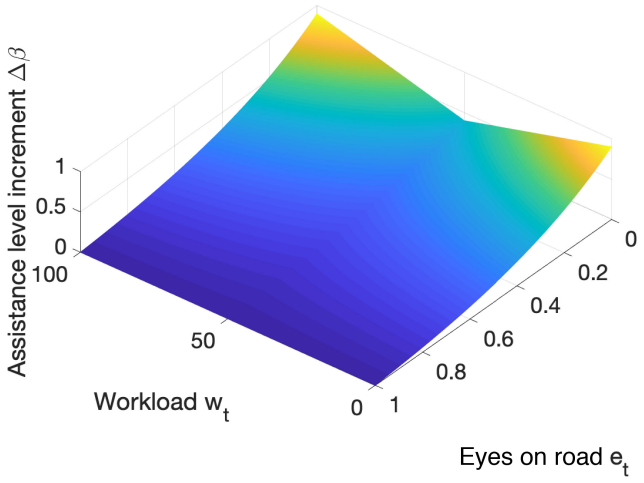


Figure 3.8: Relationship between assistance level increment $\Delta\beta$, workload w_t, and eyes on road e_t.

Table 3.2: Four Test Conditions

Condition	Surveillance task urgency		Haptic Shared Control Scheme
	First half of the track	Second half of the track	
1	1.5 s	6.5 s	Non-adaptive
2	1.5 s	6.5 s	Adaptive
3	6.5 s	1.5 s	Non-adaptive
4	6.5 s	1.5 s	Adaptive

Experimental Design

The experiment used a within-subjects design with two independent variables. The first independent variable was the haptic shared control scheme (adaptive haptic shared control vs. non-adaptive haptic shared control). The second independent variable was the surveillance task urgency (1.5 s vs. 6.5 s). Each participant experienced four tracks in the experiment. On each track, one type of haptic shared control scheme was used. Each track was segmented into two portions, one portion with high urgency surveillance task (1.5 s) and the other with low urgency surveillance task (6.5 s). The resulting four test conditions are shown in Table 3.2. The presentation of test conditions followed a 4×4 Latin square design to eliminate potential order effects.

Measures

Five dependent variables were collected in the experiment: participants' self-reported workload and trust in the shared control autonomy, participants' control effort, driving task performance, and surveillance task performance. After each track, participants reported their workload and trust for the first and second half of the track using two uni-dimensional scales. The NASA TLX survey (Hart and Staveland, 1988) and the Jian's trust survey (Jian et al., 2000) were presented to the participants such that they understood the meaning of workload and trust. Participants' control effort was calculated as the average torque that a participant applied on the steering wheel. Driving task performance was evaluated by lane-keeping error. The lane-keeping error is calculated as the mean of the absolute deviation of the vehicle's position from the centerline. The surveillance task performance was measured using the detection accuracy.

Experimental Procedure

Participants provided a signed informed consent and filled in a demographic survey. After that, they were assisted to wear the eye tracker with calibration. The experimenter measured each participants' baseline pupil diameter twice each about 30 s before the training with the normal room light and without any specific tasks.

During the training session, the participants first performed two trials of driving task only, one with the non-adaptive haptic shared control and one with the adaptive haptic shared control. Each trial took approximately 1.5 min. Then the participants performed three trials of the surveillance task only. Each trial took approximately 60 s. After that, the participants performed four trials of the combined driving and surveillance task.

During the official experiment, participants performed the driving task and the surveillance task on four different tracks with different test cases as described in Table 3.2. Each

Table 3.3: Mean and Standard Error (SE) of workload, trust, lane keeping error, detection accuracy and torque

Metrics	N	Surveillance task urgency			
		1.5 s		6.5 s	
		Adaptive	Non-adaptive	Adaptive	Non-adaptive
Workload	12	13.96 ± 0.82	14.08 ± 0.87	7.83 ± 0.81	8.71 ± 0.97
Trust	12	4.04 ± 0.37	3.63 ± 0.30	3.92 ± 0.32	3.29 ± 0.38
Lane keeping error (m)	12	0.28 ± 0.033	0.36 ± 0.045	0.21 ± 0.03	0.26 ± 0.04
Detection accuracy (%)	12	93.43 ± 1.38	91.86 ± 1.13	94.30 ± 1.77	96.54 ± 1.18
Torque (Nm)	12	0.36 ± 0.03	0.73 ± 0.03	0.30 ± 0.02	0.79 ± 0.01

trial took approximately 3 min. After each trial, participants filled a post-survey about the workload and trust during each portion of the track.

3.3.4 Results

Two-way repeated measures Analysis of Variance (ANOVAs) were conducted with the shared control scheme and the surveillance task urgency as the within-subjects variables. Results are reported as significant for $\alpha < .05$. Table 3.3 summarizes the mean and standard error (SE) values of the participants' self-reported workload and trust as well as driving task performance, surveillance task performance and their exerted torque.

Participants' Workload

Both control scheme and surveillance task urgency influence participants' self-reported workload. With the adaptive shared control, participants reported lower workload ($F(1, 11) = 5.18$, $p = .044$). When the surveillance task was less urgent, participants reported lower workload ($F(1, 11) = 20.26$, $p < .001$). See Figure 3.9.

Trust in Automation

Participants trusted the shared control autonomy more when the autonomy was adaptive ($F(1, 11) = 12.76$, $p = .004$). The effect of surveillance task urgency on trust was not significant. See Figure 3.10.

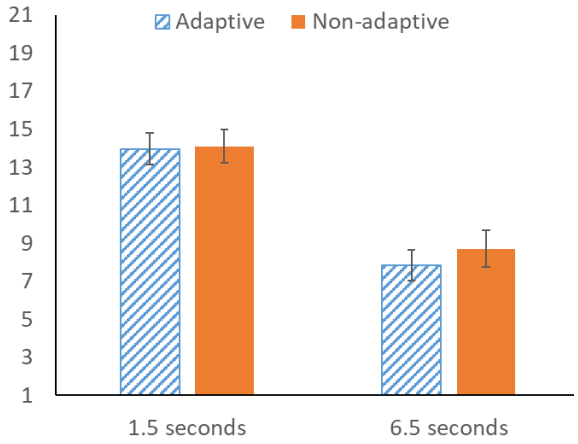


Figure 3.9: Mean and standard error (SE) values of self-reported workload.

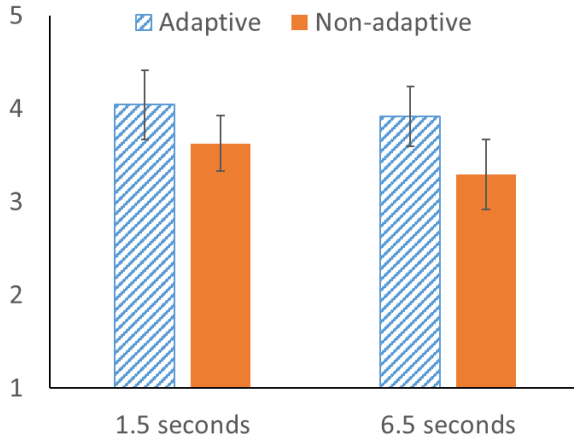


Figure 3.10: Mean and standard error (SE) values of self-reported trust.

Driving Task Performance

Results revealed that the haptic shared control scheme and the surveillance task urgency significantly affected the driving task performance. Participants had smaller lane keeping errors when using the adaptive shared control autonomy ($F(1, 11) = 7.593$, $p = .019$), and when the surveillance task was less urgent ($F(1, 11) = 96.33$, $p < 0.001$) (Figurge 3.11). There was also an interactive effect between the control scheme and surveillance task urgency ($F(1, 11) = 6.141$, $p = .031$). Using adaptive shared control led to a large reduction in lane keeping error when the surveillance task was more urgent.

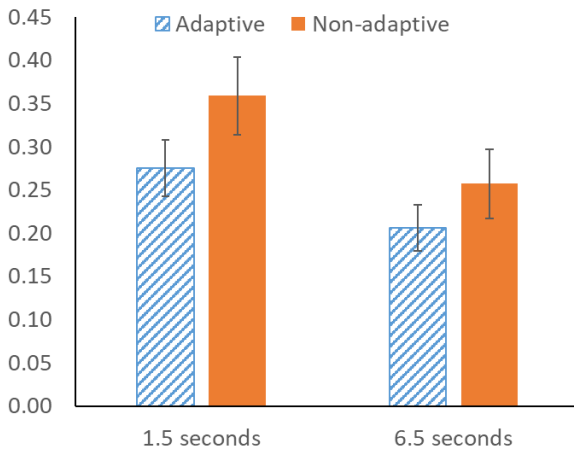


Figure 3.11: Mean and standard error (SE) values of lane keeping error (m).

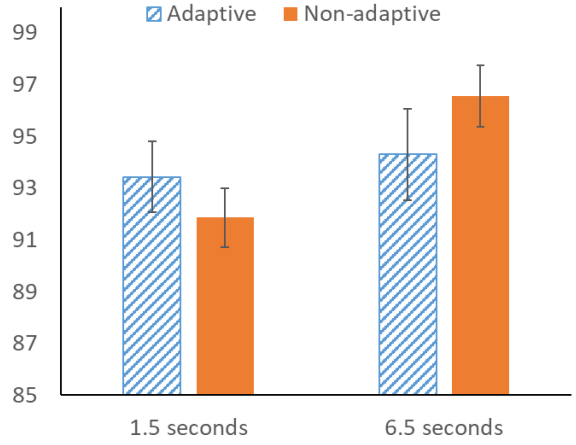


Figure 3.12: Mean and standard error (SE) values of surveillance task detection accuracy (%).

Surveillance Task Performance

For the surveillance task, task urgency significantly influenced the detection accuracy ($F(1, 11) = 6.73$, $p = .025$). Detection accuracy was higher when the task was less urgent. The effect of the shared control scheme was non-significant (Figure 3.12).

Participants' Control Effort

There was a significant effect of shared control scheme on participants' control effort ($F(1, 11) = 217.66$, $p < .001$). With adaptive shared control, participants exerted signif-

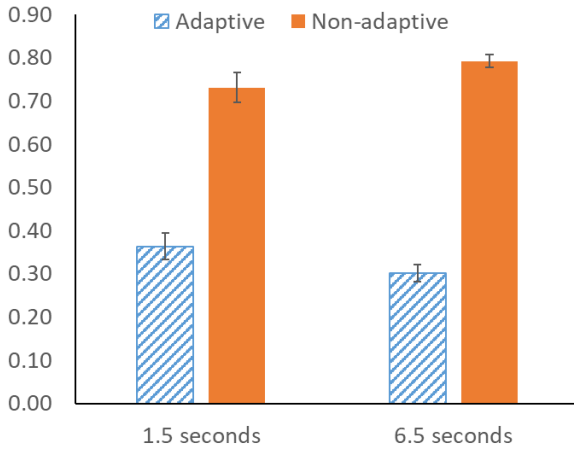


Figure 3.13: Mean and standard error (SE) values of participants' torque (Nm)

icantly less control effort. The effect of surveillance task urgency on participants' control effort was non-significant. In addition, results revealed a significant interaction effect between control scheme and surveillance task urgency ($F(1, 11) = 11.42$, $p = .006$). When the surveillance task was less urgent (6.5 s), the adaptive shared control scheme led to a larger drop in torque. See Figure 3.13.

3.3.5 Discussion

Participants' Workload

Participants' self-reported workload decreased when using the adaptive shared control scheme and when the surveillance task became less urgent. The results could have resulted from the following reasons. First, the 6.5 s surveillance task urgency imposed a smaller temporal demand on participants than 1.5 s surveillance task urgency. Second, the participants' control effort was smaller with the adaptive control scheme. Third, participants' driving task performance was higher with the adaptive control scheme and when the surveillance task was less urgent.

Trust in Automation

Our result is consistent with prior research that human operators' trust in automation is determined by the autonomy's performance (Yang et al., 2017; Du et al., 2020a; Guo and Yang, 2020). Human operators perceived both the driving and the surveillance task performance continuously, based on which they adjusted their trust in automation. As the driving task performance increased with the adaptive control scheme, trust increased accordingly.

Driving Task Performance

The results showed that the adaptive shared control scheme benefited the driving task performance, especially when participants were under a high workload. Based on the design of the adaptive haptic shared control scheme, with the same input torque, when the human operator has a high workload and focuses on the surveillance task, the assistance level is increased (average $\beta = 1.03$ with adaptive shared control scheme compared with $\beta = 1$ with non-adaptive shared control scheme when surveillance task is more urgent). The increment in the assistance level is expected to aid the driving task and reduce the lane keeping error. This design principle was supported by the experimental results.

Surveillance Task Performance

As the surveillance task became more urgent and more demanding, the surveillance task performance decreased significantly. This result is consistent with prior research that when workload increased from moderate to high level, task performance would decrease (Lu et al., 2019).

Participants' Control Effort

Our results indicate that with adaptive shared control participants exerted significantly less amount of control effort in both low and high workload conditions. The results can be explained as follows: First, as the participants' trust toward the adaptive shared control scheme is significantly higher than the non-adaptive control scheme, participants had a higher tendency to yield to the autonomy, resulting in smaller input torque. Second, according to the design of the adaptive shared control scheme, with the same input torque, when the human operator experiences moderate workload and focuses on the driving task, the assistance level is reduced (average $\beta = 0.82$ with adaptive shared control scheme compared with $\beta = 1$ with non-adaptive shared control scheme when surveillance task is less urgent). With a reduced assistance level, regardless of whether the human yields to or fights with the autonomy, the human operator's torque is expected to be smaller.

3.4 Conclusion

In this chapter, we conducted two human subject experiments with 24 participants in total. In Experiment 1, we collected a dataset with human gaze trajectories and pupil sizes under different workload levels imposed by different surveillance task urgency. We used the Hidden Markov Model for gaze trajectory feature to estimate human workload and achieved 0.66 F_1 score. In Experiment 2, we proposed a workload-adaptive haptic shared control scheme together with our collaborators. The human subject experiment indicated that the workload-adaptive haptic shared control scheme can reduce human workload, increase their trust in the system, improve driving performance, and reduce human control effort without sacrificing surveillance task performance. The results indicated that the human-automation interaction system can benefit from adapting to real-time human workload.

The findings should be viewed in light of the following limitations. First, we only used Hidden Markov Model for gaze trajectory to estimate human workload. It is unclear about the performance of other machine learning models with different eye-related features to estimate human workload. In addition, it is unclear how to combine the different machine learning models that work best for different features. We addressed these issues in Chapter IV.

Second, we tried to manipulate human workload levels by varying surveillance task urgency. However, there are other factors that could impose different workload levels. We investigated the performance of our proposed workload estimation model on these different factors could impose different workload levels.

CHAPTER IV

Bayesian Inference Model for Workload Estimation

4.1 Introduction

In Experiment 1, we used the Hidden Markov Model (HMM) for gaze trajectory to estimate human workload and showed that adapting to the estimated workload can help the design of a haptic shared control scheme. In this chapter, we explored other eye-related features for workload estimation and propose a Bayesian inference model that can leverage the different machine learning models that work best for different features (i.e., support-vector machines (SVMs) for pupil size change, HMM for gaze trajectory, SVMs for fixation feature, and Gaussian mixture models (GMMs) for fixation trajectory). We evaluated our proposed Bayesian inference model using data collected from 24 participants. The first 12 participants were from Experiment 1, and the second 12 participants were newly recruited.

4.2 Bayesian Inference Model for Workload Estimation

In Experiment 1, we considered only human gaze trajectory to estimate human workload. However, while other eye-related features have been shown to be useful for workload estimation, they work with different machine learning models. For example, previous studies showed that SVMs could be used with human pupil dilation (Kosch et al., 2018a) and

fixation features (i.e., fixation duration) (Liang et al., 2007) to estimate human workload. In addition, different kernels are suitable for different features (i.e., the linear kernel for pupil dilation (Kosch et al., 2018a) and the radial basis function (RBF) kernel for fixation duration (Liang et al., 2007)). Therefore, we proposed a Bayesian inference model that can leverage the different machine learning models that work best for different features. Figure 4.1 shows the graphical representation of our proposed Bayesian inference model, where W_L is human workload; $M_1, M_2, ..., M_n$ represent the workload estimated by different machine learning models; and $X_1, X_2, ..., X_n$ represent the different features for different machine learning models. The shaded circles represent the observed data, and the unshaded circles represent the hidden states. $W_L, M_1, M_2, ..., M_n$ are discrete random variables, representing different workload levels. The maximum a posteriori (MAP) estimate of workload is used to compute $\arg\max_{W_L} p(W_L|X_1, X_2, ..., X_n)$. Given the probabilistic graphical model, we had the following equations based on the Bayes' rule and the law of total probability:

$$
\begin{aligned}
& p(W_L|X_1, X_2, ..., X_n) \\
\propto\ & p(X_1, X_2, ..., X_n|W_L)p(W_L) \\
=\ & p(W_L) \sum_{M_1, M_2, ..., M_n} p(X_1, X_2, ..., X_n, M_1, M_2, ..., M_n|W_L) \\
=\ & p(W_L) \sum_{M_1, M_2, ..., M_n} p(X_1, X_2, ..., X_n|M_1, M_2, ..., M_n, W_L)P(M_1, M_2, ..., M_n|W_L) \\
=\ & p(W_L) \sum_{M_1, M_2, ..., M_n} \{\prod_{M_i} p(M_i|W_L)p(X_i|M_i)\} \\
=\ & p(W_L) \prod_{M_i} \{\sum_{M_i} p(M_i|W_L)p(X_i|M_i)\}
\end{aligned}
\tag{4.1}
$$

$p(W_L)$ is the prior distribution of the human workload. $p(M_i|W_L)$ is the prior knowledge of the performance of the machine learning model M_i. $p(X_i|M_i)$ is the likelihood of

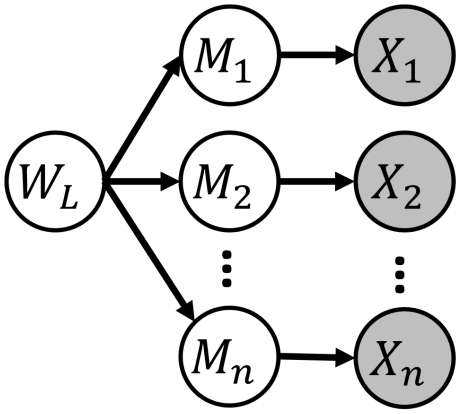


Figure 4.1: A graphical representation of the Bayesian inference model. W_L is the human's workload. M_i represents the workload estimated by different machine learning models. X_i is the feature for the different machine learning models.

each feature X_i given the machine learning model M_i. Both $p(W_L)$ and $p(M_i|W_L)$ could be obtained by manual design based on prior knowledge or from the training data. We used the frequency in the training data to determine $p(W_L)$. For $p(M_i|W_L)$, we segmented the training data into a validation set and a training set and used the performance of M_i on the validation set as $p(M_i|W_L)$.

In this chapter, we investigated four different eye-related features. We selected three features from the existing literature, including gaze trajectory (Fridman et al., 2018), pupil size change (Halverson et al., 2012), and fixation feature (Halverson et al., 2012). In addition, we proposed a new feature – the fixation trajectory feature. For each feature, we selected a machine learning model that has been shown to work well for this feature. Details are in the following sections.

4.2.1 Support-vector Machines (SVMs) for Pupil Size Change

In Experiment 1, we used the Tobii Pro Glasses to measure a human pupil size. Upon each participant's arrival, we measured their baseline pupil size D_B. We asked the participants to relax while looking at a white wall and then measured their pupil sizes for 30 seconds twice. The baseline pupil size D_B is the average pupil size during this time period

for each participant.

The pupil size change feature is the relative changes in the human pupil size. Given a sequence of pupil sizes $D = \{D_1, ..., D_T\}$, the pupil size change feature vector is $X_1 = \{\frac{D_t - D_B}{D_B}\}_{t=1,2,...,T}$. Previous literature used SVMs to estimate human workload using the pupil size change feature (Halverson et al., 2012; Hogervorst et al., 2014; Kosch et al., 2018a). The SVMs is a supervised learning algorithm that aims to find the optimal hyperplane that separates data points into different clusters. We found that using a radial basis function (RBF) kernel can achieve better performance for the pupil size change feature. We can use pairwise coupling to estimate probability $p(X_1|M_1)$ for a multi-class classification problem, where each class represents each workload level (Wu et al., 2004).

4.2.2 Hidden Markov Model (HMM) for Gaze Trajectory

Gaze trajectory X_2 is a time series of gaze points, and $X_2 = \{(g_x^t, g_y^t)\}_{t=1,2,...,T}$, where (g_x^t, g_y^t) is the human gaze point location mapped to the world frame at time t captured by the eye tracker. Previous literature used the HMM to model human gaze trajectory to estimate human workload (Fridman et al., 2018). We need to learn an HMM for each level of human workload using the expectation-maximization algorithm. Thus, we can compute the likelihood $p(X_2|M_2)$ by the standard forward algorithm. See Section 3.2.1 for more details.

As $p(X_2|M_2)$ is the probability density of the gaze trajectory, the longer the trajectory is, the smaller this value is. To eliminate the influence of trajectory length, one can use a geometric mean of the probability density of a trajectory (Luo et al., 2018), shown as follows:

$$\hat{p}(X_2|M_2 = w) = \sqrt[T]{P(\mathcal{O}|\tilde{\Theta}_w)} \tag{4.2}$$

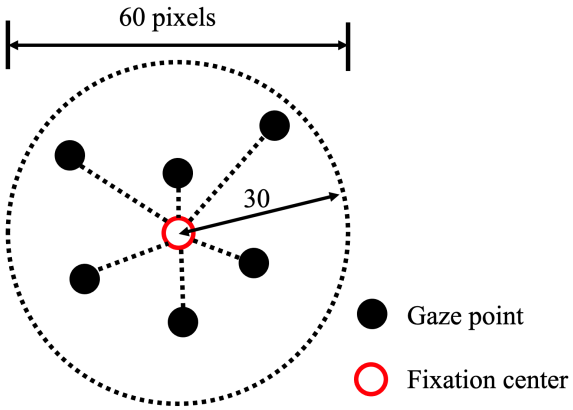


Figure 4.2: Fixation definition. The gaze points are constrained in a circle with a 60-pixel diameter. The fixation center is the common mean location of the gaze points.

4.2.3 Support-vector Machines (SVMs) for Fixation Feature

Human eye movement can be modeled as two types of phases: fixations and saccades. Fixations are the phases, in which humans maintain their gaze points at a location for a time period and gather new information from the area they are examing (Jacob, 1995; Rayner, 1995, 2009). Saccades are the rapid eye movements between fixations (Jacob, 1995; Salvucci and Goldberg, 2000; Jacob and Karn, 2003). Given a sequence of gaze points, researchers have proposed different criteria to determine a fixation. The center of a fixation is typically within $2-3^{\circ}$ (Robinson, 1979), and the fixations last at least 100 - 150 ms. We used the criterion that the fixations were constrained in a 3° spatial area and lasted at least 100 ms, in line with Goldberg and Kotval (1999). As we mapped our gaze points to the world image (as in Figure 2.4b), a 3° spatial area from the eyes is roughly a circle with a 60-pixel diameter in the world image. Figure 4.2 shows the definition of fixation. The fixation center is the common mean location of the gaze points, and all gaze points are within the circle with a 60-pixel diameter. As we down-sampled the gaze points to 30 Hz, 100 ms contained roughly four gaze points. Therefore, each fixation should contain at least four gaze points in the present study. Figure 4.3 illustrates the fixations

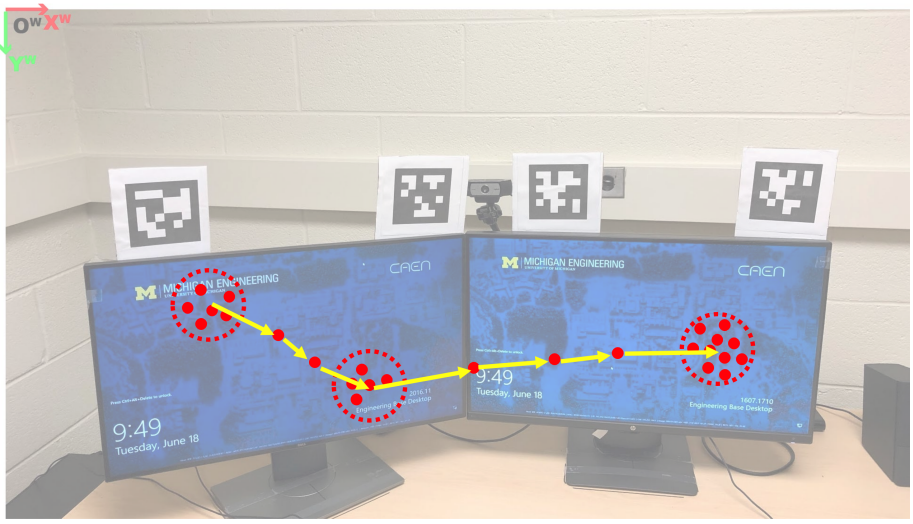


Figure 4.3: Illustration of fixations and saccades mapped on the world image. Red dots are gaze points. Red dashed circles are fixations. Yellow arrows are saccades.

and saccades mapped on the world image. The red dots are the gaze points. The red dashed circles are the fixations. The yellow arrows are the saccades between fixations. We used the same fixation-clustering algorithm as in Goldberg and Kotval (1999) to determine fixations and saccades given a sequence of gaze points, as shown in Algorithm 1.

Algorithm 1: Fixation cluster algorithm

```
Input: G: Sequence of gaze points
Initialization: cluster := { first gaze point p }
for gaze point p in G do
    μ ← ComputeCommonMeanLocation(cluster, p);
    if Distance(μ, p) ≤ 30 then
        cluster ← p;
    else
        n ← Size(cluster);          // number of points in cluster
        if n ≥ 4 then
            cluster is classified as a FIXATION of n × 33.33 ms duration;
        else
            cluster is classified as a SACCADE of n × 33.33 ms duration;
        end
        cluster := {p };
    end
end
```

Researchers have found that different measurements related to fixations and saccades can indicate human workload (Recarte and Nunes, 2000; Moacdieh et al., 2020). Fixation feature X_3 is a vector of these measurements. In our experiment, we defined $X_3 = (n_f, t_f, r, l)$, where n_f is the number of fixations within the time window T; t_f is the total fixation duration in the time window T; $r = \frac{t_f}{t_s}$ is the ratio between fixation duration and saccade duration; and l is the mean saccadic amplitude. The mean saccadic amplitude is the sum of the distances between consecutive fixations divided by the number of fixations minus one within the time window T.

Previous studies have used SVMs for the fixation feature to estimate human workload (Liang et al., 2007). We found that using a linear kernel can achieve better performance for the fixation feature. Similar to pupil size change, we can use the pairwise coupling method to estimate $p(X_3|M_3)$.

4.2.4 Gaussian Mixture Models (GMMs) for Fixation Trajectory

The fixation feature X_3 ignores the spatial information of the fixations. Therefore, we proposed a new feature: fixation trajectory. Fixation trajectory X_4 is a series of fixation centers and their durations, such as $X_4 = \{(f_x^l, f_y^l, dt^l)\}_{l=1,2,...,L}$, where (f_x^l, f_y^l) is the center of a fixation, dt^l is the duration for this fixation, and L is the length of the fixation trajectory, which is the number of fixations within the time window $T = 4$ s. As the number of fixations L during a time window varies, the length of each feature vector varies. Also, the order of the fixations does not matter. Therefore, we can use GMMs to model the fixation trajectory. Similar to the HMM, we need to learn a GMMs for each level of workload M_4^w, where w represents different workload levels. Given an observation X_4, the output of a GMMs is the likelihood $p(X_4|M_4^w)$.

Each GMMs M_4^w is a combination of K multivariate Gaussians gc_k for $k = 1, 2, 3, ..., K$.

Let $\xi^l = (f_x^l, f_y^l, dt^l)^T$ be the l th fixation in the fixation trajectory X_4. The probability of ξ^l in GMMs M_4^w represented by K multivariate Gaussians is given by:

$$p(\xi^l|M_4^w) = \sum_{k=1}^{K} p(gc_k|M_4^w)p(\xi^l|gc_k, M_4^w) \tag{4.3}$$

where ξ^l is the l th fixation in the fixation trajectory X_4, and $p(gc_k|M_4^w) = \pi_k$ is the prior probability of component gc_k in M_4^w. The probability of ξ^l given gc_k and M_4^w is defined as follows:

$$\begin{aligned} p(\xi^l|gc_k, M_4^w) &= \mathcal{N}(\mu_k, \Sigma_k) \\ &= \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} e^{-\frac{1}{2}(\xi^l-\mu_k)^T\Sigma_k^{-1}(\xi^l-\mu_k)} \end{aligned} \tag{4.4}$$

where $\{\mu_k, \Sigma_k\}$ are the mean and covariance parameters of the Gaussian component gc_k, and D is the dimension of ξ^l, which is 3 in the present study. Thus, the probability of trajectory X_4 in M_4^w is defined as follows:

$$\hat{p}(X_4|M_4^w) = \prod_{l=1}^{L} p(\xi^l|M_4^w) \tag{4.5}$$

Similar to the HMM, $p(X_4|M_4^w)$ is the probability density of the fixation trajectory. Therefore, to eliminate the influence of trajectory length, we also used the geometric mean of the probability density of a trajectory (Luo et al., 2018), shown as follows:

$$p(X_4|M_4^w) = \sqrt[L]{\prod_{l=1}^{L} p(\xi^l|M_4^w)} \tag{4.6}$$

Similar to the HMM, we also used the Bayesian information criterion (BIC) (Schwarz et al., 1978; Calinon and Billard, 2005) to determine the best number of Gaussians K, and we found that $K = 3$ is the best fit. The parameters of GMMs $\{\pi, \mu_k, \Sigma_k\}^w$ were trained using the expectation–maximization (EM) algorithm.

4.3 Results

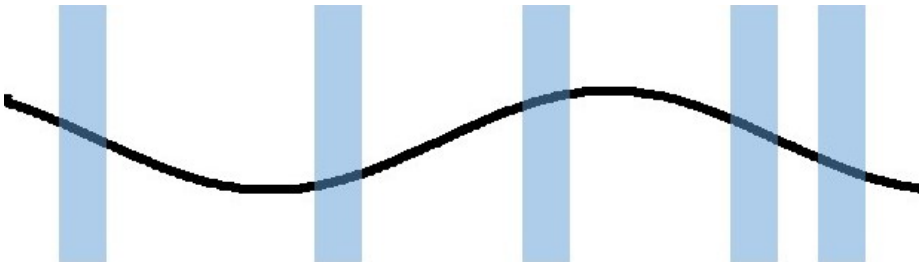
4.3.1 Data Processing

In this chapter, we made use of the data collected in Experiment 1. In addition, we recruited another 12 participants (age: mean = 25.1 years, *SD* = 3.7 years; females: 4, males: 8) for the same experiment. The 12 participants had normal or corrected-to-normal vision and hearing and an average of 4.8 years of driving experience (*SD* = 2.3 years). Therefore, we evaluated our proposed Bayesian inference model (BI) with 24 participants (age: mean = 25.9 years, *SD* = 3.4 years; females: 10, males: 14). All participants had an average of 6.5 years of driving experience (*SD* = 3.9 years).

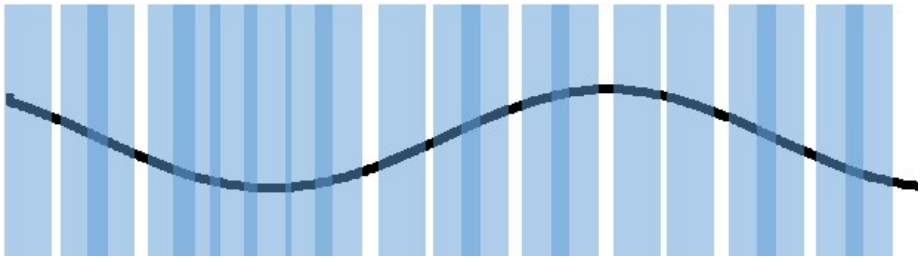
We used the same data processing method described in Experiment 1, except that we extracted the four features from each data sequence. Each participant experienced high and low workload portions in six trials. We evaluated our proposed Bayesian inference model against other single models in two different evaluation methods: cross-participants evaluation and within-participants evaluation. For the cross-participants evaluation, we randomly selected five sequences of data from each portion in each trial, with each sequence lasting 4 s. For the within-participants evaluation, we randomly selected 20 sequences of data from each portion in each trial. Figure 4.4 shows an example of sequences of data selected in one example portion of a track using cross-participants evaluation and within-participants evaluation. The blue boxes represent for the selected sequences of data. Note that, as we selected more sequences of data for within-participants evaluation, the probability of overlapping between the randomly selected sequences of data is higher.

4.3.2 Cross-participants Evaluation

The cross-participants evaluation separates the training data and testing data across the participants (i.e., data from some participants are treated as training data and data



(a) Cross-participants evaluation: 5 sequences of data from each portion



(b) Within-participants evaluation: 20 sequences of data from each portion

Figure 4.4: An example for sequences of data selected from a portion. Blue boxes represents the randomly selected sequences of data, each lasts 4 s. (a) 5 sequences of data were selected for each portion using cross-participants evaluation. (b) 20 sequences of data were selected for each portion using within-participants evaluation.

from other participants are treated as testing data). We used the leave-one-out evaluation method for cross-participants evaluation. Specifically, we randomly selected the data of six participants as the testing dataset and the data of the remaining 18 participants as the training dataset in each run of the holdout. We ran 50 holdouts to evaluate the performance of our proposed Bayesian inference model (BI) and the four single models. In each round of holdouts, we computed the means (μ_i) and standard deviations (σ_i) for every feature (X_i) using the training dataset, and then normalized all the data using these means and standard deviations, i.e., $\hat{X}_i = \frac{X_i - \mu_i}{\sigma_i}$. To obtain the prior knowledge $p(M_i|W_L)$ of each machine learning model M_i, we ran 10 rounds of leave-one-out evaluation over the training dataset with 18 participants. In each round, we randomly selected 12 participants from the 18 participants as prior training data and the remaining six participants as validation data. We then computed the confusion matrix of each machine learning model on the validation

data to obtain the estimated prior knowledge $p(M_i|W_L)$.

For cross-participants evaluation, we computed two types of performance, overall performance and individual performance, for our proposed Bayesian inference model and other single models. For overall performance, we computed the performance (i.e., F_1 score, precision, and recall) for the entire testing dataset in each round of holdouts and the overall performance shown in Table 4.1 is the mean and standard error over the 50 rounds of holdouts. For individual performance, we computed the performance for each individual participant in the testing dataset in each round of holdouts and the individual performance shown in Table 4.3 is the mean and standard error for every participant when they were in the testing dataset. As we randomly determine the training dataset and testing dataset in each round, the numbers of times when each participant was in the testing dataset were different. For the individual performance, we also computed the average individual performance over all participants at the end of Table 4.3. Table 4.2 shows the pairwise t-test results for the overall performance between our proposed Bayesian inference model (BI) and other single models for cross-participants evaluation.

The results indicated that our proposed Bayesian inference model significantly outperforms the single models alone for both overall performance and average individual performance using cross-participants evaluation. Our proposed Bayesian inference model achieved 0.823 ± 0.004 and 0.83 ± 0.02 F_1 scores for overall performance and average individual performance using cross-participants, respectively. Note that not every participant had a similar performance as shown in Figure 4.3. For example, participant 9 only had a 0.65 F_1 score; however, participant 4 achieved an F_1 score as high as 0.95.

Table 4.1: Overall performance of the Bayesian inference model and other single models for cross-participants evaluation.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.823 ± 0.004	0.772 ± 0.006	0.653 ± 0.005	0.745 ± 0.003	0.674 ± 0.005
Precision	0.824 ± 0.004	0.773 ± 0.006	0.656 ± 0.005	0.749 ± 0.003	0.679 ± 0.006
Recall	0.821 ± 0.004	0.771 ± 0.006	0.650 ± 0.005	0.741 ± 0.003	0.668 ± 0.005

Table 4.2: Pairwise t-tests between Bayesian inference model (BI) and other single models.

	BI vs. SVM pupil size change	BI vs. HMM gaze trajectory	BI vs. SVM fixation feature	BI vs. GMMs fixation trajectory
F_1 score	$t(49) = 10.66, p < .001$	$t(49) = 37.85, p < .001$	$t(49) = 22.99, p < .001$	$t(49) = 32.17, p < .001$
Precision	$t(49) = 10.95, p < .001$	$t(49) = 35.24, p < .001$	$t(49) = 21.41, p < .001$	$t(49) = 29.97, p < .001$
Recall	$t(49) = 10.34, p < .001$	$t(49) = 39.38, p < .001$	$t(49) = 24.12, p < .001$	$t(49) = 32.70, p < .001$

Table 4.3: Individual performance (F_1 score, precision, and recall) of the Bayesian inference model (BI) and other single models for cross-participants evaluation.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.81 ± 0.00	0.78 ± 0.00	0.60 ± 0.00	0.74 ± 0.01	0.67 ± 0.00
	0.81 ± 0.01	0.78 ± 0.00	0.60 ± 0.00	0.75 ± 0.01	0.69 ± 0.01
	0.80 ± 0.00	0.78 ± 0.00	0.60 ± 0.00	0.72 ± 0.01	0.65 ± 0.00
P2	0.86 ± 0.01	0.92 ± 0.01	0.71 ± 0.00	0.72 ± 0.00	0.72 ± 0.01
	0.88 ± 0.00	0.92 ± 0.01	0.75 ± 0.00	0.75 ± 0.00	0.74 ± 0.01
	0.85 ± 0.01	0.92 ± 0.01	0.68 ± 0.01	0.69 ± 0.00	0.71 ± 0.01
P3	0.74 ± 0.01	0.68 ± 0.01	0.53 ± 0.01	0.69 ± 0.01	0.58 ± 0.01
	0.75 ± 0.01	0.68 ± 0.01	0.55 ± 0.01	0.69 ± 0.01	0.59 ± 0.02
	0.73 ± 0.01	0.68 ± 0.01	0.52 ± 0.01	0.69 ± 0.01	0.58 ± 0.01
P4	0.95 ± 0.00	0.94 ± 0.00	0.69 ± 0.00	0.80 ± 0.01	0.81 ± 0.01
	0.96 ± 0.00	0.94 ± 0.00	0.69 ± 0.01	0.81 ± 0.00	0.82 ± 0.01
	0.95 ± 0.00	0.94 ± 0.01	0.69 ± 0.00	0.80 ± 0.01	0.79 ± 0.01
P5	0.88 ± 0.01	0.85 ± 0.01	0.76 ± 0.01	0.80 ± 0.00	0.74 ± 0.02
	0.89 ± 0.01	0.85 ± 0.01	0.76 ± 0.01	0.81 ± 0.00	0.80 ± 0.01
	0.88 ± 0.01	0.85 ± 0.01	0.76 ± 0.01	0.79 ± 0.01	0.69 ± 0.02
P6	0.81 ± 0.01	0.78 ± 0.01	0.58 ± 0.01	0.72 ± 0.01	0.52 ± 0.02
	0.82 ± 0.01	0.83 ± 0.00	0.59 ± 0.01	0.72 ± 0.01	0.52 ± 0.03
	0.79 ± 0.01	0.74 ± 0.01	0.57 ± 0.01	0.72 ± 0.01	0.52 ± 0.02
P7	0.74 ± 0.01	0.56 ± 0.01	0.76 ± 0.00	0.71 ± 0.01	0.70 ± 0.01
	0.75 ± 0.01	0.56 ± 0.01	0.77 ± 0.01	0.71 ± 0.01	0.71 ± 0.01
	0.74 ± 0.01	0.56 ± 0.01	0.75 ± 0.00	0.71 ± 0.01	0.69 ± 0.01
P8	0.83 ± 0.02	0.75 ± 0.03	0.74 ± 0.01	0.74 ± 0.00	0.77 ± 0.01
	0.83 ± 0.02	0.75 ± 0.03	0.74 ± 0.01	0.74 ± 0.00	0.78 ± 0.01

Continued on next page

Table 4.3 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.82 ± 0.02	0.75 ± 0.03	0.73 ± 0.01	0.74 ± 0.00	0.77 ± 0.01
P9	0.65 ± 0.01	0.56 ± 0.01	0.65 ± 0.01	0.67 ± 0.00	0.58 ± 0.02
	0.65 ± 0.01	0.56 ± 0.01	0.66 ± 0.01	0.68 ± 0.00	0.59 ± 0.02
	0.65 ± 0.01	0.56 ± 0.01	0.65 ± 0.01	0.66 ± 0.00	0.58 ± 0.02
P10	0.81 ± 0.01	0.77 ± 0.01	0.70 ± 0.00	0.74 ± 0.00	0.61 ± 0.01
	0.83 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.80 ± 0.00	0.65 ± 0.02
	0.79 ± 0.01	0.75 ± 0.01	0.63 ± 0.00	0.68 ± 0.00	0.58 ± 0.01
P11	0.79 ± 0.00	0.73 ± 0.01	0.55 ± 0.01	0.78 ± 0.00	0.60 ± 0.03
	0.81 ± 0.00	0.73 ± 0.01	0.56 ± 0.01	0.79 ± 0.00	0.61 ± 0.03
	0.78 ± 0.00	0.72 ± 0.01	0.55 ± 0.01	0.78 ± 0.00	0.60 ± 0.03
P12	0.81 ± 0.01	0.85 ± 0.01	0.63 ± 0.01	0.73 ± 0.00	0.65 ± 0.01
	0.84 ± 0.01	0.85 ± 0.01	0.67 ± 0.01	0.77 ± 0.00	0.66 ± 0.01
	0.79 ± 0.01	0.84 ± 0.01	0.59 ± 0.01	0.68 ± 0.00	0.63 ± 0.01
P13	0.93 ± 0.01	0.80 ± 0.01	0.73 ± 0.00	0.76 ± 0.01	0.67 ± 0.01
	0.93 ± 0.01	0.80 ± 0.01	0.73 ± 0.00	0.79 ± 0.01	0.71 ± 0.01
	0.92 ± 0.01	0.79 ± 0.01	0.73 ± 0.00	0.74 ± 0.01	0.63 ± 0.01
P14	0.79 ± 0.01	0.81 ± 0.01	0.68 ± 0.01	0.66 ± 0.00	0.67 ± 0.01
	0.79 ± 0.01	0.81 ± 0.01	0.70 ± 0.01	0.67 ± 0.00	0.69 ± 0.01
	0.78 ± 0.01	0.80 ± 0.02	0.67 ± 0.01	0.66 ± 0.00	0.65 ± 0.01
P15	0.87 ± 0.01	0.81 ± 0.01	0.70 ± 0.01	0.81 ± 0.00	0.78 ± 0.01
	0.87 ± 0.01	0.81 ± 0.01	0.70 ± 0.01	0.81 ± 0.00	0.80 ± 0.01
	0.87 ± 0.01	0.81 ± 0.01	0.69 ± 0.01	0.80 ± 0.00	0.77 ± 0.01
P16	0.91 ± 0.01	0.85 ± 0.01	0.66 ± 0.01	0.88 ± 0.01	0.63 ± 0.02
	0.92 ± 0.01	0.86 ± 0.01	0.68 ± 0.01	0.88 ± 0.01	0.63 ± 0.02
	0.91 ± 0.01	0.85 ± 0.01	0.63 ± 0.01	0.88 ± 0.01	0.62 ± 0.02
P17	0.81 ± 0.01	0.87 ± 0.01	0.53 ± 0.04	0.69 ± 0.00	0.58 ± 0.02
	0.83 ± 0.01	0.88 ± 0.01	0.60 ± 0.07	0.70 ± 0.00	0.59 ± 0.02
	0.80 ± 0.01	0.86 ± 0.01	0.51 ± 0.00	0.68 ± 0.00	0.58 ± 0.02
P18	0.90 ± 0.01	0.80 ± 0.01	0.71 ± 0.00	0.90 ± 0.00	0.80 ± 0.01
	0.91 ± 0.01	0.80 ± 0.01	0.78 ± 0.01	0.91 ± 0.00	0.81 ± 0.01
	0.90 ± 0.01	0.80 ± 0.01	0.65 ± 0.01	0.89 ± 0.00	0.80 ± 0.01
P19	0.92 ± 0.01	0.87 ± 0.00	0.83 ± 0.01	0.85 ± 0.01	0.78 ± 0.01
	0.92 ± 0.01	0.87 ± 0.00	0.83 ± 0.01	0.86 ± 0.00	0.78 ± 0.01
	0.92 ± 0.01	0.86 ± 0.00	0.82 ± 0.01	0.85 ± 0.01	0.78 ± 0.01
P20	0.85 ± 0.01	0.68 ± 0.02	0.71 ± 0.01	0.82 ± 0.01	0.79 ± 0.02
	0.85 ± 0.01	0.68 ± 0.02	0.71 ± 0.01	0.83 ± 0.01	0.80 ± 0.02
	0.84 ± 0.01	0.67 ± 0.03	0.71 ± 0.01	0.82 ± 0.01	0.78 ± 0.02
P21	0.81 ± 0.01	0.79 ± 0.01	0.67 ± 0.01	0.71 ± 0.00	0.75 ± 0.01
	0.81 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.71 ± 0.00	0.76 ± 0.01
	0.80 ± 0.01	0.79 ± 0.01	0.61 ± 0.00	0.71 ± 0.00	0.74 ± 0.02

Continued on next page

Table 4.3 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P22	0.95 ± 0.00	0.87 ± 0.02	0.83 ± 0.01	0.91 ± 0.00	0.75 ± 0.01
	0.95 ± 0.00	0.87 ± 0.02	0.86 ± 0.01	0.91 ± 0.00	0.75 ± 0.02
	0.95 ± 0.00	0.87 ± 0.02	0.81 ± 0.01	0.91 ± 0.00	0.74 ± 0.01
P23	0.83 ± 0.01	0.77 ± 0.01	0.59 ± 0.03	0.75 ± 0.00	0.69 ± 0.01
	0.86 ± 0.01	0.77 ± 0.01	0.70 ± 0.05	0.77 ± 0.01	0.74 ± 0.01
	0.81 ± 0.01	0.77 ± 0.01	0.52 ± 0.00	0.73 ± 0.00	0.64 ± 0.01
P24	0.70 ± 0.01	0.60 ± 0.01	0.60 ± 0.02	0.64 ± 0.00	0.54 ± 0.02
	0.72 ± 0.01	0.60 ± 0.01	0.70 ± 0.03	0.68 ± 0.00	0.56 ± 0.03
	0.68 ± 0.01	0.60 ± 0.01	0.52 ± 0.00	0.60 ± 0.00	0.52 ± 0.02
Avg	0.83 ± 0.02	0.78 ± 0.02	0.67 ± 0.02	0.76 ± 0.02	0.68 ± 0.02
	0.84 ± 0.02	0.78 ± 0.02	0.70 ± 0.02	0.77 ± 0.01	0.70 ± 0.02
	0.82 ± 0.02	0.77 ± 0.02	0.65 ± 0.02	0.75 ± 0.02	0.67 ± 0.02

4.3.3 Within-participants Evaluation

The within-participants evaluation evaluates the performance for each participant (i.e., a personalized model), and it separates the training data and testing data across the trials for each participant (i.e., data from some trials are treated as training data and data from other trials are treated as testing data). We used k-fold cross validation for the within-participants evaluation, where k equals 6 as there was 6 trials for each workload level. Specifically, we used data from one of the six trials as testing data and data from the other trials as training data. Similar to the cross-participants evaluation, we used the training data to obtain the estimated prior knowledge $p(M_i|W_L)$, except that we used five-fold cross validation over the five training trials.

Table 4.4 shows the performance (i.e., F_1 score, precision and recall) of our proposed Bayesian inference model and other single models for each participant and the average performance. The results indicated that our proposed Bayesian inference model achieved a 0.85 F_1 score on average using within-participants evaluation.

Table 4.4: Performance (F_1 score, precision, and recall) of the Bayesian inference model (BI) and other single models for within-participants evaluation.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.78 ± 0.03	0.77 ± 0.02	0.69 ± 0.07	0.67 ± 0.05	0.67 ± 0.04
	0.79 ± 0.03	0.77 ± 0.02	0.70 ± 0.07	0.69 ± 0.05	0.68 ± 0.04
	0.77 ± 0.03	0.76 ± 0.03	0.67 ± 0.07	0.66 ± 0.05	0.65 ± 0.04
P2	0.95 ± 0.02	0.95 ± 0.02	0.62 ± 0.07	0.72 ± 0.03	0.67 ± 0.03
	0.95 ± 0.02	0.95 ± 0.02	0.62 ± 0.08	0.74 ± 0.03	0.69 ± 0.03
	0.95 ± 0.02	0.95 ± 0.02	0.62 ± 0.07	0.70 ± 0.03	0.66 ± 0.03
P3	0.82 ± 0.05	0.81 ± 0.04	0.74 ± 0.07	0.75 ± 0.05	0.73 ± 0.07
	0.83 ± 0.04	0.83 ± 0.04	0.78 ± 0.07	0.76 ± 0.04	0.74 ± 0.07
	0.80 ± 0.05	0.80 ± 0.05	0.71 ± 0.07	0.73 ± 0.05	0.71 ± 0.07
P4	0.94 ± 0.01	0.93 ± 0.01	0.75 ± 0.05	0.87 ± 0.02	0.81 ± 0.03
	0.94 ± 0.01	0.93 ± 0.01	0.76 ± 0.05	0.88 ± 0.02	0.82 ± 0.04
	0.94 ± 0.01	0.92 ± 0.01	0.74 ± 0.05	0.86 ± 0.03	0.81 ± 0.03
P5	0.90 ± 0.02	0.86 ± 0.02	0.68 ± 0.05	0.86 ± 0.03	0.81 ± 0.02
	0.90 ± 0.02	0.87 ± 0.03	0.69 ± 0.06	0.86 ± 0.03	0.82 ± 0.02
	0.89 ± 0.02	0.86 ± 0.02	0.67 ± 0.05	0.85 ± 0.03	0.80 ± 0.03
P6	0.80 ± 0.02	0.76 ± 0.02	0.52 ± 0.04	0.73 ± 0.03	0.60 ± 0.05
	0.80 ± 0.02	0.77 ± 0.02	0.52 ± 0.04	0.74 ± 0.03	0.61 ± 0.05
	0.79 ± 0.02	0.76 ± 0.02	0.52 ± 0.04	0.72 ± 0.03	0.60 ± 0.05
P7	0.78 ± 0.04	0.59 ± 0.05	0.69 ± 0.07	0.77 ± 0.03	0.68 ± 0.07
	0.78 ± 0.04	0.59 ± 0.05	0.69 ± 0.07	0.79 ± 0.03	0.68 ± 0.07
	0.77 ± 0.03	0.59 ± 0.05	0.69 ± 0.07	0.76 ± 0.04	0.67 ± 0.07
P8	0.82 ± 0.03	0.82 ± 0.03	0.71 ± 0.04	0.76 ± 0.08	0.76 ± 0.06
	0.83 ± 0.03	0.83 ± 0.03	0.76 ± 0.04	0.78 ± 0.08	0.78 ± 0.06
	0.81 ± 0.03	0.82 ± 0.03	0.68 ± 0.04	0.75 ± 0.08	0.74 ± 0.06
P9	0.74 ± 0.07	0.65 ± 0.02	0.69 ± 0.05	0.67 ± 0.06	0.70 ± 0.04
	0.75 ± 0.07	0.65 ± 0.02	0.70 ± 0.06	0.67 ± 0.06	0.71 ± 0.04
	0.74 ± 0.07	0.65 ± 0.02	0.67 ± 0.05	0.66 ± 0.05	0.69 ± 0.04
P10	0.90 ± 0.02	0.85 ± 0.01	0.78 ± 0.04	0.75 ± 0.03	0.86 ± 0.04
	0.90 ± 0.02	0.86 ± 0.01	0.79 ± 0.04	0.75 ± 0.03	0.86 ± 0.04
	0.90 ± 0.02	0.84 ± 0.02	0.77 ± 0.05	0.74 ± 0.03	0.86 ± 0.04
P11	0.84 ± 0.06	0.66 ± 0.06	0.69 ± 0.06	0.81 ± 0.05	0.77 ± 0.05
	0.85 ± 0.06	0.67 ± 0.06	0.70 ± 0.05	0.81 ± 0.05	0.78 ± 0.05
	0.84 ± 0.06	0.66 ± 0.05	0.67 ± 0.06	0.80 ± 0.05	0.75 ± 0.05
P12	0.94 ± 0.03	0.93 ± 0.02	0.76 ± 0.04	0.76 ± 0.03	0.83 ± 0.06
	0.94 ± 0.03	0.93 ± 0.02	0.78 ± 0.04	0.78 ± 0.03	0.84 ± 0.06
	0.94 ± 0.03	0.93 ± 0.02	0.74 ± 0.04	0.75 ± 0.03	0.82 ± 0.06
P13	0.86 ± 0.03	0.75 ± 0.05	0.67 ± 0.03	0.85 ± 0.03	0.61 ± 0.04
	0.87 ± 0.03	0.76 ± 0.05	0.68 ± 0.03	0.86 ± 0.03	0.61 ± 0.04

Continued on next page

Table 4.4 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.86 ± 0.03	0.75 ± 0.06	0.66 ± 0.03	0.84 ± 0.03	0.60 ± 0.04
P14	0.79 ± 0.04	0.74 ± 0.06	0.64 ± 0.05	0.58 ± 0.05	0.76 ± 0.05
	0.80 ± 0.04	0.74 ± 0.06	0.65 ± 0.05	0.58 ± 0.05	0.76 ± 0.05
	0.79 ± 0.04	0.73 ± 0.06	0.62 ± 0.04	0.58 ± 0.05	0.76 ± 0.05
P15	0.88 ± 0.03	0.76 ± 0.04	0.60 ± 0.04	0.84 ± 0.03	0.73 ± 0.05
	0.89 ± 0.03	0.76 ± 0.04	0.64 ± 0.05	0.84 ± 0.03	0.74 ± 0.05
	0.88 ± 0.03	0.75 ± 0.04	0.57 ± 0.03	0.83 ± 0.02	0.72 ± 0.05
P16	0.84 ± 0.03	0.79 ± 0.05	0.73 ± 0.05	0.81 ± 0.03	0.75 ± 0.05
	0.84 ± 0.03	0.80 ± 0.05	0.74 ± 0.04	0.82 ± 0.03	0.75 ± 0.05
	0.83 ± 0.03	0.79 ± 0.05	0.72 ± 0.05	0.80 ± 0.03	0.75 ± 0.05
P17	0.88 ± 0.03	0.85 ± 0.03	0.67 ± 0.06	0.73 ± 0.06	0.67 ± 0.04
	0.88 ± 0.03	0.86 ± 0.03	0.67 ± 0.06	0.74 ± 0.06	0.68 ± 0.04
	0.87 ± 0.03	0.85 ± 0.04	0.67 ± 0.06	0.72 ± 0.06	0.67 ± 0.04
P18	0.88 ± 0.02	0.86 ± 0.01	0.66 ± 0.06	0.82 ± 0.03	0.76 ± 0.04
	0.89 ± 0.02	0.87 ± 0.01	0.67 ± 0.06	0.82 ± 0.03	0.76 ± 0.04
	0.88 ± 0.02	0.86 ± 0.01	0.66 ± 0.06	0.82 ± 0.03	0.75 ± 0.04
P19	0.86 ± 0.03	0.77 ± 0.04	0.64 ± 0.04	0.80 ± 0.01	0.75 ± 0.02
	0.87 ± 0.03	0.78 ± 0.04	0.65 ± 0.04	0.81 ± 0.01	0.76 ± 0.03
	0.86 ± 0.03	0.76 ± 0.04	0.63 ± 0.04	0.80 ± 0.01	0.74 ± 0.02
P20	0.85 ± 0.02	0.69 ± 0.03	0.77 ± 0.04	0.82 ± 0.02	0.80 ± 0.02
	0.85 ± 0.02	0.70 ± 0.03	0.79 ± 0.04	0.83 ± 0.02	0.81 ± 0.02
	0.84 ± 0.03	0.68 ± 0.03	0.75 ± 0.05	0.80 ± 0.02	0.78 ± 0.02
P21	0.90 ± 0.03	0.88 ± 0.03	0.70 ± 0.05	0.76 ± 0.04	0.66 ± 0.04
	0.90 ± 0.03	0.88 ± 0.03	0.72 ± 0.05	0.77 ± 0.03	0.68 ± 0.05
	0.90 ± 0.03	0.88 ± 0.03	0.68 ± 0.05	0.75 ± 0.04	0.65 ± 0.04
P22	0.92 ± 0.03	0.83 ± 0.04	0.66 ± 0.07	0.89 ± 0.02	0.80 ± 0.03
	0.92 ± 0.03	0.84 ± 0.04	0.67 ± 0.09	0.90 ± 0.02	0.81 ± 0.03
	0.92 ± 0.03	0.82 ± 0.04	0.67 ± 0.05	0.89 ± 0.02	0.80 ± 0.03
P23	0.86 ± 0.02	0.80 ± 0.02	0.68 ± 0.10	0.81 ± 0.04	0.81 ± 0.05
	0.87 ± 0.02	0.80 ± 0.02	0.67 ± 0.11	0.82 ± 0.04	0.82 ± 0.05
	0.85 ± 0.02	0.80 ± 0.02	0.70 ± 0.08	0.80 ± 0.04	0.80 ± 0.05
P24	0.69 ± 0.08	0.69 ± 0.06	0.64 ± 0.05	0.70 ± 0.08	0.68 ± 0.04
	0.69 ± 0.09	0.70 ± 0.07	0.64 ± 0.05	0.70 ± 0.08	0.68 ± 0.04
	0.69 ± 0.08	0.69 ± 0.06	0.63 ± 0.05	0.69 ± 0.08	0.67 ± 0.04
Avg	0.85 ± 0.01	0.79 ± 0.02	0.68 ± 0.01	0.77 ± 0.01	0.74 ± 0.01
	0.86 ± 0.01	0.80 ± 0.02	0.69 ± 0.01	0.78 ± 0.01	0.74 ± 0.01
	0.85 ± 0.01	0.79 ± 0.02	0.67 ± 0.01	0.76 ± 0.01	0.73 ± 0.01

4.4 Discussion

As shown in the results, our proposed Bayesian inference model outperforms other single models alone in both the cross-participants evaluation and the within-participants evaluation.

In the cross-participants evaluation, the overall performance of SVMs for pupil size change is better than other single models (i.e., the HMM for gaze trajectory, SVMs for fixation feature, and GMMs for fixation trajectory, as shown in Table 4.1). However, this may not hold for each individual participant. For example, for participant 20, the SVMs for pupil size change had an F_1 score of 0.68 ± 0.02, which is lower than the SVMs for fixation features (0.82 ± 0.01) and the GMMs for fixation trajectory (0.79 ± 0.02) in Table 4.3 for the individual performance using cross-participants evaluation. This indicates the necessity of leveraging different machine learning models for different features. Therefore, our proposed Bayesian inference model is useful for achieving reliable workload estimation for different people.

The performance for SVMs for pupil size change, HMM for gaze trajectory, and SVMs for fixation feature using cross-participants evaluation is similar with those using within-participants evaluation as shown in Table 4.3, Table 4.1, and Table 4.4. However, the performance for GMMs for fixation trajectory using within-participants evaluation (0.74 ± 0.01 F_1 score) is better than using cross-participants evaluation (0.68 ± 0.02 F_1 score). This finding indicates that analyzing fixation trajectory may not work well for the cross-participants evaluation. This may be due to the different ways people use to acquire and process information on the screen, i.e., balancing driving task and surveillance task as well as finding potential threats in the four images in the surveillance task. Different ways to acquire and process information may lead to different patterns of fixations and saccades.

The cross-participants evaluation and the within-participants evaluation have its advantages and disadvantages, and therefore are particularly suitable for certain contexts. The cross-participants evaluation can be considered a "population-based" model. Its underlying premise is that a "population" model is generalizable to any human operator. This approach, on one hand, is convenient to use. On the other hand, it requires another set of data for training the model. For instance, Experiment 1 was conducted first to train the workload estimation model, which was used later in Experiment 2. The within-participants evaluation can be considered a "personalized" model. Using this approach, a portion of data collected from one participant was used to train a model that work well for this particular participant. On average, within-participants evaluation provides better performance than cross-participants evaluation. In addition, within-participants evaluation does not need an extra experiment beforehand to collect data to train the build. However, this approach requires more trials for each participant and hence much longer experiment time.

Table 4.3 shows that the performance for Bayesian inference model is worse than at least one single model alone for six participants using the cross-participants evaluation, i.e., P2, P7, P9, P12, P14, and P17. However, the performance for Bayesian inference model is equal or better than all single models alone for all participants using the within-participants evaluation. This indicates that using within-participants evaluation, we could acquire more accurate prior distribution $p(M_i|W_L)$, resulting in higher chance to outperform other single models. We speculate that the performance increment using within-participants is because of the more accurate prior distribution and better single model performance (GMMs for fixation trajectory).

Although the results indicated that our proposed Bayesian inference model can achieve an overall 0.82 F_1 score using cross-participants evaluation and an average 0.85 F_1 score

using within-participants evaluation for workload estimation, there are still limitations to the study. In this chapter, the different levels of human workload is induced by manipulating the surveillance task urgency. The results indicated that our proposed Bayesian inference model is able to distinguish the different workload levels caused by different surveillance task urgencies. However, it is unclear if the proposed Bayesian inference model can classify different workload levels caused by other factors, such as different driving task conditions. We address this research question in Chapter V.

CHAPTER V

Generalizability for Bayesian Inference Model for Workload estimation

5.1 Introduction

In Chapter IV, we evaluated the proposed Bayesian inference model for workload estimation using data collected in Experiment 1 in Chapter III, and the different workload levels were induced by manipulating the surveillance task urgency. In this chapter, we investigate the generalizability of the proposed Bayesian inference model for workload estimation by conducting two more experiments: Experiment 3 and Experiment 4. In Experiment 3, we introduce obstacle avoidance to the driving task to investigate the effects of different obstacle parameters (i.e., headway and size) on workload estimation performance. In Experiment 4, we investigate the effects of a vehicle's speed on workload estimation performance in the driving task.

5.2 Experiment 3: Effects of Obstacle Avoidance on Workload Estimation Performance

5.2.1 Method

Participants

A total of 20 students participated in the experiment. Data of eight participants were discarded due to an equipment malfunction. The remaining 12 participants were on aver-

age 22.7 years old (SD = 2.6) and had an average of 4.5 years of driving experience (SD = 2.2). There were 5 females and 7 males in the remaining 12 participants. All participants had normal or corrected-to-normal vision.

Experimental Apparatus and Stimuli

We used a similar dual-task platform as the one mentioned in Section 2.2 except for the following changes:

1) In the surveillance task, we changed the fixed time limits for the detection period to a combined pace design: If the participant responded within 8 seconds, there would be a gap until the display of the next set of four images; if the participant responded after 8 seconds, the next set of images would be displayed immediately. There was an auditory alert every 3 seconds to remind the participant of the secondary task. Figure 5.1 illustrates the combined pace design of the surveillance task, where R_t denotes the human operator's response time, $A_t = 3$ s is the alert time, $a = 8$ s is a parameter that limits the participant's pace on the visual search task, and W_t is the gap between the display of the current set of images and the display of the next set of images. We define that $W_t = \max(a - R_t, 0)$. Thus, if the human operator's response time R_t is smaller than $a = 8$ s, a white image will be shown for $a - R_t$ seconds; if the human operator's response time R_t is larger than $a = 8$ s, the next set of images will appear immediately. We modified the surveillance task from fixed time limits for the detection period to the combined pace design as we aimed to manipulate human workload in the driving task and gave human the control of the surveillance task pace.

2) In the driving task, we introduced some obstacles. We disabled the obstacle avoidance function in the shared control autonomy to simulate the perception difficulties that autonomy has and therefore the human operator was responsible for the obstacle avoid-

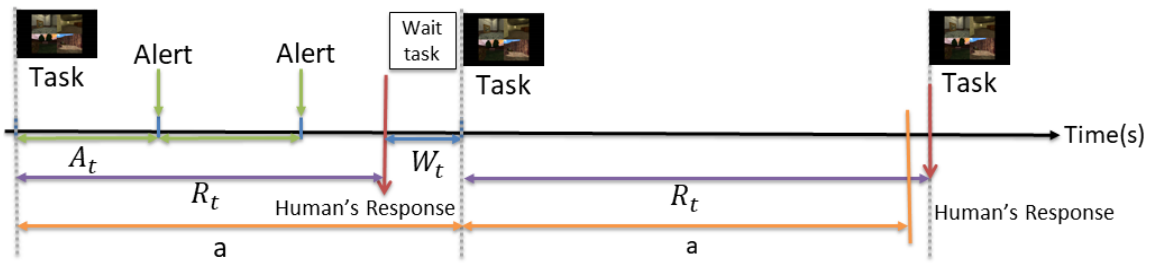


Figure 5.1: Illustration of the combined pace design for the surveillance task.

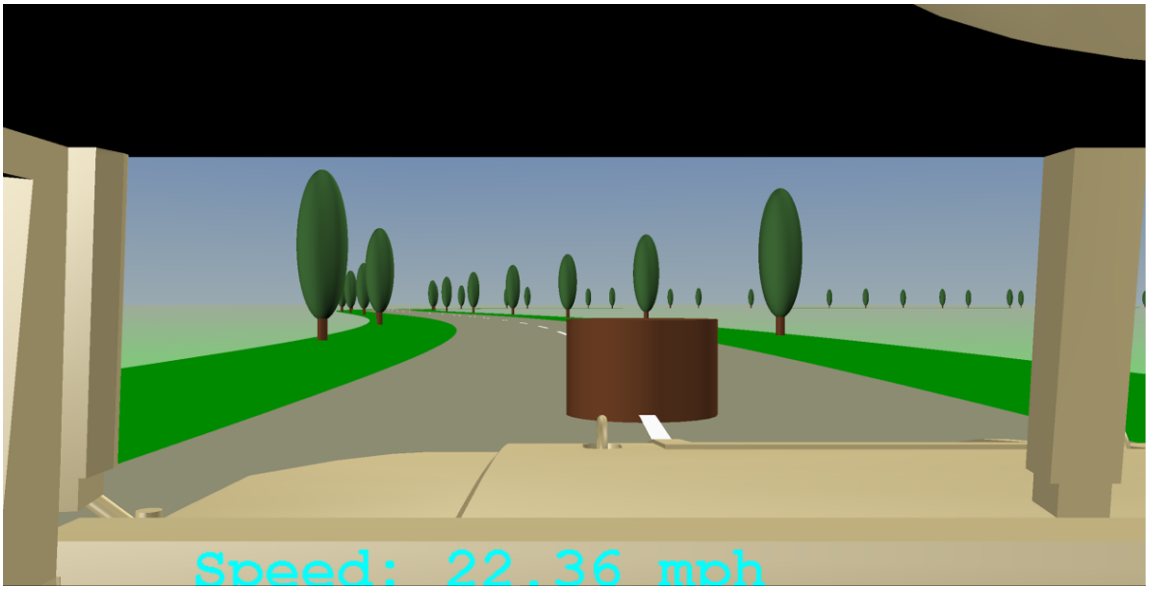


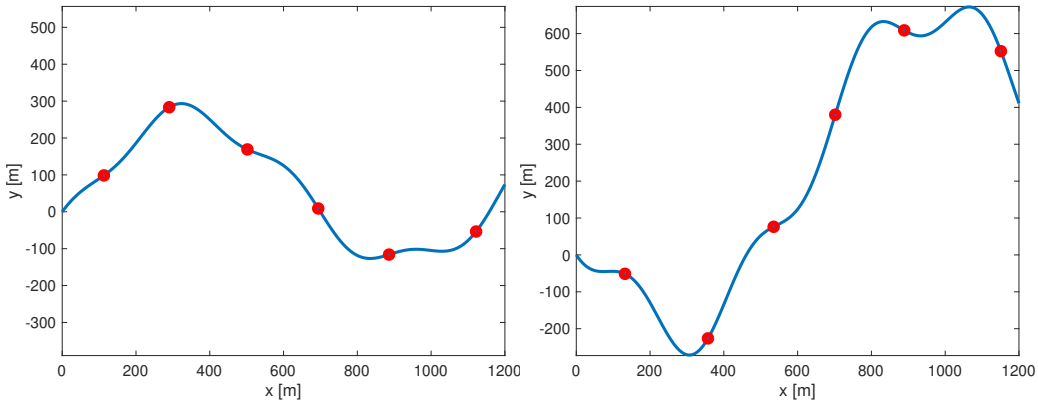
Figure 5.2: Example of an obstacle in the driving task.

ance task. We removed the localization offset for the autonomy mentioned in Figure 2.5 from Section 2.4.1. The obstacles were cylinders in the center of the road, as shown in Figure 5.2. The non-adaptive shared control scheme was applied to control the speed of the vehicle at 10 m/s (around 22 mph). The human operator and the autonomy shared the control of the steering wheel.

Experimental Design

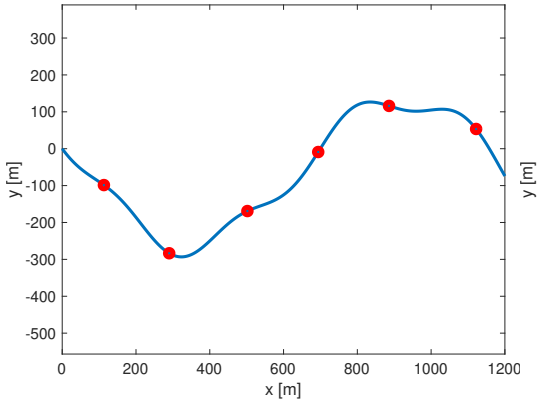
The experiment used a within-subjects design. Each participant drove on four different tracks, and each track was 1,000 m long. There were six different obstacles on the track, with varying obstacle sizes (1, 3, 5 meters in diameter diameter) and headways (2.5 and

8 seconds). The headway indicates how far away the participant can see the obstacle and perform obstacle avoidance. To eliminate potential order effects, the presentation order of the six obstacles followed a 6×6 Latin square. Figure 5.3 shows the four tracks and the obstacle locations in the experiment. The blue curves are the tracks, and the red dots indicate the locations of the obstacles.

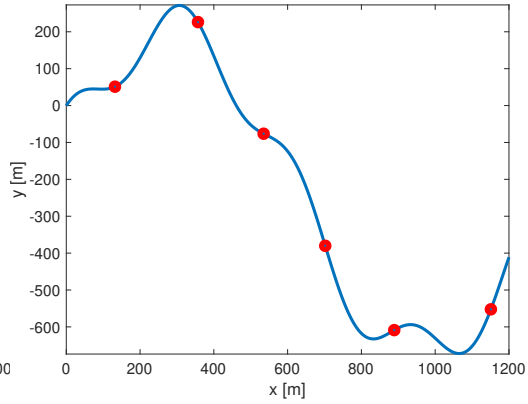


(a) Track 1

(b) Track 2



(c) Track 3



(d) Track 4

Figure 5.3: Four tracks in the formal experiment (blue curves). Red dots indicate the locations of the obstacles.

Measures

During the experiment, the participants wore the Tobii Pro Glasses 2 to gather their eye-related data (i.e., gaze points and pupil sizes). After avoiding each obstacle, participants reported perceived difficulty on a 7-point Likert scale.

Experimental Procedure

The participants provided informed consent and filled in a demographic survey prior to the experiment. The participants' baseline pupil sizes were then collected by asking them to look at a white wall twice, each time for 30 seconds. Four training trials were provided to them before the real experiment: (1) driving on a track without obstacles; (2) driving on a track with obstacles; (3) performing the surveillance task; (4) performing the driving task and the surveillance task on a track with obstacles. Participants then performed the formal experiment on four different tracks. A debriefing survey was completed at the end of the experiment.

5.2.2 Results

Effects of Obstacle Types on Perceived Difficulty

Two-way repeated measures analysis of variance (ANOVAs) were conducted with the obstacle headway and the obstacle size as the within-subjects variables. Results are reported as significant for $\alpha < .05$.

The results revealed a significant effect of obstacle headway ($F(1, 11) = 101.928$, $p < .001$) and obstacle size ($F(2, 22) = 17.025$, $p < .001$) on perceived difficulty. Figure 5.4 shows the mean and standard error (SE) values of perceived difficulty. We then performed a series of t tests between different pairs of obstacles. The results revealed significant differences between each pair of obstacles except for the obstacle with 8 s headway/1 m diameter and the obstacle with 8 s headway/3 m diameter.

Data Preparation

Given the results above, we considered the event of avoiding obstacles with a 2.5-second headway as imposing a high workload on human operators and the event of avoiding obstacles with an 8-second headway as imposing a low workload. We defined the

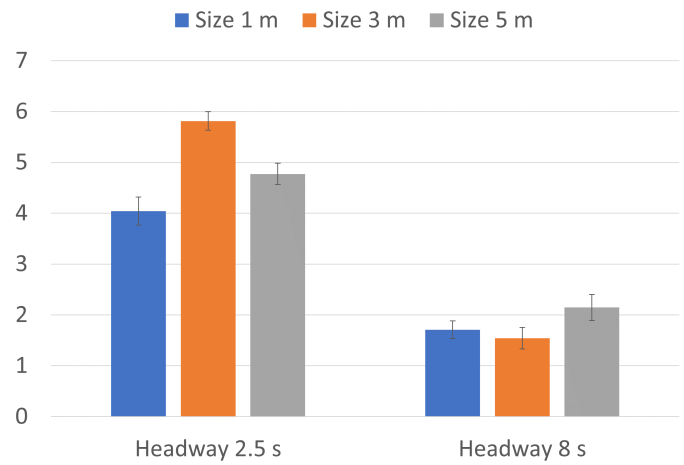


Figure 5.4: Mean and standard error (SE) values of perceived difficulty.

period of the obstacle-avoidance event as from 2 seconds prior to revealing an obstacle to 2 seconds after the vehicle passed the obstacle and crossed the centerline, as shown in Figure 5.5. Black curve represents the track and blue dotted line represents the vehicle trajectory. Yellow cross mark indicates when the obstacle is first revealed to the participants. Yellow circle represents 2 seconds before obstacle revealed. Green cross mark indicates the first time that the vehicle passed the centerline after obstacle avoidance. Green circle represents 2 seconds after vehicle passed the obstacle and the centerline. The obstacle avoidance event is the period between the yellow circle and the green circle.

The sampling rate for the Tobii Pro Glasses 2 is 50 Hz. In the case of data dropout, we down-sampled the data to a 30 Hz sampling rate. We used data of 4 seconds in the middle of each obstacle avoidance event (shown as the blue box in Figure 5.5) and had 288 data points (12 participants $\times$ 4 tracks $\times$ 6 obstacles).

Workload Estimation Performance

As each participant only has 12 data points for each workload level, we only used cross-participants evaluation in this experiment. Due to the small dataset (12 participants), we used the holdout method (Kim, 2009) for cross-validation to test the performance of

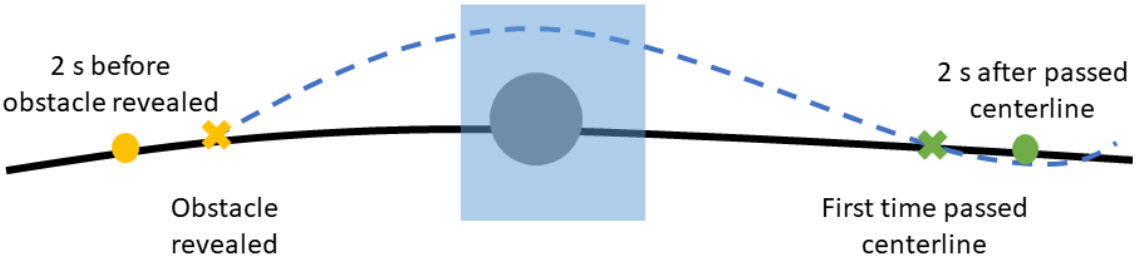


Figure 5.5: An example of obstacle avoidance event. Black curve: track. Grey circle: obstacle. Blue dotted curve: vehicle trajectory. The region between the yellow circle and the green circle indicates the obstacle avoidance event. Blue box: selected sequence of data.

Table 5.1: Performance of the Bayesian inference model and other single models for different obstacle headways

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.683 ± 0.011	0.601 ± 0.007	0.687 ± 0.008	0.609 ± 0.010	0.626 ± 0.009
Precision	0.688 ± 0.011	0.604 ± 0.007	0.691 ± 0.008	0.613 ± 0.011	0.631 ± 0.009
Recall	0.679 ± 0.011	0.598 ± 0.007	0.683 ± 0.008	0.606 ± 0.010	0.621 ± 0.008

our proposed method.

We randomly selected the data of three participants as the testing dataset and the data of the remaining nine participants as the training dataset in each run of the holdout. We ran 50 holdouts to evaluate the performance of our proposed Bayesian inference model (BI) and the four single models, and Table 4.1 shows the results. Our proposed Bayesian inference model achieved a 0.696 ± 0.012 F_1 score for the workload imposed by different obstacle headways.

5.2.3 Discussion

In this experiment, we tried to manipulate human workload by varying driving task difficulty. The results showed that our proposed Bayesian inference model can achieve an F_1 score of 0.683 ± 0.011 when estimating the workload imposed by different obstacle headways. When comparing these results with the workload induced by surveillance task urgency (Table 4.1 in Chapter IV), in which we achieved a 0.823 ± 0.004 F_1 score, our

proposed Bayesian inference model performed worse when distinguishing workload imposed by different obstacle headways. The main reason for this result is the performance decrement for SVMs for pupil size change (F_1 score decreased from 0.772 ± 0.006 to 0.611 ± 0.007) and SVMs for fixation feature (F_1 score decreased from 0.745 ± 0.003 to 0.603 ± 0.011). One potential explanation for this performance decrement is due to the different workload dimensions used by the surveillance task and obstacle avoidance in the driving task. In the surveillance task, participants tried to identify potential threats in the image feeds, which contributed to mental demand. Therefore, when varying surveillance task urgency, the mental demand changed substantially. However, avoiding obstacles with a shorter headway may trigger more temporal demand and physical demand, as the human must steer the vehicle harder and faster. Previous literature showed that pupil dilation and fixation duration are sensitive, diagnostic, and selective to cognitive workload (Heard et al., 2018). Thus, SVMs for pupil size change and SVMs for fixation feature had better performance when workload was imposed by varying surveillance task urgency.

The findings from this experiment should also be viewed in light of the following limitations.

First, we only manipulated the driving task by varying the obstacle headways. However, there are different factors for the driving task that may affect human workload, for example driving speed. We addressed this issue in Experiment 4.

Second, in this experiment, we utilized a combined pace surveillance task. It is unclear the performance of our proposed Bayesian inference model for workload estimation when varying both the surveillance task urgency and the driving task conditions. In Experiment 4, we manipulated both the surveillance task urgency and the driving speed for the driving task.

5.3 Experiment 4: Effects of Driving Speed on Workload Estimation Performance

5.3.1 Method

Participants

A total of 12 students participated in Experiment 4. The 12 participants were on average 25.4 years old (SD = 3.1) and had an average of 5.9 years of driving experience (SD = 3.7). There were 1 female and 11 males in the 12 participants. All participants had normal or corrected-to-normal vision.

Experimental Apparatus and Stimuli

We used an updated version of the dual-task shared control simulation platform, with a high-fidelity visualization system as shown in Figure 5.6. The high-fidelity visualization system depicted a more realistic off-road driving scenario. The goal for the driving task was to drive as close to the reference line as possible. The green dots in Figure 5.6 indicated the reference line. The red gasoline barrel in the middle of the road indicated the obstacle in Figure 5.6. Similar to Experiment 3, we disabled the obstacle avoidance capability in the autonomy. The obstacle headway was 20 s.

Experimental Design

The experiment used a within-subjects design with two independent variables: vehicle speed (low speed 12.5 m/s vs. high speed 22.5 m/s) and surveillance task urgency (low urgency 6.5 s vs. high urgency 1.5 s). Therefore, there were four different cases, as shown in Table 5.2. Each participant drove on four different tracks (each track was 2,700 m long) and with one of the four cases. The four tracks were Track ID 1, 7, 8, and 9, as shown in Figure 2.6 in Section 2.4. There were two obstacles on each track, and each obstacle had a 20 s headway and 1 m diameter. Figure 5.7 shows the four tracks and the locations of

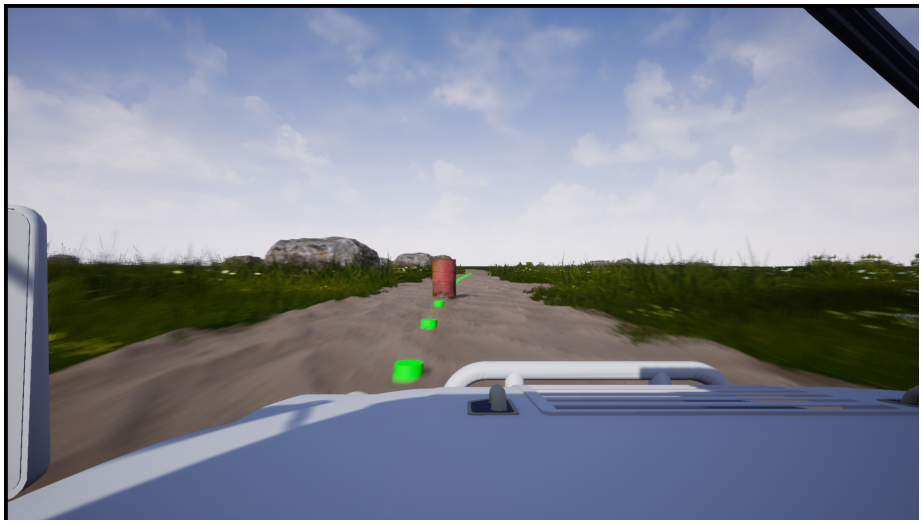


Figure 5.6: High-fidelity visualization system for the driving task.

the obstacles. To eliminate potential order effects, the presentation order of the cases was determined by using 4×4 Latin square.

Table 5.2: Four different cases.

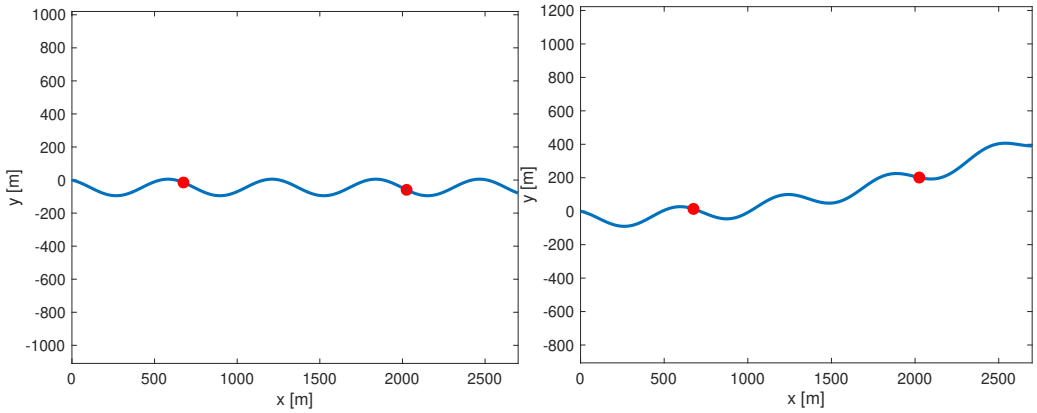
Case ID	Case
1	Low urgency + low speed
2	Low urgency + high speed
3	High urgency + low speed
4	High urgency + high speed

Measures

The participants' gaze points and pupil sizes were recorded by the Tobii Pro Glasses 2 at 30 Hz. After each track, the participants reported their perceived workload using the NASA TLX survey.

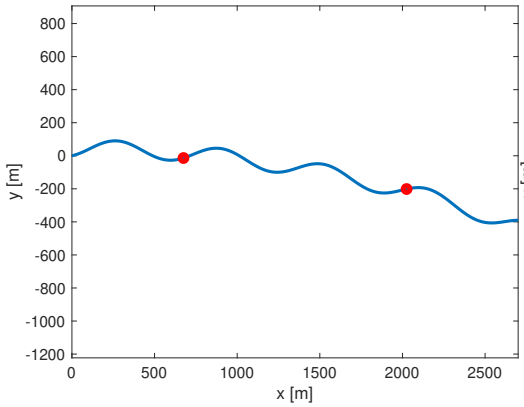
Experimental Procedure

Participants provided signed informed consent and filled out a demographic survey. Then, they received an instruction session. After that, they were assisted to wear the eye tracker (Tobii Pro Glasses 2) and underwent a calibration before the training session.

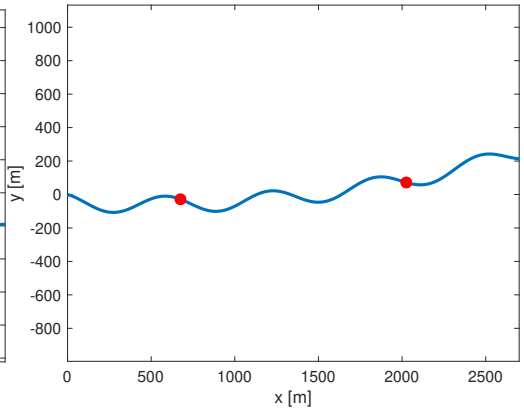


(a) Track 1 (ID: 1)

(b) Track 2 (ID: 7)



(c) Track 3 (ID: 8)



(d) Track 4 (ID: 9)

Figure 5.7: Four tracks in the formal experiment. Red dots indicate the locations of the obstacles.

With the normal room light and without any specific tasks, the experimenter measured each participant's baseline pupil diameter twice, each for about 120 s. During the training session, participants first performed the driving task alone on two different tracks, with low speed and high speed. Then, each participant performed the surveillance task alone, with low urgency first and then high urgency. In the end, each participant performed the dual-task scenario on two tracks, with low speed and high speed. Both tracks contain two portions: the first portion is for the low-urgency surveillance task, and the second portion is for the high-urgency surveillance task. After the training session, participants performed on four different tracks with different cases in the formal experiment. After each track, the participants reported their perceived workload using the NASA TLX survey. A debriefing

survey was completed after the experiment.

5.3.2 Results

Effects of Driving Speed on Human Workload

Two-way repeated measures Analysis of Variance (ANOVAs) were conducted with driving speed and surveillance task urgency as the within-subjects variables. The results indicated significant main effects of driving speed ($F(1, 33) = 10.226$, $p = .003$) and surveillance task urgency ($F(1, 33) = 81.12$, $p < .001$) on self-reported workload (see Figure 5.8). The interaction effect was non-significant.

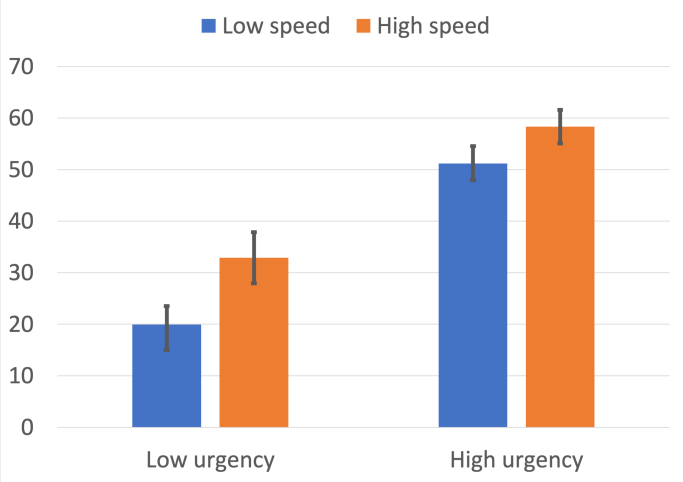


Figure 5.8: Mean and standard error (SE) values of self-reported workload.

We then performed a series of t tests between each pair of different cases. The results revealed significant differences between each pair of different cases.

Data Preparation

We prepared the dataset for the cross-participants evaluation and the within-participants evaluation separately. For the cross-participants evaluation, the participants experienced each case (low/high speed $\times$ low/high urgency) on one track. We randomly selected 30 sequences of data in each track, and each had a 4-second length. We then extracted differ-

ent features, as in Chapter IV. For within-participants evaluation, as each participant only performed each case on one track, we needed to segment each track into four portions equally. In each portion, we randomly selected 20 sequences of data. Therefore, when we performed the within-participants evaluation, we could use data from one portion as testing data and data from the remaining portions as training data.

Given the four different cases, we applied three different ground truth labeling criteria: 1) each case represents one workload level (four levels in total); 2) using driving speed as ground truth labels (two levels) under two conditions (low urgency and high urgency); 3) using surveillance task urgency as ground truth labels (two levels) under two conditions (low speed and high speed).

Four Different Cases as Ground Truth Labels

Table 5.3 and Table 5.5 show the performance of our proposed Bayesian inference model and other single models for workload estimation using four different cases as ground truth labels using cross-participants and within-participants evaluation methods, respectively. Table 5.4 shows the individual performance using cross-participants evaluation, i.e., for each participant, we used the performance of s/he when s/he was selected as testing data to compute the individual performance of her/him.

Note that the random guess for the four-category classification problem is 0.25. The results indicated that for cross-participants evaluation, all methods cannot estimate workload well. Our proposed Bayesian inference model achieved a 0.396 ± 0.006 F_1 score using cross-participants evaluation. However, within-participants evaluation can estimate workload at four different levels with a 0.56 ± 0.03 F_1 score on average, which is much better than the random guess but still not good enough.

Table 5.3: Overall performance for cross-participants evaluation for four different cases as ground truth labels.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.396 ± 0.006	0.339 ± 0.004	0.327 ± 0.004	0.389 ± 0.005	0.386 ± 0.004
Precision	0.381 ± 0.008	0.341 ± 0.005	0.328 ± 0.006	0.388 ± 0.005	0.388 ± 0.004
Recall	0.416 ± 0.005	0.337 ± 0.004	0.328 ± 0.004	0.390 ± 0.005	0.385 ± 0.004

Table 5.4: Individual performance for cross-participants evaluation (F_1 score, precision, recall) for four different cases as ground truth labels.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.42 ± 0.02	0.40 ± 0.02	0.20 ± 0.01	0.41 ± 0.01	0.26 ± 0.01
	0.41 ± 0.03	0.42 ± 0.02	0.17 ± 0.01	0.40 ± 0.01	0.27 ± 0.01
	0.43 ± 0.01	0.38 ± 0.02	0.22 ± 0.00	0.41 ± 0.01	0.26 ± 0.01
P2	0.38 ± 0.02	0.32 ± 0.01	0.35 ± 0.01	0.35 ± 0.00	0.37 ± 0.01
	0.37 ± 0.03	0.32 ± 0.01	0.33 ± 0.01	0.36 ± 0.01	0.37 ± 0.01
	0.40 ± 0.01	0.32 ± 0.01	0.37 ± 0.00	0.34 ± 0.00	0.36 ± 0.01
P3	0.29 ± 0.01	0.24 ± 0.01	0.27 ± 0.01	0.40 ± 0.01	0.29 ± 0.01
	0.30 ± 0.03	0.26 ± 0.01	0.25 ± 0.01	0.47 ± 0.02	0.27 ± 0.01
	0.29 ± 0.00	0.22 ± 0.01	0.29 ± 0.00	0.35 ± 0.01	0.32 ± 0.01
P4	0.35 ± 0.01	0.33 ± 0.01	0.33 ± 0.01	0.39 ± 0.01	0.40 ± 0.01
	0.31 ± 0.01	0.34 ± 0.01	0.29 ± 0.01	0.39 ± 0.01	0.40 ± 0.01
	0.39 ± 0.00	0.33 ± 0.00	0.38 ± 0.01	0.38 ± 0.01	0.41 ± 0.01
P5	0.42 ± 0.02	0.27 ± 0.02	0.23 ± 0.01	0.40 ± 0.01	0.39 ± 0.02
	0.39 ± 0.02	0.26 ± 0.02	0.22 ± 0.02	0.42 ± 0.01	0.38 ± 0.03
	0.45 ± 0.01	0.28 ± 0.02	0.23 ± 0.01	0.37 ± 0.01	0.40 ± 0.02
P6	0.47 ± 0.02	0.23 ± 0.01	0.25 ± 0.04	0.47 ± 0.02	0.40 ± 0.02
	0.49 ± 0.03	0.20 ± 0.01	0.29 ± 0.07	0.44 ± 0.02	0.42 ± 0.03
	0.46 ± 0.01	0.27 ± 0.01	0.26 ± 0.01	0.49 ± 0.01	0.40 ± 0.01
P7	0.46 ± 0.02	0.38 ± 0.01	0.38 ± 0.02	0.49 ± 0.01	0.39 ± 0.01
	0.46 ± 0.03	0.37 ± 0.01	0.38 ± 0.02	0.49 ± 0.01	0.40 ± 0.01
	0.47 ± 0.01	0.39 ± 0.01	0.38 ± 0.01	0.49 ± 0.01	0.38 ± 0.01
P8	0.35 ± 0.01	0.38 ± 0.01	0.30 ± 0.01	0.29 ± 0.01	0.37 ± 0.01
	0.33 ± 0.01	0.38 ± 0.01	0.29 ± 0.01	0.29 ± 0.01	0.36 ± 0.01
	0.38 ± 0.01	0.37 ± 0.01	0.32 ± 0.00	0.29 ± 0.01	0.38 ± 0.01
P9	0.36 ± 0.02	0.36 ± 0.01	0.31 ± 0.00	0.35 ± 0.01	0.38 ± 0.02
	0.34 ± 0.03	0.38 ± 0.01	0.28 ± 0.00	0.38 ± 0.01	0.40 ± 0.03
	0.39 ± 0.01	0.35 ± 0.01	0.35 ± 0.01	0.33 ± 0.01	0.36 ± 0.01
P10	0.44 ± 0.01	0.40 ± 0.00	0.22 ± 0.01	0.39 ± 0.01	0.35 ± 0.01
	0.43 ± 0.02	0.40 ± 0.00	0.20 ± 0.01	0.37 ± 0.01	0.36 ± 0.02
	0.45 ± 0.01	0.40 ± 0.00	0.23 ± 0.01	0.41 ± 0.01	0.35 ± 0.01

Continued on next page

Table 5.4 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P11	0.46 ± 0.02	0.33 ± 0.01	0.35 ± 0.02	0.41 ± 0.01	0.50 ± 0.02
	0.43 ± 0.03	0.34 ± 0.01	0.38 ± 0.02	0.39 ± 0.02	0.50 ± 0.02
	0.50 ± 0.01	0.33 ± 0.01	0.34 ± 0.01	0.44 ± 0.01	0.49 ± 0.02
P12	0.39 ± 0.01	0.34 ± 0.01	0.38 ± 0.01	0.43 ± 0.01	0.42 ± 0.01
	0.36 ± 0.02	0.34 ± 0.01	0.37 ± 0.02	0.43 ± 0.01	0.43 ± 0.01
	0.42 ± 0.01	0.34 ± 0.01	0.39 ± 0.01	0.42 ± 0.01	0.42 ± 0.01
Avg	0.40 ± 0.02	0.33 ± 0.02	0.30 ± 0.02	0.40 ± 0.02	0.38 ± 0.02
	0.39 ± 0.02	0.33 ± 0.02	0.29 ± 0.02	0.40 ± 0.02	0.38 ± 0.02
	0.42 ± 0.02	0.33 ± 0.02	0.31 ± 0.02	0.39 ± 0.02	0.38 ± 0.02

Table 5.5: Within-participants evaluation (F_1 score, precision, recall) for four different cases as ground truth labels.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.53 ± 0.06	0.40 ± 0.02	0.38 ± 0.04	0.43 ± 0.01	0.37 ± 0.02
	0.51 ± 0.07	0.41 ± 0.02	0.38 ± 0.07	0.43 ± 0.01	0.37 ± 0.02
	0.55 ± 0.05	0.40 ± 0.03	0.40 ± 0.01	0.44 ± 0.01	0.37 ± 0.02
P2	0.61 ± 0.02	0.41 ± 0.03	0.52 ± 0.05	0.54 ± 0.04	0.55 ± 0.06
	0.61 ± 0.01	0.42 ± 0.03	0.55 ± 0.04	0.53 ± 0.04	0.55 ± 0.05
	0.62 ± 0.02	0.40 ± 0.04	0.51 ± 0.06	0.54 ± 0.04	0.55 ± 0.06
P3	0.42 ± 0.05	0.32 ± 0.03	0.39 ± 0.05	0.30 ± 0.05	0.47 ± 0.04
	0.45 ± 0.05	0.31 ± 0.03	0.37 ± 0.06	0.30 ± 0.05	0.48 ± 0.04
	0.40 ± 0.05	0.33 ± 0.03	0.42 ± 0.02	0.30 ± 0.04	0.47 ± 0.04
P4	0.54 ± 0.03	0.36 ± 0.04	0.56 ± 0.03	0.42 ± 0.01	0.58 ± 0.03
	0.56 ± 0.03	0.36 ± 0.04	0.57 ± 0.03	0.41 ± 0.01	0.58 ± 0.03
	0.53 ± 0.04	0.36 ± 0.03	0.55 ± 0.04	0.43 ± 0.01	0.58 ± 0.02
P5	0.40 ± 0.04	0.41 ± 0.05	0.42 ± 0.07	0.40 ± 0.02	0.43 ± 0.02
	0.39 ± 0.05	0.41 ± 0.06	0.44 ± 0.08	0.39 ± 0.02	0.44 ± 0.03
	0.42 ± 0.03	0.42 ± 0.04	0.41 ± 0.06	0.40 ± 0.02	0.42 ± 0.02
P6	0.47 ± 0.03	0.32 ± 0.04	0.47 ± 0.04	0.51 ± 0.05	0.45 ± 0.04
	0.47 ± 0.04	0.32 ± 0.04	0.49 ± 0.04	0.51 ± 0.06	0.45 ± 0.04
	0.49 ± 0.03	0.32 ± 0.04	0.46 ± 0.05	0.51 ± 0.05	0.46 ± 0.04
P7	0.70 ± 0.05	0.56 ± 0.02	0.64 ± 0.05	0.44 ± 0.05	0.54 ± 0.04
	0.71 ± 0.05	0.57 ± 0.02	0.65 ± 0.05	0.43 ± 0.05	0.54 ± 0.04
	0.69 ± 0.05	0.55 ± 0.02	0.63 ± 0.05	0.44 ± 0.04	0.54 ± 0.04

Continued on next page

Table 5.5 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P8	0.65 ± 0.04	0.37 ± 0.03	0.58 ± 0.03	0.37 ± 0.03	0.59 ± 0.07
	0.67 ± 0.04	0.37 ± 0.03	0.58 ± 0.03	0.38 ± 0.03	0.60 ± 0.09
	0.64 ± 0.03	0.38 ± 0.03	0.58 ± 0.04	0.37 ± 0.04	0.58 ± 0.06
P9	0.59 ± 0.04	0.52 ± 0.04	0.41 ± 0.05	0.54 ± 0.03	0.54 ± 0.01
	0.60 ± 0.05	0.52 ± 0.05	0.41 ± 0.06	0.55 ± 0.03	0.55 ± 0.02
	0.58 ± 0.04	0.52 ± 0.04	0.41 ± 0.05	0.54 ± 0.02	0.53 ± 0.01
P10	0.56 ± 0.04	0.41 ± 0.03	0.42 ± 0.05	0.41 ± 0.06	0.50 ± 0.02
	0.58 ± 0.05	0.41 ± 0.03	0.42 ± 0.05	0.39 ± 0.06	0.51 ± 0.02
	0.54 ± 0.03	0.40 ± 0.03	0.42 ± 0.05	0.43 ± 0.06	0.49 ± 0.02
P11	0.76 ± 0.02	0.52 ± 0.01	0.54 ± 0.04	0.54 ± 0.02	0.58 ± 0.03
	0.77 ± 0.01	0.52 ± 0.01	0.55 ± 0.04	0.54 ± 0.03	0.60 ± 0.03
	0.75 ± 0.03	0.51 ± 0.01	0.53 ± 0.04	0.53 ± 0.02	0.56 ± 0.04
P12	0.47 ± 0.04	0.34 ± 0.02	0.37 ± 0.05	0.39 ± 0.05	0.51 ± 0.05
	0.48 ± 0.04	0.34 ± 0.03	0.36 ± 0.06	0.38 ± 0.04	0.51 ± 0.05
	0.46 ± 0.03	0.33 ± 0.02	0.38 ± 0.05	0.41 ± 0.05	0.51 ± 0.04
Avg	0.56 ± 0.03	0.41 ± 0.02	0.47 ± 0.03	0.44 ± 0.02	0.51 ± 0.02
	0.57 ± 0.03	0.41 ± 0.02	0.48 ± 0.03	0.44 ± 0.02	0.51 ± 0.02
	0.55 ± 0.03	0.41 ± 0.02	0.47 ± 0.02	0.44 ± 0.02	0.51 ± 0.02

Driving Speed as Ground Truth Labels

When using driving speed as ground truth labels, we must evaluate the performance of workload estimation models conditioning on high urgency and low urgency. For example, when conditioning on low urgency, we are trying to classify the data from low urgency into two categories: low urgency + low speed (Case ID 1) and low urgency + high speed (Case ID 2). When conditioning on high urgency, we are trying to classify the data from high surveillance task urgency into two categories: high urgency + low speed (Case ID 3) and high urgency + high speed (Case ID 4).

When conditioning on low urgency, Table 5.6 and Table 5.8 show the performance of our proposed Bayesian inference model and other single models using cross-participants evaluation and within-participants evaluation, respectively. Table 5.7 shows the individual

performance using cross-participants evaluation.

The results indicated that both models cannot distinguish workload imposed by driving speed under low surveillance task urgency using cross-participants evaluation. Our proposed Bayesian inference model achieved an overall 0.512 ± 0.007 F_1 score and an average individual performance with 0.52 ± 0.02 F_1 score using cross-participants evaluation.

For the within-participants evaluation, the estimation was able to reach a reasonable performance for a few participants (i.e., participants 4, 7, 10, and 11). On average, our proposed Bayesian inference model achieved a 0.62 ± 0.04 F_1 score using the within-participants evaluation under low surveillance task urgency.

Table 5.6: Overall performance for cross-participants evaluation for driving speed as ground truth labels conditioned on low surveillance task urgency.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.512 ± 0.007	0.529 ± 0.004	0.521 ± 0.007	0.535 ± 0.005	0.517 ± 0.007
Precision	0.517 ± 0.009	0.529 ± 0.004	0.524 ± 0.008	0.537 ± 0.006	0.517 ± 0.007
Recall	0.510 ± 0.004	0.528 ± 0.004	0.518 ± 0.006	0.533 ± 0.005	0.517 ± 0.007

Table 5.7: Individual performance for cross-participants evaluation (F_1 score, precision, recall) for driving speed as ground truth labels conditioned on low surveillance task urgency.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.56 ± 0.03	0.49 ± 0.02	0.37 ± 0.04	0.54 ± 0.01	0.51 ± 0.01
	0.59 ± 0.04	0.49 ± 0.02	0.32 ± 0.05	0.54 ± 0.01	0.51 ± 0.02
	0.54 ± 0.02	0.49 ± 0.02	0.46 ± 0.01	0.54 ± 0.01	0.51 ± 0.01
P2	0.53 ± 0.02	0.57 ± 0.01	0.59 ± 0.01	0.57 ± 0.01	0.56 ± 0.02
	0.53 ± 0.03	0.57 ± 0.01	0.61 ± 0.01	0.58 ± 0.01	0.56 ± 0.02
	0.54 ± 0.01	0.57 ± 0.01	0.58 ± 0.01	0.56 ± 0.00	0.55 ± 0.02
P3	0.42 ± 0.03	0.53 ± 0.01	0.45 ± 0.01	0.51 ± 0.02	0.42 ± 0.03
	0.39 ± 0.05	0.53 ± 0.01	0.44 ± 0.01	0.54 ± 0.04	0.40 ± 0.04
	0.48 ± 0.01	0.53 ± 0.01	0.46 ± 0.01	0.50 ± 0.00	0.45 ± 0.02

Continued on next page

Table 5.7 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P4	0.49 ± 0.02	0.50 ± 0.01	0.39 ± 0.03	0.49 ± 0.01	0.52 ± 0.02
	0.48 ± 0.02	0.50 ± 0.01	0.35 ± 0.05	0.49 ± 0.01	0.52 ± 0.02
	0.51 ± 0.01	0.50 ± 0.01	0.51 ± 0.00	0.49 ± 0.00	0.51 ± 0.02
P5	0.54 ± 0.02	0.57 ± 0.01	0.47 ± 0.03	0.59 ± 0.02	0.53 ± 0.03
	0.56 ± 0.04	0.58 ± 0.01	0.47 ± 0.04	0.59 ± 0.02	0.55 ± 0.05
	0.53 ± 0.02	0.57 ± 0.01	0.47 ± 0.02	0.58 ± 0.02	0.51 ± 0.02
P6	0.59 ± 0.03	0.51 ± 0.02	0.45 ± 0.02	0.61 ± 0.02	0.52 ± 0.05
	0.63 ± 0.04	0.51 ± 0.02	0.43 ± 0.03	0.62 ± 0.02	0.53 ± 0.06
	0.56 ± 0.02	0.51 ± 0.02	0.48 ± 0.01	0.60 ± 0.01	0.52 ± 0.04
P7	0.51 ± 0.02	0.56 ± 0.02	0.64 ± 0.02	0.54 ± 0.01	0.48 ± 0.01
	0.52 ± 0.03	0.56 ± 0.02	0.66 ± 0.02	0.55 ± 0.01	0.48 ± 0.02
	0.50 ± 0.01	0.56 ± 0.02	0.62 ± 0.02	0.53 ± 0.01	0.48 ± 0.01
P8	0.42 ± 0.02	0.61 ± 0.02	0.46 ± 0.01	0.45 ± 0.01	0.40 ± 0.02
	0.42 ± 0.02	0.61 ± 0.02	0.46 ± 0.01	0.44 ± 0.01	0.40 ± 0.02
	0.43 ± 0.02	0.60 ± 0.02	0.47 ± 0.00	0.45 ± 0.01	0.40 ± 0.02
P9	0.57 ± 0.04	0.49 ± 0.02	0.54 ± 0.04	0.60 ± 0.01	0.58 ± 0.03
	0.57 ± 0.04	0.49 ± 0.02	0.59 ± 0.07	0.61 ± 0.01	0.59 ± 0.03
	0.56 ± 0.03	0.49 ± 0.02	0.53 ± 0.01	0.60 ± 0.01	0.58 ± 0.03
P10	0.54 ± 0.02	0.52 ± 0.01	0.44 ± 0.01	0.52 ± 0.01	0.57 ± 0.02
	0.56 ± 0.02	0.52 ± 0.01	0.41 ± 0.02	0.52 ± 0.01	0.59 ± 0.02
	0.53 ± 0.01	0.52 ± 0.01	0.48 ± 0.00	0.52 ± 0.01	0.56 ± 0.02
P11	0.54 ± 0.03	0.44 ± 0.01	0.72 ± 0.02	0.61 ± 0.01	0.66 ± 0.04
	0.57 ± 0.04	0.44 ± 0.01	0.80 ± 0.01	0.64 ± 0.02	0.66 ± 0.04
	0.53 ± 0.03	0.44 ± 0.01	0.66 ± 0.02	0.58 ± 0.01	0.65 ± 0.04
P12	0.49 ± 0.02	0.55 ± 0.02	0.43 ± 0.02	0.58 ± 0.01	0.50 ± 0.02
	0.48 ± 0.03	0.55 ± 0.02	0.41 ± 0.02	0.58 ± 0.01	0.50 ± 0.02
	0.51 ± 0.01	0.55 ± 0.02	0.44 ± 0.01	0.58 ± 0.01	0.50 ± 0.02
Avg	0.52 ± 0.02	0.53 ± 0.01	0.50 ± 0.03	0.55 ± 0.02	0.52 ± 0.02
	0.52 ± 0.02	0.53 ± 0.01	0.50 ± 0.04	0.56 ± 0.02	0.52 ± 0.02
	0.52 ± 0.01	0.53 ± 0.01	0.51 ± 0.02	0.54 ± 0.01	0.52 ± 0.02

Table 5.8: Within-participants evaluation (F_1 score, precision, recall) for driving speed as ground truth labels conditioned on low surveillance task urgency.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.60 ± 0.08	0.47 ± 0.06	0.40 ± 0.07	0.49 ± 0.04	0.37 ± 0.02
P1	0.61 ± 0.09	0.46 ± 0.06	0.39 ± 0.08	0.49 ± 0.04	0.34 ± 0.04
	0.58 ± 0.06	0.47 ± 0.05	0.42 ± 0.05	0.49 ± 0.04	0.41 ± 0.01
	0.54 ± 0.03	0.54 ± 0.01	0.63 ± 0.09	0.65 ± 0.07	0.56 ± 0.08
P2	0.54 ± 0.03	0.54 ± 0.01	0.63 ± 0.09	0.65 ± 0.07	0.56 ± 0.08
	0.54 ± 0.03	0.54 ± 0.01	0.63 ± 0.09	0.64 ± 0.07	0.56 ± 0.08
	0.57 ± 0.10	0.46 ± 0.05	0.56 ± 0.04	0.38 ± 0.06	0.61 ± 0.05
P3	0.57 ± 0.12	0.46 ± 0.05	0.57 ± 0.05	0.36 ± 0.06	0.62 ± 0.06
	0.59 ± 0.07	0.46 ± 0.05	0.56 ± 0.04	0.39 ± 0.05	0.60 ± 0.05
	0.72 ± 0.03	0.49 ± 0.02	0.79 ± 0.05	0.48 ± 0.02	0.72 ± 0.03
P4	0.75 ± 0.03	0.49 ± 0.02	0.79 ± 0.05	0.47 ± 0.02	0.74 ± 0.03
	0.70 ± 0.04	0.49 ± 0.02	0.78 ± 0.05	0.48 ± 0.01	0.71 ± 0.04
	0.52 ± 0.07	0.51 ± 0.06	0.47 ± 0.07	0.46 ± 0.11	0.53 ± 0.05
P5	0.53 ± 0.11	0.50 ± 0.07	0.45 ± 0.08	0.45 ± 0.11	0.53 ± 0.05
	0.54 ± 0.02	0.51 ± 0.06	0.50 ± 0.04	0.46 ± 0.10	0.53 ± 0.04
	0.44 ± 0.06	0.59 ± 0.10	0.64 ± 0.02	0.45 ± 0.07	0.53 ± 0.07
P6	0.43 ± 0.07	0.59 ± 0.10	0.65 ± 0.03	0.44 ± 0.08	0.52 ± 0.07
	0.46 ± 0.06	0.59 ± 0.10	0.63 ± 0.01	0.46 ± 0.06	0.53 ± 0.06
	0.75 ± 0.09	0.56 ± 0.03	0.78 ± 0.07	0.49 ± 0.03	0.76 ± 0.09
P7	0.76 ± 0.10	0.56 ± 0.03	0.79 ± 0.07	0.49 ± 0.03	0.76 ± 0.09
	0.74 ± 0.09	0.56 ± 0.03	0.78 ± 0.07	0.49 ± 0.03	0.76 ± 0.09
	0.62 ± 0.05	0.45 ± 0.05	0.70 ± 0.03	0.49 ± 0.10	0.69 ± 0.03
P8	0.65 ± 0.06	0.45 ± 0.05	0.73 ± 0.04	0.49 ± 0.11	0.71 ± 0.01
	0.60 ± 0.04	0.45 ± 0.05	0.68 ± 0.04	0.49 ± 0.10	0.68 ± 0.03
	0.62 ± 0.02	0.58 ± 0.04	0.67 ± 0.05	0.65 ± 0.04	0.60 ± 0.02
P9	0.62 ± 0.02	0.58 ± 0.04	0.68 ± 0.05	0.66 ± 0.04	0.60 ± 0.02
	0.62 ± 0.02	0.57 ± 0.04	0.66 ± 0.05	0.64 ± 0.04	0.60 ± 0.02
	0.71 ± 0.04	0.71 ± 0.04	0.60 ± 0.02	0.56 ± 0.05	0.67 ± 0.03
P10	0.73 ± 0.04	0.71 ± 0.04	0.61 ± 0.02	0.56 ± 0.05	0.68 ± 0.03
	0.70 ± 0.04	0.71 ± 0.04	0.59 ± 0.02	0.56 ± 0.05	0.67 ± 0.03
	0.86 ± 0.03	0.55 ± 0.04	0.85 ± 0.02	0.61 ± 0.03	0.74 ± 0.05
P11	0.86 ± 0.03	0.55 ± 0.05	0.86 ± 0.02	0.62 ± 0.03	0.75 ± 0.05
	0.86 ± 0.03	0.54 ± 0.04	0.84 ± 0.02	0.61 ± 0.03	0.74 ± 0.05
	0.48 ± 0.06	0.51 ± 0.06	0.57 ± 0.03	0.42 ± 0.03	0.45 ± 0.07
P12	0.47 ± 0.09	0.51 ± 0.06	0.58 ± 0.04	0.41 ± 0.03	0.46 ± 0.08
	0.51 ± 0.04	0.51 ± 0.05	0.56 ± 0.03	0.43 ± 0.03	0.44 ± 0.06
	0.62 ± 0.04	0.53 ± 0.02	0.64 ± 0.04	0.51 ± 0.03	0.60 ± 0.04
Avg	0.63 ± 0.04	0.53 ± 0.02	0.64 ± 0.04	0.51 ± 0.03	0.61 ± 0.04

Continued on next page

Table 5.8 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.62 ± 0.03	0.53 ± 0.02	0.63 ± 0.04	0.51 ± 0.02	0.60 ± 0.03

When conditioning on high surveillance task urgency, Table 5.9 and Table 5.11 show the performance of our proposed Bayesian inference model and other single models using cross-participants evaluation and within-participants evaluation, respectively. Table 5.10 shows the individual performance using cross-participants evaluation.

The results indicated that using the cross-participants evaluation cannot distinguish the workload imposed by the driving speed under high surveillance task urgency (a 0.526 ± 0.010 F_1 score for overall performance and a 0.50 ± 0.01 F_1 score for average individual performance). However, using the within-participants evaluation can achieve a 0.70 ± 0.03 F_1 score on average to distinguish human workload imposed by a driving speed under high surveillance task urgency.

Table 5.9: Overall performance for cross-participants evaluation for driving speed as ground truth labels conditioned on high surveillance task urgency.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.526 ± 0.010	0.483 ± 0.006	0.574 ± 0.006	0.492 ± 0.006	0.547 ± 0.005
Precision	0.535 ± 0.015	0.483 ± 0.006	0.580 ± 0.007	0.493 ± 0.006	0.548 ± 0.005
Recall	0.528 ± 0.005	0.484 ± 0.005	0.569 ± 0.006	0.492 ± 0.006	0.546 ± 0.005

Table 5.10: Individual performance for cross-participants evaluation (F_1 score, precision, recall) for driving speed as ground truth labels conditioned on high surveillance task urgency.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.50 ± 0.02	0.47 ± 0.03	0.50 ± 0.01	0.53 ± 0.01	0.48 ± 0.01
P1	0.50 ± 0.02	0.47 ± 0.03	0.50 ± 0.01	0.53 ± 0.01	0.49 ± 0.01
	0.49 ± 0.02	0.47 ± 0.03	0.50 ± 0.01	0.53 ± 0.01	0.48 ± 0.01
	0.60 ± 0.03	0.52 ± 0.02	0.59 ± 0.01	0.48 ± 0.01	0.55 ± 0.02
P2	0.61 ± 0.03	0.52 ± 0.02	0.60 ± 0.01	0.48 ± 0.01	0.56 ± 0.02
	0.58 ± 0.02	0.52 ± 0.01	0.57 ± 0.01	0.48 ± 0.01	0.55 ± 0.02
	0.50 ± 0.02	0.44 ± 0.03	0.45 ± 0.04	0.50 ± 0.02	0.53 ± 0.03
P3	0.51 ± 0.04	0.44 ± 0.03	0.45 ± 0.07	0.50 ± 0.03	0.56 ± 0.05
	0.51 ± 0.01	0.44 ± 0.03	0.51 ± 0.00	0.50 ± 0.02	0.53 ± 0.02
	0.54 ± 0.02	0.49 ± 0.01	0.59 ± 0.01	0.49 ± 0.02	0.56 ± 0.01
P4	0.55 ± 0.03	0.49 ± 0.01	0.60 ± 0.01	0.49 ± 0.02	0.56 ± 0.01
	0.54 ± 0.01	0.49 ± 0.01	0.58 ± 0.01	0.49 ± 0.02	0.56 ± 0.01
	0.41 ± 0.03	0.50 ± 0.02	0.45 ± 0.01	0.33 ± 0.01	0.49 ± 0.03
P5	0.39 ± 0.04	0.50 ± 0.02	0.43 ± 0.02	0.33 ± 0.01	0.50 ± 0.04
	0.46 ± 0.02	0.50 ± 0.02	0.47 ± 0.01	0.33 ± 0.01	0.49 ± 0.02
	0.49 ± 0.05	0.38 ± 0.01	0.52 ± 0.03	0.47 ± 0.03	0.51 ± 0.04
P6	0.51 ± 0.07	0.34 ± 0.01	0.54 ± 0.04	0.47 ± 0.03	0.51 ± 0.04
	0.48 ± 0.03	0.44 ± 0.01	0.51 ± 0.02	0.47 ± 0.02	0.51 ± 0.04
	0.51 ± 0.02	0.44 ± 0.02	0.56 ± 0.01	0.49 ± 0.01	0.50 ± 0.02
P7	0.50 ± 0.03	0.44 ± 0.02	0.57 ± 0.01	0.49 ± 0.01	0.50 ± 0.02
	0.53 ± 0.01	0.44 ± 0.02	0.56 ± 0.01	0.49 ± 0.01	0.50 ± 0.01
	0.50 ± 0.03	0.49 ± 0.01	0.46 ± 0.01	0.53 ± 0.02	0.59 ± 0.01
P8	0.51 ± 0.04	0.49 ± 0.01	0.45 ± 0.01	0.55 ± 0.02	0.59 ± 0.01
	0.53 ± 0.01	0.49 ± 0.01	0.47 ± 0.01	0.52 ± 0.01	0.59 ± 0.01
	0.46 ± 0.03	0.48 ± 0.02	0.58 ± 0.02	0.40 ± 0.03	0.50 ± 0.01
P9	0.44 ± 0.04	0.48 ± 0.02	0.60 ± 0.02	0.40 ± 0.03	0.50 ± 0.01
	0.50 ± 0.01	0.48 ± 0.02	0.57 ± 0.02	0.41 ± 0.03	0.50 ± 0.01
	0.46 ± 0.04	0.52 ± 0.01	0.69 ± 0.02	0.49 ± 0.01	0.56 ± 0.02
P10	0.42 ± 0.05	0.52 ± 0.01	0.71 ± 0.02	0.49 ± 0.02	0.56 ± 0.03
	0.55 ± 0.02	0.52 ± 0.01	0.67 ± 0.02	0.49 ± 0.01	0.55 ± 0.02
	0.47 ± 0.04	0.52 ± 0.02	0.50 ± 0.02	0.52 ± 0.01	0.61 ± 0.02
P11	0.47 ± 0.06	0.52 ± 0.02	0.50 ± 0.02	0.54 ± 0.03	0.64 ± 0.03
	0.52 ± 0.01	0.52 ± 0.02	0.51 ± 0.01	0.51 ± 0.01	0.59 ± 0.02
	0.52 ± 0.04	0.47 ± 0.01	0.75 ± 0.02	0.53 ± 0.01	0.57 ± 0.01
P12	0.51 ± 0.05	0.47 ± 0.01	0.76 ± 0.02	0.53 ± 0.01	0.57 ± 0.01
	0.55 ± 0.02	0.47 ± 0.01	0.74 ± 0.02	0.53 ± 0.01	0.56 ± 0.01
	0.50 ± 0.01	0.48 ± 0.01	0.55 ± 0.03	0.48 ± 0.02	0.54 ± 0.01
Avg					

Continued on next page

Table 5.10 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.49 ± 0.02	0.47 ± 0.01	0.56 ± 0.03	0.48 ± 0.02	0.54 ± 0.01
	0.52 ± 0.01	0.48 ± 0.01	0.55 ± 0.02	0.48 ± 0.02	0.53 ± 0.01

Table 5.11: Within-participants evaluation (F_1 score) for driving speed as ground truth labels conditioned on high surveillance task urgency.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
	0.72 ± 0.05	0.72 ± 0.04	0.54 ± 0.03	0.60 ± 0.03	0.56 ± 0.06
P1	0.73 ± 0.05	0.73 ± 0.04	0.54 ± 0.03	0.60 ± 0.03	0.57 ± 0.07
	0.71 ± 0.05	0.70 ± 0.04	0.53 ± 0.03	0.59 ± 0.03	0.55 ± 0.05
	0.87 ± 0.03	0.53 ± 0.03	0.81 ± 0.06	0.64 ± 0.01	0.85 ± 0.07
P2	0.88 ± 0.02	0.53 ± 0.03	0.84 ± 0.05	0.64 ± 0.01	0.85 ± 0.07
	0.86 ± 0.03	0.53 ± 0.03	0.79 ± 0.08	0.64 ± 0.01	0.84 ± 0.06
	0.75 ± 0.04	0.68 ± 0.02	0.64 ± 0.05	0.68 ± 0.04	0.59 ± 0.08
P3	0.76 ± 0.04	0.68 ± 0.01	0.65 ± 0.05	0.70 ± 0.04	0.59 ± 0.09
	0.74 ± 0.04	0.68 ± 0.02	0.63 ± 0.05	0.66 ± 0.04	0.59 ± 0.07
	0.72 ± 0.07	0.45 ± 0.02	0.73 ± 0.06	0.62 ± 0.05	0.66 ± 0.04
P4	0.73 ± 0.07	0.45 ± 0.02	0.74 ± 0.06	0.62 ± 0.05	0.67 ± 0.05
	0.71 ± 0.07	0.45 ± 0.02	0.72 ± 0.07	0.61 ± 0.05	0.65 ± 0.04
	0.45 ± 0.09	0.57 ± 0.09	0.65 ± 0.04	0.51 ± 0.08	0.51 ± 0.07
P5	0.42 ± 0.10	0.55 ± 0.11	0.68 ± 0.06	0.51 ± 0.09	0.52 ± 0.08
	0.50 ± 0.07	0.60 ± 0.06	0.63 ± 0.03	0.51 ± 0.08	0.51 ± 0.07
	0.68 ± 0.06	0.49 ± 0.03	0.69 ± 0.05	0.68 ± 0.08	0.67 ± 0.07
P6	0.68 ± 0.06	0.49 ± 0.03	0.69 ± 0.05	0.69 ± 0.08	0.68 ± 0.07
	0.68 ± 0.06	0.49 ± 0.03	0.68 ± 0.05	0.68 ± 0.07	0.66 ± 0.06
	0.77 ± 0.04	0.76 ± 0.04	0.65 ± 0.04	0.56 ± 0.05	0.46 ± 0.04
P7	0.78 ± 0.03	0.77 ± 0.04	0.66 ± 0.05	0.56 ± 0.05	0.46 ± 0.04
	0.76 ± 0.04	0.76 ± 0.04	0.65 ± 0.04	0.56 ± 0.05	0.46 ± 0.04
	0.80 ± 0.09	0.40 ± 0.03	0.83 ± 0.08	0.47 ± 0.07	0.78 ± 0.09
P8	0.81 ± 0.09	0.40 ± 0.03	0.84 ± 0.08	0.46 ± 0.07	0.78 ± 0.09
	0.80 ± 0.09	0.41 ± 0.03	0.83 ± 0.08	0.47 ± 0.06	0.78 ± 0.09
	0.76 ± 0.04	0.70 ± 0.05	0.64 ± 0.03	0.69 ± 0.03	0.69 ± 0.03
P9	0.77 ± 0.04	0.70 ± 0.04	0.69 ± 0.05	0.70 ± 0.03	0.70 ± 0.03
	0.76 ± 0.04	0.69 ± 0.05	0.60 ± 0.03	0.69 ± 0.03	0.68 ± 0.04

Continued on next page

Table 5.11 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P10	0.52 ± 0.08	0.48 ± 0.06	0.74 ± 0.08	0.47 ± 0.06	0.70 ± 0.05
	0.49 ± 0.09	0.48 ± 0.06	0.74 ± 0.08	0.48 ± 0.06	0.71 ± 0.05
	0.56 ± 0.05	0.48 ± 0.05	0.73 ± 0.08	0.47 ± 0.06	0.68 ± 0.05
P11	0.67 ± 0.11	0.52 ± 0.02	0.71 ± 0.14	0.61 ± 0.01	0.64 ± 0.06
	0.66 ± 0.14	0.52 ± 0.02	0.70 ± 0.16	0.61 ± 0.01	0.64 ± 0.07
	0.69 ± 0.06	0.52 ± 0.02	0.74 ± 0.10	0.61 ± 0.01	0.63 ± 0.06
P12	0.71 ± 0.06	0.52 ± 0.08	0.68 ± 0.08	0.66 ± 0.06	0.73 ± 0.07
	0.71 ± 0.06	0.52 ± 0.09	0.69 ± 0.08	0.67 ± 0.06	0.74 ± 0.06
	0.70 ± 0.06	0.52 ± 0.07	0.67 ± 0.07	0.66 ± 0.06	0.72 ± 0.07
Avg	0.70 ± 0.03	0.57 ± 0.03	0.69 ± 0.02	0.60 ± 0.02	0.65 ± 0.03
	0.70 ± 0.04	0.57 ± 0.03	0.71 ± 0.02	0.60 ± 0.02	0.66 ± 0.03
	0.71 ± 0.03	0.57 ± 0.03	0.68 ± 0.02	0.60 ± 0.02	0.65 ± 0.03

Surveillance Task Urgency as Ground Truth Labels

Similar to using driving speed as ground truth labels, when using surveillance task urgency as ground truth labels, we evaluated the performance conditioning on different driving speeds (i.e., low speed and high speed). Specifically, under low driving speed, we are trying to classify the data from low driving speed into two categories: low urgency + low speed (Case ID 1) and high urgency + low speed (Case ID 3). Table 5.12 and Table5.14 show the performance of our proposed Bayesian inference model and other single models using cross-participants evaluation and within-participants evaluation, respectively. Table 5.13 shows the individual performance using cross-participants evaluation.

The results indicated that our proposed Bayesian inference model achieved an overall 0.765 ± 0.007 F_1 score and an average individual performance with 0.77 ± 0.02 F_1 score using cross-participants evaluation, as well as a 0.87 ± 0.02 F_1 score on average using within-participants evaluation under the low driving speed condition. In addition, our proposed Bayesian inference model outperforms other single models alone when using

both cross-participants and within-participants evaluation methods.

Table 5.12: Overall performance for cross-participants evaluation for surveillance task urgency as ground truth labels conditioned on low driving speed.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.765 ± 0.007	0.653 ± 0.007	0.663 ± 0.009	0.745 ± 0.006	0.711 ± 0.005
Precision	0.770 ± 0.006	0.658 ± 0.008	0.678 ± 0.009	0.752 ± 0.006	0.717 ± 0.005
Recall	0.759 ± 0.007	0.649 ± 0.008	0.650 ± 0.009	0.739 ± 0.006	0.705 ± 0.006

Table 5.13: Individual performance for cross-participants evaluation (F_1 score, precision, recall) for surveillance task urgency as ground truth labels conditioned on low driving speed.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.78 ± 0.01	0.72 ± 0.01	0.46 ± 0.03	0.83 ± 0.00	0.70 ± 0.04
	0.80 ± 0.01	0.74 ± 0.01	0.45 ± 0.03	0.86 ± 0.00	0.71 ± 0.04
	0.76 ± 0.01	0.70 ± 0.01	0.47 ± 0.03	0.81 ± 0.01	0.69 ± 0.04
P2	0.66 ± 0.01	0.58 ± 0.01	0.64 ± 0.00	0.66 ± 0.01	0.63 ± 0.01
	0.68 ± 0.01	0.58 ± 0.01	0.73 ± 0.01	0.69 ± 0.00	0.63 ± 0.01
	0.64 ± 0.01	0.58 ± 0.01	0.56 ± 0.01	0.64 ± 0.01	0.62 ± 0.01
P3	0.69 ± 0.00	0.57 ± 0.01	0.66 ± 0.01	0.68 ± 0.00	0.66 ± 0.01
	0.78 ± 0.00	0.58 ± 0.01	0.72 ± 0.01	0.78 ± 0.00	0.68 ± 0.02
	0.62 ± 0.00	0.56 ± 0.01	0.61 ± 0.01	0.61 ± 0.01	0.64 ± 0.01
P4	0.81 ± 0.00	0.55 ± 0.01	0.74 ± 0.02	0.79 ± 0.00	0.79 ± 0.01
	0.83 ± 0.00	0.56 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.79 ± 0.01
	0.80 ± 0.00	0.54 ± 0.00	0.70 ± 0.02	0.78 ± 0.00	0.79 ± 0.01
P5	0.71 ± 0.01	0.59 ± 0.02	0.78 ± 0.01	0.75 ± 0.01	0.64 ± 0.04
	0.72 ± 0.01	0.61 ± 0.02	0.78 ± 0.01	0.75 ± 0.01	0.65 ± 0.04
	0.71 ± 0.01	0.57 ± 0.02	0.78 ± 0.01	0.75 ± 0.01	0.64 ± 0.04
P6	0.81 ± 0.01	0.64 ± 0.01	0.53 ± 0.06	0.83 ± 0.01	0.80 ± 0.02
	0.81 ± 0.01	0.73 ± 0.02	0.58 ± 0.10	0.83 ± 0.01	0.81 ± 0.02
	0.81 ± 0.01	0.57 ± 0.01	0.53 ± 0.01	0.83 ± 0.01	0.79 ± 0.02
P7	0.89 ± 0.01	0.69 ± 0.01	0.64 ± 0.01	0.86 ± 0.00	0.78 ± 0.01
	0.89 ± 0.00	0.69 ± 0.01	0.64 ± 0.01	0.86 ± 0.00	0.79 ± 0.01
	0.88 ± 0.01	0.69 ± 0.01	0.64 ± 0.01	0.86 ± 0.00	0.77 ± 0.01
P8	0.71 ± 0.01	0.73 ± 0.01	0.78 ± 0.03	0.66 ± 0.01	0.72 ± 0.01
	0.71 ± 0.01	0.76 ± 0.01	0.80 ± 0.03	0.67 ± 0.01	0.73 ± 0.01
	0.70 ± 0.01	0.70 ± 0.01	0.78 ± 0.02	0.65 ± 0.01	0.71 ± 0.01

Continued on next page

Table 5.13 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P9	0.72 ± 0.01	0.75 ± 0.01	0.67 ± 0.01	0.69 ± 0.01	0.58 ± 0.02
	0.77 ± 0.01	0.77 ± 0.01	0.68 ± 0.01	0.74 ± 0.00	0.60 ± 0.02
	0.68 ± 0.01	0.73 ± 0.01	0.66 ± 0.01	0.64 ± 0.01	0.57 ± 0.02
P10	0.84 ± 0.01	0.74 ± 0.01	0.51 ± 0.02	0.82 ± 0.01	0.69 ± 0.01
	0.84 ± 0.01	0.74 ± 0.01	0.51 ± 0.02	0.83 ± 0.01	0.70 ± 0.01
	0.84 ± 0.01	0.74 ± 0.01	0.50 ± 0.01	0.82 ± 0.00	0.69 ± 0.01
P11	0.83 ± 0.01	0.79 ± 0.01	0.61 ± 0.02	0.73 ± 0.01	0.72 ± 0.01
	0.84 ± 0.01	0.81 ± 0.01	0.62 ± 0.03	0.73 ± 0.01	0.74 ± 0.01
	0.82 ± 0.01	0.78 ± 0.01	0.61 ± 0.01	0.72 ± 0.01	0.71 ± 0.01
P12	0.76 ± 0.01	0.58 ± 0.01	0.82 ± 0.01	0.78 ± 0.01	0.74 ± 0.01
	0.76 ± 0.01	0.59 ± 0.01	0.83 ± 0.01	0.78 ± 0.01	0.75 ± 0.01
	0.76 ± 0.01	0.58 ± 0.01	0.81 ± 0.01	0.77 ± 0.01	0.74 ± 0.01
Avg	0.77 ± 0.02	0.66 ± 0.02	0.65 ± 0.03	0.76 ± 0.02	0.70 ± 0.02
	0.78 ± 0.02	0.68 ± 0.03	0.68 ± 0.03	0.78 ± 0.02	0.71 ± 0.02
	0.75 ± 0.02	0.64 ± 0.02	0.64 ± 0.03	0.74 ± 0.02	0.69 ± 0.02

Table 5.14: Within-participants evaluation (F_1 score, precision, recall) for surveillance task urgency as ground truth labels conditioned on low driving speed.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.91 ± 0.02	0.74 ± 0.06	0.68 ± 0.06	0.89 ± 0.01	0.80 ± 0.03
	0.92 ± 0.02	0.74 ± 0.06	0.73 ± 0.05	0.89 ± 0.01	0.81 ± 0.03
	0.91 ± 0.02	0.73 ± 0.06	0.65 ± 0.07	0.89 ± 0.02	0.79 ± 0.03
P2	0.84 ± 0.03	0.73 ± 0.05	0.74 ± 0.04	0.78 ± 0.02	0.83 ± 0.04
	0.84 ± 0.03	0.73 ± 0.05	0.79 ± 0.02	0.78 ± 0.02	0.83 ± 0.04
	0.83 ± 0.03	0.72 ± 0.05	0.71 ± 0.06	0.77 ± 0.02	0.83 ± 0.04
P3	0.71 ± 0.04	0.75 ± 0.04	0.65 ± 0.12	0.58 ± 0.04	0.55 ± 0.12
	0.75 ± 0.03	0.75 ± 0.04	0.66 ± 0.14	0.58 ± 0.04	0.54 ± 0.13
	0.68 ± 0.06	0.74 ± 0.04	0.66 ± 0.09	0.57 ± 0.04	0.58 ± 0.09
P4	0.93 ± 0.01	0.67 ± 0.02	0.79 ± 0.02	0.88 ± 0.02	0.87 ± 0.04
	0.93 ± 0.01	0.68 ± 0.02	0.80 ± 0.02	0.88 ± 0.01	0.87 ± 0.04
	0.93 ± 0.01	0.67 ± 0.02	0.78 ± 0.02	0.88 ± 0.02	0.87 ± 0.04
P5	0.80 ± 0.05	0.63 ± 0.09	0.72 ± 0.06	0.80 ± 0.03	0.73 ± 0.02
	0.81 ± 0.05	0.63 ± 0.09	0.73 ± 0.07	0.81 ± 0.04	0.74 ± 0.02
	0.79 ± 0.05	0.63 ± 0.09	0.70 ± 0.06	0.79 ± 0.03	0.72 ± 0.02

Continued on next page

Table 5.14 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P6	0.89 ± 0.03	0.62 ± 0.05	0.79 ± 0.02	0.87 ± 0.02	0.86 ± 0.03
	0.90 ± 0.03	0.63 ± 0.05	0.79 ± 0.02	0.87 ± 0.02	0.86 ± 0.03
	0.89 ± 0.03	0.62 ± 0.05	0.78 ± 0.02	0.86 ± 0.02	0.85 ± 0.03
P7	0.93 ± 0.03	0.79 ± 0.04	0.93 ± 0.02	0.89 ± 0.02	0.90 ± 0.04
	0.93 ± 0.03	0.80 ± 0.04	0.94 ± 0.02	0.90 ± 0.02	0.90 ± 0.04
	0.93 ± 0.03	0.78 ± 0.04	0.92 ± 0.02	0.89 ± 0.02	0.89 ± 0.04
P8	0.92 ± 0.01	0.88 ± 0.01	0.88 ± 0.03	0.70 ± 0.02	0.88 ± 0.04
	0.93 ± 0.01	0.89 ± 0.01	0.88 ± 0.03	0.71 ± 0.03	0.88 ± 0.04
	0.92 ± 0.01	0.87 ± 0.02	0.88 ± 0.03	0.69 ± 0.03	0.88 ± 0.04
P9	0.94 ± 0.01	0.88 ± 0.01	0.67 ± 0.06	0.78 ± 0.01	0.83 ± 0.03
	0.94 ± 0.01	0.88 ± 0.01	0.69 ± 0.07	0.79 ± 0.01	0.84 ± 0.03
	0.94 ± 0.01	0.88 ± 0.01	0.64 ± 0.06	0.77 ± 0.01	0.82 ± 0.03
P10	0.85 ± 0.02	0.67 ± 0.04	0.57 ± 0.06	0.86 ± 0.01	0.61 ± 0.03
	0.85 ± 0.02	0.67 ± 0.04	0.59 ± 0.06	0.87 ± 0.01	0.62 ± 0.03
	0.84 ± 0.02	0.67 ± 0.04	0.56 ± 0.05	0.85 ± 0.02	0.61 ± 0.03
P11	0.85 ± 0.04	0.80 ± 0.06	0.73 ± 0.03	0.83 ± 0.03	0.78 ± 0.02
	0.86 ± 0.04	0.81 ± 0.06	0.74 ± 0.03	0.85 ± 0.02	0.79 ± 0.02
	0.84 ± 0.05	0.80 ± 0.07	0.72 ± 0.03	0.81 ± 0.03	0.77 ± 0.03
P12	0.92 ± 0.04	0.62 ± 0.02	0.83 ± 0.04	0.69 ± 0.05	0.88 ± 0.03
	0.92 ± 0.04	0.63 ± 0.02	0.85 ± 0.03	0.71 ± 0.06	0.89 ± 0.03
	0.91 ± 0.04	0.62 ± 0.02	0.82 ± 0.04	0.68 ± 0.04	0.88 ± 0.03
Avg	0.87 ± 0.02	0.73 ± 0.03	0.75 ± 0.03	0.80 ± 0.03	0.79 ± 0.03
	0.88 ± 0.02	0.74 ± 0.03	0.77 ± 0.03	0.80 ± 0.03	0.80 ± 0.03
	0.87 ± 0.02	0.73 ± 0.03	0.74 ± 0.03	0.79 ± 0.03	0.79 ± 0.03

When conditioning on high driving speed, we are trying to classify the data from high driving speed into two categories: low urgency + high speed (Case ID 2) and high urgency + high speed (Case ID 4). Table 5.15 and Table 5.17 show the performance of our proposed Bayesian inference model and other single models using cross-participants evaluation and within-participants evaluation, respectively. Table 5.16 shows the individual performance using cross-participants evaluation.

The results indicated that our proposed Bayesian inference model achieved an overall 0.808 ± 0.008 F_1 score and an average individual performance with a 0.82 ± 0.02 F_1 score

using cross-participants evaluation as well as a 0.86 ± 0.02 F_1 score on average using within-participants evaluation under the high driving speed condition. In addition, our proposed Bayesian inference model outperforms other single models alone when using both cross-participants and within-participants evaluation methods.

Table 5.15: Overall cross-participants evaluation for surveillance task urgency as ground truth labels conditioned on high driving speed.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
F_1 score	0.808 ± 0.008	0.679 ± 0.007	0.593 ± 0.008	0.777 ± 0.009	0.720 ± 0.007
Precision	0.815 ± 0.008	0.682 ± 0.007	0.601 ± 0.009	0.783 ± 0.009	0.725 ± 0.007
Recall	0.801 ± 0.008	0.676 ± 0.007	0.586 ± 0.008	0.772 ± 0.009	0.714 ± 0.007

Table 5.16: Individual performance for cross-participants evaluation (F_1 score, precision, recall) for surveillance task urgency as ground truth labels conditioned on high driving speed.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.81 ± 0.02	0.68 ± 0.01	0.39 ± 0.01	0.88 ± 0.01	0.52 ± 0.02
	0.81 ± 0.02	0.71 ± 0.01	0.36 ± 0.01	0.89 ± 0.01	0.53 ± 0.03
	0.81 ± 0.02	0.66 ± 0.01	0.42 ± 0.01	0.88 ± 0.01	0.52 ± 0.02
P2	0.71 ± 0.01	0.64 ± 0.01	0.71 ± 0.01	0.70 ± 0.01	0.67 ± 0.01
	0.73 ± 0.01	0.64 ± 0.01	0.76 ± 0.01	0.73 ± 0.01	0.67 ± 0.01
	0.69 ± 0.01	0.64 ± 0.01	0.67 ± 0.01	0.68 ± 0.02	0.67 ± 0.01
P3	0.68 ± 0.00	0.48 ± 0.01	0.49 ± 0.03	0.68 ± 0.00	0.51 ± 0.01
	0.78 ± 0.00	0.48 ± 0.01	0.52 ± 0.05	0.78 ± 0.00	0.52 ± 0.01
	0.60 ± 0.01	0.48 ± 0.01	0.47 ± 0.02	0.61 ± 0.00	0.51 ± 0.01
P4	0.79 ± 0.01	0.70 ± 0.01	0.71 ± 0.02	0.73 ± 0.01	0.78 ± 0.00
	0.79 ± 0.01	0.71 ± 0.01	0.74 ± 0.01	0.74 ± 0.01	0.78 ± 0.00
	0.79 ± 0.01	0.70 ± 0.01	0.68 ± 0.02	0.73 ± 0.01	0.77 ± 0.01
P5	0.90 ± 0.00	0.53 ± 0.02	0.55 ± 0.02	0.89 ± 0.01	0.80 ± 0.02
	0.90 ± 0.01	0.55 ± 0.02	0.55 ± 0.03	0.90 ± 0.01	0.80 ± 0.02
	0.90 ± 0.00	0.52 ± 0.01	0.54 ± 0.02	0.88 ± 0.02	0.79 ± 0.02
P6	0.87 ± 0.00	0.52 ± 0.01	0.45 ± 0.02	0.86 ± 0.00	0.71 ± 0.01
	0.87 ± 0.00	0.53 ± 0.01	0.45 ± 0.02	0.86 ± 0.00	0.73 ± 0.02
	0.86 ± 0.00	0.52 ± 0.00	0.45 ± 0.02	0.86 ± 0.00	0.69 ± 0.02

Continued on next page

Table 5.16 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P7	0.93 ± 0.00	0.80 ± 0.01	0.74 ± 0.01	0.91 ± 0.00	0.83 ± 0.01
	0.93 ± 0.00	0.81 ± 0.01	0.75 ± 0.01	0.91 ± 0.00	0.84 ± 0.01
	0.93 ± 0.00	0.79 ± 0.01	0.72 ± 0.01	0.90 ± 0.00	0.81 ± 0.01
P8	0.72 ± 0.02	0.74 ± 0.01	0.68 ± 0.01	0.63 ± 0.02	0.74 ± 0.01
	0.72 ± 0.02	0.75 ± 0.01	0.69 ± 0.01	0.63 ± 0.02	0.75 ± 0.01
	0.71 ± 0.02	0.74 ± 0.01	0.67 ± 0.01	0.62 ± 0.02	0.72 ± 0.01
P9	0.82 ± 0.00	0.70 ± 0.01	0.65 ± 0.01	0.84 ± 0.00	0.73 ± 0.01
	0.85 ± 0.00	0.71 ± 0.01	0.68 ± 0.01	0.86 ± 0.00	0.77 ± 0.01
	0.79 ± 0.00	0.68 ± 0.01	0.63 ± 0.02	0.81 ± 0.00	0.69 ± 0.02
P10	0.89 ± 0.01	0.72 ± 0.01	0.60 ± 0.02	0.89 ± 0.00	0.66 ± 0.01
	0.89 ± 0.01	0.72 ± 0.01	0.61 ± 0.02	0.89 ± 0.00	0.67 ± 0.02
	0.89 ± 0.01	0.72 ± 0.01	0.59 ± 0.02	0.89 ± 0.00	0.65 ± 0.01
P11	0.88 ± 0.01	0.72 ± 0.00	0.53 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
	0.89 ± 0.01	0.76 ± 0.01	0.56 ± 0.02	0.83 ± 0.01	0.81 ± 0.01
	0.87 ± 0.01	0.68 ± 0.01	0.51 ± 0.00	0.78 ± 0.02	0.80 ± 0.01
P12	0.84 ± 0.00	0.71 ± 0.01	0.45 ± 0.03	0.80 ± 0.01	0.80 ± 0.01
	0.85 ± 0.00	0.71 ± 0.01	0.46 ± 0.03	0.81 ± 0.01	0.81 ± 0.01
	0.83 ± 0.00	0.70 ± 0.01	0.44 ± 0.02	0.80 ± 0.01	0.79 ± 0.01
Avg	0.82 ± 0.02	0.66 ± 0.03	0.58 ± 0.03	0.80 ± 0.03	0.71 ± 0.03
	0.83 ± 0.02	0.67 ± 0.03	0.59 ± 0.04	0.82 ± 0.03	0.72 ± 0.03
	0.81 ± 0.03	0.65 ± 0.03	0.57 ± 0.03	0.79 ± 0.03	0.70 ± 0.03

Table 5.17: Within-participants evaluation (F_1 score, precision, recall) for surveillance task urgency as ground truth labels conditioned on high driving speed.

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P1	0.88 ± 0.01	0.59 ± 0.04	0.83 ± 0.05	0.83 ± 0.04	0.82 ± 0.03
	0.89 ± 0.01	0.59 ± 0.04	0.84 ± 0.05	0.84 ± 0.04	0.83 ± 0.03
	0.87 ± 0.02	0.59 ± 0.04	0.82 ± 0.06	0.83 ± 0.04	0.81 ± 0.03
P2	0.92 ± 0.03	0.77 ± 0.02	0.71 ± 0.07	0.87 ± 0.04	0.87 ± 0.03
	0.93 ± 0.03	0.78 ± 0.02	0.72 ± 0.08	0.88 ± 0.04	0.88 ± 0.03
	0.92 ± 0.03	0.76 ± 0.02	0.71 ± 0.07	0.86 ± 0.05	0.87 ± 0.03
P3	0.76 ± 0.04	0.62 ± 0.02	0.81 ± 0.03	0.64 ± 0.06	0.78 ± 0.06
	0.77 ± 0.04	0.62 ± 0.02	0.82 ± 0.03	0.65 ± 0.06	0.79 ± 0.06
	0.76 ± 0.04	0.61 ± 0.02	0.80 ± 0.04	0.63 ± 0.05	0.77 ± 0.05

Continued on next page

Table 5.17 – continued from previous page

	Bayesian inference	SVMs pupil size change	HMM gaze trajectory	SVMs fixation feature	GMMs fixation trajectory
P4	0.82 ± 0.04	0.62 ± 0.05	0.72 ± 0.04	0.82 ± 0.04	0.74 ± 0.05
	0.84 ± 0.03	0.62 ± 0.05	0.74 ± 0.04	0.84 ± 0.04	0.75 ± 0.05
	0.80 ± 0.04	0.61 ± 0.05	0.71 ± 0.04	0.81 ± 0.05	0.74 ± 0.05
P5	0.90 ± 0.02	0.71 ± 0.04	0.74 ± 0.04	0.91 ± 0.01	0.74 ± 0.06
	0.91 ± 0.02	0.71 ± 0.04	0.76 ± 0.05	0.92 ± 0.01	0.75 ± 0.06
	0.90 ± 0.02	0.71 ± 0.04	0.72 ± 0.04	0.91 ± 0.01	0.73 ± 0.06
P6	0.85 ± 0.04	0.65 ± 0.07	0.70 ± 0.07	0.86 ± 0.05	0.68 ± 0.03
	0.87 ± 0.03	0.66 ± 0.07	0.71 ± 0.07	0.86 ± 0.04	0.68 ± 0.03
	0.84 ± 0.05	0.64 ± 0.07	0.69 ± 0.07	0.85 ± 0.05	0.68 ± 0.03
P7	0.90 ± 0.04	0.85 ± 0.02	0.82 ± 0.03	0.89 ± 0.04	0.83 ± 0.04
	0.90 ± 0.04	0.86 ± 0.02	0.82 ± 0.03	0.89 ± 0.04	0.83 ± 0.05
	0.89 ± 0.04	0.84 ± 0.03	0.81 ± 0.03	0.89 ± 0.04	0.82 ± 0.04
P8	0.85 ± 0.04	0.82 ± 0.02	0.58 ± 0.05	0.71 ± 0.05	0.70 ± 0.04
	0.85 ± 0.03	0.83 ± 0.02	0.58 ± 0.05	0.71 ± 0.05	0.74 ± 0.04
	0.84 ± 0.04	0.81 ± 0.02	0.57 ± 0.05	0.71 ± 0.05	0.68 ± 0.05
P9	0.84 ± 0.05	0.76 ± 0.05	0.57 ± 0.10	0.83 ± 0.02	0.76 ± 0.08
	0.84 ± 0.05	0.76 ± 0.05	0.59 ± 0.13	0.85 ± 0.02	0.76 ± 0.08
	0.84 ± 0.05	0.75 ± 0.05	0.58 ± 0.07	0.81 ± 0.03	0.76 ± 0.08
P10	0.91 ± 0.04	0.73 ± 0.03	0.69 ± 0.05	0.88 ± 0.04	0.82 ± 0.04
	0.91 ± 0.04	0.73 ± 0.03	0.71 ± 0.06	0.88 ± 0.04	0.82 ± 0.04
	0.91 ± 0.04	0.72 ± 0.03	0.67 ± 0.05	0.88 ± 0.04	0.81 ± 0.04
P11	0.93 ± 0.01	0.92 ± 0.01	0.61 ± 0.03	0.88 ± 0.04	0.86 ± 0.05
	0.93 ± 0.01	0.92 ± 0.01	0.62 ± 0.03	0.90 ± 0.03	0.87 ± 0.05
	0.93 ± 0.01	0.91 ± 0.01	0.59 ± 0.02	0.86 ± 0.05	0.84 ± 0.06
P12	0.80 ± 0.04	0.69 ± 0.05	0.56 ± 0.03	0.80 ± 0.05	0.84 ± 0.04
	0.81 ± 0.04	0.70 ± 0.05	0.57 ± 0.03	0.81 ± 0.04	0.86 ± 0.04
	0.79 ± 0.04	0.69 ± 0.05	0.56 ± 0.03	0.79 ± 0.05	0.82 ± 0.04
Avg	0.86 ± 0.02	0.73 ± 0.03	0.70 ± 0.03	0.83 ± 0.02	0.79 ± 0.02
	0.87 ± 0.01	0.73 ± 0.03	0.71 ± 0.03	0.84 ± 0.02	0.80 ± 0.02
	0.86 ± 0.02	0.72 ± 0.03	0.69 ± 0.03	0.82 ± 0.02	0.78 ± 0.02

5.3.3 Discussion

In Experiment 4, we investigated the effects of driving speed and surveillance task urgency on workload estimation performance. Table 5.18 summarizes the F_1 scores for our proposed Bayesian inference model using different ground truth labels when using

Table 5.18: Summary of F_1 score for Bayesian inference model.

Ground Truth Label	Condition	Cross-participants Evaluation	Within-participants Evaluation
Four different cases (4 levels)	N/A	0.396 ± 0.006	0.56 ± 0.03
Driving speed (2 levels)	Low urgency	0.512 ± 0.007	0.62 ± 0.04
	High urgency	0.526 ± 0.010	0.70 ± 0.03
Surveillance task urgency (2 levels)	Low speed	0.765 ± 0.007	0.87 ± 0.02
	High speed	0.808 ± 0.008	0.86 ± 0.02

cross-participants and within-participants evaluation methods. On average, the within-participant evaluation method outperforms the cross-participants evaluation method for every conditions. Cross-participants evaluation cannot distinguish workload imposed by driving speed. However, within-participants evaluation can distinguish the human workload imposed by different driving speeds under high surveillance task urgency but not low surveillance task urgency. One potential reason for this is that a human operator has more resources to deal with the additional workload imposed by high driving speed under low surveillance task urgency than under high surveillance task urgency. Therefore, under low surveillance task urgency, the effects of different driving speeds on physiological measurements are too small to be modeled by the machine learning models. However, human operators can still feel the differences in driving speed and report a higher workload for high speed.

When using surveillance task urgency as the ground truth label, both cross-participants evaluation and within-participants evaluation can achieve F_1 scores greater than 0.8, except for cross-participants evaluation under low driving speed, which achieved 0.765 ± 0.007. This is consistent with Chapter IV, even though the visualization systems are totally different and obstacle avoidance is introduced in the driving task.

Both the cross-participants and within-participants evaluation methods did not perform well in classifying workload into four different levels. Note that the F_1 score should be

around 0.25 for random guess to classify workload into four different levels. We speculate that the imperfect performance for four different levels is due to the challenges to distinguish workload introduced by driving speed.

5.4 Conclusion

In this chapter, we investigated whether our proposed Bayesian inference model for workload estimation can generalize to different factors to impose human workload and to different scenarios. In Experiment 3, we introduced obstacle avoidance to the driving task. The results indicated that our proposed Bayesian inference model can distinguish the workload imposed by different obstacle headways. In Experiment 4, we updated the visualization system to a high-fidelity visualization system. Our proposed Bayesian inference model could still distinguish the workload imposed by surveillance task urgency even with the high-fidelity visualization system. However, estimating the workload imposed by driving speed is a challenging problem. Our proposed Bayesian inference model can still distinguish the different workload levels introduced by different driving speeds under high surveillance task urgency.

CHAPTER VI

Conclusion

6.1 Summary

Automated vehicles (AVs) have the potential to reduce driving-related injuries and deaths. However, autonomous driving technology is currently limited in its scope and reliability, giving rise to the semi-autonomous driving model, in which the autonomy and the human shared control of the vehicle. Existing studies have developed haptic shared control schemes for semi-autonomous vehicles that adapt to different factors. Workload, as an important human factor for human-automation interaction, has not yet been considered for adaptation in the shared control.

Different physiological measurements have been used to estimate human workload including brain signals, galvanic skin responses, heart rate-related measures, and eye-related measures. However, existing studies primarily adopted either a single-model-single-feature or a single-model-all-feature approach. It is unclear how to leverage the different machine learning models that work best for different features to improve overall performance.

To address real-time workload estimation problem, as well as its application in the haptic shared control of ground vehicles, this dissertation research was focused on using non-intrusive physiological measures, in particular, eye-related measures. The aims of this

dissertation were to:

(1) Investigate whether and to what extent haptic shared control performance can be improved by incorporating drivers' workload.

(2) Explore different eye-related features for workload estimation and their corresponding machine learning models. Propose a computational model to leverage different machine learning models for these features to improve the workload estimation performance.

(3) Investigate the generalizability of the proposed method for workload estimation in different scenarios and different factors that impose human workload.

For Aim 1, together with our collaborators from the Department of Mechanical Engineering - Yifan Weng, Dr. Tulga Ersal, and Prof. Jeffrey Stein, we developed a teleoperated dual-task shared control platform, where the human operator shares control of a ground vehicle with autonomy while performing a surveillance task alone. We conducted two pilot studies and two human subject experiments with 10, 6, 12, and 12 participants, respectively. In Pilot Study 1, we selected the tracks with similar difficulties to be used in the platform. In Pilot Study 2, we determined the time limits for the detection period in the surveillance task to impose human workload. In Experiment 1, we collected participants' pupil sizes and gaze points when they performed the dual tasks under different surveillance task urgencies. We used the Hidden Markov Model to model human gaze trajectory for workload estimation, which achieved a 0.66 F_1 score. In Experiment 2, we investigated the effects of the proposed workload-adaptive haptic shared control scheme on human performance in the dual-task scenarios. The results indicated that our proposed real-time workload-adaptive shared control scheme can reduce human workload, path tracking errors, and control effort while increasing human trust in the system without sacrificing surveillance task performance.

For Aim 2, we explored pupil size change, gaze trajectory, and fixation features for workload estimation. In addition, we proposed a new feature: fixation trajectory, which contains spatial information for the fixations. We proposed a Bayesian inference model that can leverage the different machine learning models for different features: SVMs for pupil size change, the HMM for gaze trajectory, SVMs for fixation feature, and GMMs for fixation trajectory. We used both cross-participants and within-participants evaluation methods to evaluate the performance of our proposed Bayesian inference model. The training data and testing data in the cross-participants evaluation method are from different participants. However, the training data and testing data are from the same participants but different trials in the within-participants evaluation method. On the data set collected in Experiment 1 and additional 12 participants, our proposed Bayesian inference model achieved 0.82 and 0.85 F_1 scores using cross-participants and within-participants evaluation methods, respectively.

For Aim 3, we conducted another two human subject experiments to investigate the generalizability of our proposed Bayesian inference model. In Experiment 3, we introduced obstacle avoidance to the driving task and varied human workload by manipulating the obstacle headway. Our proposed Bayesian inference model was able to distinguish the workload induced by different obstacle headways with a 0.68 F_1 score using cross-participants evaluation method. In Experiment 4, we used a higher-fidelity simulator and varied the human workload by both surveillance task urgency and driving speed. The results indicated that our proposed Bayesian inference model were still able to distinguish the workload induced by the surveillance task urgency with 0.77 (cross-participants) and 0.87 (within-participants) F_1 scores under low driving speed, as well as 0.81 (cross-participants) and 0.86 F_1 (within-participants) scores under high driving speed. However, estimating the human workload induced by driving speed was less promising. Our pro-

posed Bayesian inference model obtained a 0.51 F_1 scores under low surveillance task urgency and a 0.53 F_1 score under high surveillance task urgency when using cross-participants evaluation. In addition, when using the within-participants evaluation, our proposed Bayesian inference model achieved a 0.62 F_1 score under low surveillance task urgency and a 0.70 F_1 score under high surveillance task urgency.

6.2 Intellectual Merit and Broad Impact

The proposed research will contribute to the knowledge in real-time human workload estimation and its application in the haptic shared control of a ground vehicle.

First, we showed that adapting to human operators' workload led to better driving performance, lower workload, higher trust in the automation, and smaller control effort from the human. Our research is beneficial for the automotive industry and can help it build adaptive Guardian systems for semi-autonomous vehicles. For example, Toyota's Guardian systems can safely blend the vehicle control between the driver and the autonomy to take best advantage of their individual skills (Toyota, 2020). With an in-vehicle camera, the driver's pupil diameters and gaze directions can be obtained to estimate driver workload. Therefore, Guardian system can use estimated workload to determine the control authority from the driver.

Second, our proposed Bayesian inference model can leverage the different machine learning models that work best for different features. Although this dissertation research was focused on eye-related measurements, the model could be generalized to incorporate other physiological measurements.

Third, we showed that our proposed Bayesian inference can be generalized to different scenarios with different factors to manipulate human workload. The results indicated that workload induced by certain factors can be easily distinguished, but others not. Our re-

search suggested that we should be careful about the factors that induce human workload for applying real-time workload estimation to other human-machine systems.

Finally, although our Bayesian inference model was developed using the teleoperated dual-task shared control platform, it can be applied to other human-machine systems. For example, it can be applied to control centers for monitoring multiple unmanned autonomous vehicles or automated vehicles for package delivery.

6.3 Limitations and Future Work

Through a series of human subject experiments, this dissertation can enhance people's understanding of real-time workload estimation and its benefits for haptic shared control of ground vehicles. However, it is subjective to the following limitations, and a few future research directions are suggested to address them.

First, in this dissertation, we focused on the estimating human overall workload using non-intrusive physiological measurements. However, workload has different dimensions (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration according to the NASA TLX scale). Future research can focus on estimating human workload in different dimensions (i.e., using electromyography (EMG) to measure physical demand and Galvanic Skin Response (GSR) to measure temporal demand). Then, our proposed Bayesian inference model could be helpful in combining them.

Second, we did not factor in the workload dynamics in workload estimation. In this dissertation, we treated the workload estimation problem as a classification problem and segmented the time series physiological signals into sequences of data (i.e., each sequence of data lasts for 4 s time window). Therefore, we treated each sequence of data as one data point and extract feature vectors for classifiers. Future work can take the workload dynamics into account to improve the workload estimation performance. As our proposed

Bayesian inference model is based on the graphical model, it can be naturally extended to graphical model with time series data by connecting the hidden state of workload, with the workload dynamics modeled as the transition between the hidden states.

Third, we manipulated surveillance task urgency, obstacle headway, and driving speed to impose different levels of workload. However, there are other factors that can influence human workload (i.e., road curvature, surrounding traffic, weather, etc.). Future work can investigate the generalizability of our proposed Bayesian inference model for the workload imposed by these factors.

Fourth, we measured people's pupil sizes and gaze points by using an eye tracker under normal room light conditions in all our experiments. However, for outdoor driving under natural light conditions, collecting reliable pupil sizes is challenging. For instance, when it is too bright or too dark, pupil diameter can be extremely small or big, which may lead to the floor effect or ceiling effect. Therefore, the performance of our proposed Bayesian inference model in outdoor conditions is unclear.

Fifth, we showed that the performance of the within-participants evaluation was mostly better than cross-participants evaluation. However, within-participants evaluation requires a larger set of training data and trials to obtain better performance. In the future, we can use semi-supervised techniques to build a personalized model for workload estimation (i.e., build a baseline model first with some participants and fine tune the model parameters for each individual). In this way, we may reduce the number of trials for each new participant while maintaining a good within-participants performance.

Sixth, as we only manipulated the surveillance task urgency in the dual-task mission, the participants could not experience high workload and high eyes on road in Experiment 2 when we examined the effects of adapting to workload on the performance of haptic shared control scheme. In the future, we will adjust the experiment design to manipulate the

driving task difficulty to trigger the high workload and high eyes on road experience for the human.

Finally, the population of participants in our experiments were young adults. Existing studies have shown that different age groups have different patterns for certain physiological signals under different workload conditions. Future work should investigate the generalizability of the findings using other populations.

BIBLIOGRAPHY

Aasman, J., Mulder, G., and Mulder, L. J. (1987). Operator effort and the measurement of heart-rate variability. *Human factors*, 29(2):161–170. 1.3.4

Ahern, S. and Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, 205(4412):1289–1292. 1.3.4, 1.2

Ahlstrom, U. and Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International journal of industrial ergonomics*, 36(7):623–636. 1.3.4, 1.2

Alm, H. and Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis & Prevention*, 27(5):707–715. 1.3.3

Anderson, S. J., Peters, S. C., Pilutti, T. E., and Iagnemma, K. (2011). Design and development of an optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios. In Pradalier, C., Siegwart, R., and Hirzinger, G., editors, *Robotics Research*, pages 39–54, Berlin, Heidelberg. Springer Berlin Heidelberg. 1.1

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47. 1.3.4

Backs, R. W., Lenneman, J. K., Wetzel, J. M., and Green, P. (2003). Cardiac measures of driver workload during simulated driving with and without visual occlusion. *Human Factors*, 45(4):525–538. 1.3.4

Backs, R. W. and Walrath, L. C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied ergonomics*, 23(4):243–254. 1.2, 1.3.4

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., and Montanari, R. (2011). Driver workload and eye blink duration. *Transportation research part F: traffic psychology and behaviour*, 14(3):199–208. 1.2, 1.3.4

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental

workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75. 1.3.4

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. 2.3

Briggs, G. F., Hole, G. J., and Land, M. F. (2011). Emotionally involving telephone conversations lead to driver error and visual tunnelling. *Transportation research part F: traffic psychology and behaviour*, 14(4):313–323. 1.2

Calinon, S. (2016). A tutorial on task-parameterized movement learning and retrieval. *Intelligent Service Robotics*, 9(1):1–29. 3.2.1

Calinon, S. and Billard, A. (2005). Recognition and reproduction of gestures using a probabilistic framework combining pca, ica and hmm. In *Proceedings of the 22nd international conference on Machine learning*, pages 105–112. ACM. 3.2.1, 3.2.3, 4.2.4

Carmody, M. A. (1994). Current issues in the measurement of military aircrew performance: A consideration of the relationship between available metrics and operational concerns. Technical report, NAVAL AIR WARFARE CENTER AIRCRAFT DIV WARMINSTER PA AIR VEHICLE AND CREW 1.2, 1.3.4

Chang, C.-C., Boyle, L. N., Lee, J. D., and Jenness, J. (2017). Using tactile detection response tasks to assess in-vehicle voice control interactions. *Transportation research part F: traffic psychology and behaviour*, 51:38–46. 1.3.2

Chen, S. and Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer methods and programs in biomedicine*, 110(2):111–124. 1.3, 1.3.4

Chen, S., Wang, H., and Meng, Q. (2019). Designing autonomous vehicle incentive program with uncertain vehicle purchase price. *Transportation Research Part C: Emerging Technologies*, 103:226 – 245. 1.1

Chen, W., Jaques, N., Taylor, S., Sano, A., Fedor, S., and Picard, R. W. (2015). Wavelet-based motion artifact removal for electrodermal activity. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6223–6226. IEEE. 1.3.4

Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B., Servan-Schreiber, D., and Noll, D. C. (1994). Activation of the prefrontal cortex in a nonspatial working memory task with functional mri. *Human brain mapping*, 1(4):293–304. 1.3.2

De Waard, D. (1996). *The measurement of drivers' mental workload*. PhD thesis, Netherlands: University of Groningen. 1.2, 1.3.3, 1.2, 1.3.4

Demberg, V. (2013). Pupillometry: the index of cognitive activity in a dual-task study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35. 1.3.4, 1.3.4, 1.2

Di Nocera, F., Camilli, M., and Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, 1(3):271–285. 1.2, 1.3.4

Diaz-Piedra, C., Sebastián, M. V., and Di Stasi, L. L. (2020). Eeg theta power activity reflects workload among army combat drivers: an experimental study. *Brain sciences*, 10(4):199. 1.3.4

Du, N., Huang, K. Y., and Yang, X. J. (2020a). Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming. *Human Factors*, 62(6):987–1001. 3.3.5

Du, N., Kim, J., Zhou, F., Pulver, E., Tilbury, D. M., Robert, L. P., Pradhan, A. K., and Yang, X. J. (2020b). Evaluating effects of cognitive load, takeover request lead time, and traffic density on drivers' takeover performance in conditionally automated driving. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 66–73. 1.2

Du, N., Yang, X. J., and Zhou, F. (2020c). Psychophysiological responses to takeover requests in conditionally automated driving. *Accident Analysis & Prevention*, 148:105804. 1.3.2

Du, N., Zhang, Q., and Yang, X. J. (2018). Evaluating effects of automation reliability and reliability information on trust, dependence and dual-task performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 174–174. SAGE Publications Sage CA: Los Angeles, CA. 2.2

Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., and Yang, X. J. (2020d). Predicting driver takeover performance in conditionally automated driving. *Accident Analysis & Prevention*, 148:105748. 1.3.2

Eby, D. W., Molnar, L. J., Zhang, L., Louis, R. M. S., Zanier, N., Kostyniuk, L. P., and Stanciu, S. (2016). Use, perceptions, and benefits of automotive technologies among aging drivers. *Injury epidemiology*, 3(1):28. 1.1

Eggemeier, F., Biers, D., Wickens, C., Andre, A., Vreuls, D., Billman, E., and Schueren, J. (1990). Performance assessment and workload evaluation systems: Analysis of candidate measures. Technical report, Technical Report No HSD-TR-90. 1.2, 1.3.4

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395. 2.3

Fridman, L., Reimer, B., Mehler, B., and Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 652. ACM. 1.3, 1.3.4, 4.2, 4.2.2

Goldberg, J. H. and Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International journal of industrial ergonomics*, 24(6):631–645. 4.2.3

Griffiths, P. G. and Gillespie, R. B. (2005). Sharing control between humans and automation using haptic interface: Primary and secondary task performance benefits. *Human Factors*, 47(3):574–590. 1.1

Guo, Y. and Yang, X. J. (2020). Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics*. 3.3.5

Halverson, T., Estepp, J., Christensen, J., and Monnin, J. (2012). Classifying workload with eye movements in a complex task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):168–172. 1.3.4, 1.3, 3.2.4, 4.2, 4.2.1

Hancock, P. A., Parasuraman, R., and Byrne, E. A. (1996). Driver-centered issues in advanced automation. *Automation and human performance*. 1.2

Hancock, P. A., Wulf, G., Thom, D., and Fassnacht, P. (1990). Driver workload during differing driving maneuvers. *Accident Analysis & Prevention*, 22(3):281–290. 1.2, 1.2, 1.3.3

Hankins, T. C. and Wilson, G. F. (1998). A comparison of heart rate, eye activity, eeg and subjective measures of pilot mental workload during flight. *Aviation, space, and environmental medicine*, 69(4):360. 1.3.4, 1.3.4

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier. 1.3.3, 2.5.1, 3.3.3

Heard, J., Harriott, C. E., and Adams, J. A. (2018). A Survey of Workload Assessment Algorithms. *IEEE Transactions on Human-Machine Systems*, 48(5):434–451. 1.3.4, 1.3.4, 1.3.4, 1.3.4, 5.2.3

Hendy, K. C., Liao, J., and Milgram, P. (1997). Combining time and intensity effects in assessing operator information-processing load. *Human Factors*, 39(1):30–47. PMID: 9302878. 1.2

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, 7:935. 1.3.2

Hicks, T. G. and Wierwille, W. W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors*, 21(2):129–143. 1.2, 1.3.1

Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., and Fantini, S. (2009). Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2185–2194. 1.3.4

Hogervorst, M. A., Brouwer, A.-M., and Van Erp, J. B. (2014). Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience*, 8:322. 1.3, 1.3.4, 3.2.4, 4.2.1

Hoogendoorn, R. G., Hoogendoorn, S. P., Brookhuis, K. A., and Daamen, W. (2011). Adaptation longitudinal driving behavior, mental workload, and psycho-spacing models in fog. *Transportation research record*, 2249(1):20–28. 1.2

ISO/TC 22/SC 39 (2016). Road vehicles–transport information and control systems–detection-response task (DRT) for assessing attentional effects of cognitive load in driving. Standard ISO 17488:2016, International Organization for Standardization, Geneva, CH. 1.3.2

Jacob, R. J. (1995). Eye tracking in advanced interface design. *Virtual environments and advanced interface design*, 258:288. 1.3.4, 4.2.3

Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, pages 573–605. Elsevier. 1.3.4, 4.2.3

Jahn, G., Oehme, A., Krems, J. F., and Gelau, C. (2005). Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(3):255–275. 1.2

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71. 3.3.3

Jorna, P. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, 36(9):1043–1054. 1.3.4

Kahneman, D. and Beatty, J. (1966). Pupil diameter and load on memory. *Science*, 154(3756):1583–1585. 1.3.4, 1.2

Kannala, J. and Brandt, S. (2006). A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1335–1340. 2.3

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745. 3.2.3, 5.2.2

Kim, J.-J., Kim, M. S., Lee, J. S., Lee, D. S., Lee, M. C., and Kwon, J. S. (2002). Dissociation of working memory processing associated with native and second languages: Pet investigation. *NeuroImage*, 15(4):879–891. 1.3.2

Klingner, J., Kumar, R., and Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 69–72. 1.3.4, 1.2

Kosch, T., Hassib, M., Buschek, D., and Schmidt, A. (2018a). Look into my eyes: Using pupil dilation to estimate mental workload for task complexity adaptation. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA. Association for Computing Machinery. 1.3.4, 1.3, 3.2.4, 4.2, 4.2.1

Kosch, T., Hassib, M., Woźniak, P. W., Buschek, D., and Alt, F. (2018b). Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. 1.3

Kun, A. L., Palinko, O., Medenica, Z., and Heeman, P. A. (2013). On the feasibility of using pupil diameter to estimate cognitive load changes for in-vehicle spoken dialogues. In *INTERSPEECH*, pages 3766–3770. 1.3.4, 1.3.4, 1.2

Lansdown, T. C., Brook-Carter, N., and Kersloot, T. (2004). Distraction from multiple in-vehicle secondary tasks: vehicle performance and mental workload implications. *Ergonomics*, 47(1):91–104. 1.3.1

Lee, J. D., Wickens, C. D., Liu, Y., and Boyle, L. N. (2017). *Designing for People: An introduction to human factors engineering*. CreateSpace. 1.2, 1.3.2, 1.3.4, 1.4

Li, W.-C., Chiu, F.-C., and Wu, K.-J. (2012). The evaluation of pilots performance and mental workload by eye movement. 1.2

Liang, Y., Reyes, M. L., and Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *IEEE transactions on intelligent transportation systems*, 8(2):340–350. 1.3, 1.3.4, 4.2, 4.2.3

Liu, K. (2019). *Measuring and Quantifying Driver Workload on Limited Access Roads*. PhD thesis, University of Michigan. 1.3.1

Lu, S., Zhang, M. Y., Ersal, T., and Yang, X. J. (2019). Workload management in teleoperation of unmanned ground vehicles: Effects of a delay compensation aid on human operators' workload and teleoperation performance. *International Journal of Human–Computer Interaction*, pages 1–11. 1.2, 1.3.2, 3.3.5

Luo, R., Hayne, R., and Berenson, D. (2018). Unsupervised early prediction of human reaching for human–robot collaboration in shared workspaces. *Autonomous Robots*, 42(3):631–648. 4.2.2, 4.2.4

Luo, R., Wang, Y., Weng, Y., Paul, V., Brudnak, M. J., Jayakumar, P., Reed, M., Stein, J., Ersal, T., and Yang, X. J. (2019). Toward real-time assessment of workload: A bayesian inference approach. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage WA: Seattle, WA. 2.5.1

Ma, R. and Kaber, D. B. (2005). Situation awareness and workload in driving while using adaptive cruise control and a cell phone. *International Journal of Industrial Ergonomics*, 35(10):939–953. 1.2

Marquart, G., Cabrall, C., and de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3:2854–2861. 1.2, 1.3.4

Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. 1.2, 1.3.4

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*, pages 7–7. IEEE. 1.2, 1.3.4

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., and Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta psychologica*, 75(1):75–89. 1.2

Maye, J., Furgale, P., and Siegwart, R. (2013). Self-supervised calibration for robotic systems. *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 473–480. 2.3

Miller, E. E., Boyle, L. N., Jenness, J. W., and Lee, J. D. (2018). Voice control tasks on cognitive workload and driving performance: Implications of modality, difficulty, and duration. *Transportation Research Record*, 2672(37):84–93. 1.3.2

Moacdieh, N. M., Devlin, S. P., Jundi, H., and Riggs, S. L. (2020). Effects of workload and workload transitions on attention allocation in a dual-task environment: Evidence from eye tracking metrics. *Journal of Cognitive Engineering and Decision Making*, page 1555343419892184. 1.3.4, 1.2, 1.3.4, 4.2.3

Mulder, M., Abbink, D. A., and Boer, E. R. (2008). The effect of haptic guidance on curve negotiation behavior of young, experienced drivers. In *2008 IEEE International Conference on Systems, Man and Cybernetics*, pages 804–809. 1.1

Nguyen, A.-T., Sentouh, C., and Popieul, J.-C. (2018). Sensor reduction for driver-automation shared steering control via an adaptive authority allocation strategy. *IEEE/ASME Transactions on Mechatronics*, 23(1):5–16. 1.1

Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pages 420–423. 1.3.4

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59. 1.3.2

Palinko, O. and Kun, A. (2011). Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. 1.2

Palinko, O., Kun, A. L., Shyrokov, A., and Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pages 141–144. 1.3.4, 1.3.4, 1.2

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making*, 2(2):140–160. 1.2

Parks, D. L. and Boucek, G. P. (1989). Workload prediction, diagnosis, and continuing challenges. In *Applications of human performance models to system design*, pages 47–63. Springer. 1.2

Patten, C. J., Kircher, A., Östlund, J., Nilsson, L., and Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident Analysis & Prevention*, 38(5):887–894. 1.2

Petermeijer, S. M., Abbink, D. A., and de Winter, J. C. F. (2015). Should drivers be operating within an automation-free bandwidth? evaluating haptic steering support systems with different levels of authority. *Human Factors*, 57(1):5–20. 1.1

Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. In *Studies in visual information processing*, volume 6, pages 3–22. Elsevier. 1.3.4, 4.2.3

Rayner, K. (2009). The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506. 1.3.4, 4.2.3

Rayner, K. and Morris, R. K. (1990). Do eye movements reflect higher order processes in reading? In *From eye to mind: Information acquisition in perception, search, and reading.*, Studies in visual information processing, Vol. 1., pages 179–190. North-Holland, Oxford, England. 1.2, 1.3.4

Recarte, M. A. and Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of experimental psychology: Applied*, 6(1):31. 1.2, 1.3.4, 4.2.3

Recarte, M. A. and Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, 9(2):119. 1.2

Recarte, M. Á., Pérez, E., Conchillo, Á., and Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish journal of psychology*, 11(2):374. 1.2, 1.3.4

Reid, G. B. and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology*, volume 52, pages 185–218. Elsevier. 1.3.3

Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record*, 2138(1):13–19. 1.2

Reimer, B., Mehler, B., Coughlin, J. F., Godfrey, K. M., and Tan, C. (2009). An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st international conference on automotive user interfaces and interactive vehicular applications*, pages 115–118. 1.3.4, 1.3.4

Reimer, B., Mehler, B., Coughlin, J. F., Roy, N., and Dusek, J. A. (2011). The impact of a naturalistic hands-free cellular phone task on heart rate and simulated driving performance in two age groups. *Transportation research part F: traffic psychology and behaviour*, 14(1):13–25. 1.3.4

Rerhaye, L., Blaser, T., and Alexander, T. (2018). Evaluation of the index of cognitive activity (ica) as an instrument to measure cognitive workload under differing light conditions. In *Congress of the International Ergonomics Association*, pages 350–359. Springer. 1.2, 1.3.4

Robinson, G. H. (1979). Dynamics of the eye and head during movement between displays: A qualitative and quantitative guide for designers. *Human Factors*, 21(3):343–352. 4.2.3

Rozo, L., Silverio, J., Calinon, S., and Caldwell, D. G. (2016). Learning controllers for reactive and proactive behaviors in human–robot collaboration. *Frontiers in Robotics and AI*, 3:30. 3.2.1

Salvucci, D. D. and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. 1.3.4, 4.2.3

Samms, C. and Mitchell, D. (2010). Predicting the consequences of workload management strategies with human performance modeling. *Maryland: Army Research Laboratory*. (document), 1.2, 1.1

Schneegass, S., Pfleging, B., Broy, N., Heinrich, F., and Schmidt, A. (2013). 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*, pages 150–157. 1.3.4

Schutte, P. C. (2015). How to make the most of your human: Design considerations for single pilot operations. In Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics*, pages 480–491, Cham. Springer International Publishing. (document), 1.2, 1.1

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*. 3.2.1, 3.2.3, 4.2.4

Schweitzer, J. and Green, P. (2006). Task acceptability and workload of driving urban roads, highways, and expressway: Ratings from video clips (technical report umtri-2006-19). *Ann Arbor, MI: University of Michigan Transportation Research Institute*. 1.3.3

Seong, P. H., Kang, H. G., Na, M. G., Kim, J. H., Heo, G., and Jung, Y. (2013). Advanced mmis toward substantial reduction in human errors in npps. *Nuclear Engineering and Technology*, 45(2):125–140. (document), 1.2, 1.1

Shi, Y., Ruiz, N., Taib, R., Choi, E., and Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 2651–2656. 1.3.4

Stapel, J., Mullakkal-Babu, F. A., and Happee, R. (2019). Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving. *Transportation research part F: traffic psychology and behaviour*, 60:590–605. 1.2

Sterman, M. and Mann, C. (1995). Concepts and applications of eeg analysis in aviation performance evaluation. *Biological psychology*, 40(1-2):115–130. 1.3.4

Teh, E., Jamson, S., Carsten, O., and Jamson, H. (2014). Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. *Transportation research part F: traffic psychology and behaviour*, 22:207–217. 1.2

Tobii Pro AB (2014). Tobii pro lab. Computer software. 2.3

Toyota (2020). White paper: Toyota automated driving. Technical report, Toyota. 6.2

Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., and Jung, T.-P. (2007). Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine*, 78(5):B176–B185. 1.2

Tsimhoni, O. and Green, P. (1999). Visual demand of driving curves determined by visual occlusion. In *Vision in Vehicles Conference*. 1.2

Tsimhoni, O., Yoo, H., and Green, P. (1999). Effects of workload and task complexity on driving and task performance for in-vehicle displays as assessed by visual occlusion. *University of Michigan Transportation Research Institute*. 1.3.4

van der Wel, P. and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, 25(6):2005–2015. 1.3.4

Van Orden, K. F., Limbert, W., Makeig, S., and Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human factors*, 43(1):111–121. 1.2, 1.3.4

Verwey, W. B. (2000). On-line driver workload estimation. effects of road situation and age on secondary task measures. *Ergonomics*, 43(2):187–209. 1.3.2

Vicente, K. J., Thornton, D. C., and Moray, N. (1987). Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human factors*, 29(2):171–182. 1.3.4

Vogels, J., Demberg, V., and Kray, J. (2018). The index of cognitive activity as a measure of cognitive processing load in dual task settings. *Frontiers in Psychology*, 9:2276. 1.2, 1.3.4

Wang, J. and Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2.3

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, 50(3):449–455. 1.3.2

Wickens, C. D., Hollands, J. G., Banbury, S., and Parasuraman, R. (2015). *Engineering psychology and human performance*. Psychology Press. 1.3.1

Wierwille, W., Hanowski, R., Hankey, J., Kieliszewski, C., Lee, S., Medina, A., Keisler, A., and Dingus, T. (2002). Identification of driver errors: Overview and recommendations (fhwa-rd-02-003). Technical report, US Department of Transportation, Federal Highway Administration. 1.2

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005. 4.2.1

Xing, Y., Lv, C., Cao, D., Wang, H., and Zhao, Y. (2018). Driver workload estimation using a novel hybrid method of error reduction ratio causality and support vector machine. *Measurement*, 114:390–397. 1.3.1

Yang, X. J., Unhelkar, V. V., Li, K., and Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 408–416. IEEE. 2.2, 3.3.5

Yokoyama, H., Eihata, K., Muramatsu, J., and Fujiwara, Y. (2018). Prediction of driver's workload from slow fluctuations of pupil diameter. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1775–1780. IEEE. 1.3.4, 1.3

Zenati, M. A., Kennedy-Metz, L., and Dias, R. D. (2020). Cognitive engineering to improve patient safety and outcomes in cardiothoracic surgery. In *Seminars in thoracic and cardiovascular surgery*, volume 32, pages 1–7. Elsevier. (document), 1.2, 1.1

Zhang, Y., Owechko, Y., and Zhang, J. (2008). Learning-based driver workload estimation. In *Computational intelligence in automotive applications*, pages 1–17. Springer. 1.3.1, 1.3, 1.3.4

Zijlstra, F. and Van Doorn, L. (1985). *The construction of a scale to measure perceived effort*. University of Technology. 1.3.3