This paper has been accepted for publication and is forthcoming in International Journal of Human-Computer Interaction.

Real-time Workload Estimation Using Eye Tracking: A Bayesian Inference Approach

Ruikun Luo
Robotics Institute, University of Michigan, Ann Arbor, MI

Yifan Weng
Mechanical Engineering, University of Michigan, Ann Arbor, MI

Paramsothy Jayakumar
U.S. Army Ground Vehicles System Center, Warren, MI

Mark J. Brudnak
U.S. Army Ground Vehicles System Center, Warren, MI

Victor Paul
U.S. Army Ground Vehicles System Center, Warren, MI

Vishnu R. Desaraju
Toyota Research Institute, Ann Arbor, MI, USA

Jeffrey L. Stein
Mechanical Engineering, University of Michigan, Ann Arbor, MI

Tulga Ersal
Mechanical Engineering, University of Michigan, Ann Arbor, MI

X. Jessie Yang
Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI

Corresponding author: X. Jessie Yang, 1205 Beal Avenue, Ann Arbor, MI 48109, Email: xijyang@umich.edu. **Acknowledgement:** We acknowledge the technical and financial support of the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV19-2-0001 U.S. Army Ground Vehicle Systems Center (GVSC) Warren, MI (OPSEC #: Pending)

Abstract

Workload management is a critical concern in shared control of unmanned ground vehicles. In response to this challenge, prior studies have developed methods to estimate human operators' workload by analyzing their physiological data. However, these studies have primarily adopted a single-model-single-feature or a single-model-multiple-feature approach. The present study proposes a Bayesian inference model to estimate workload, which leverages different machine learning models for different features. We conducted a human subject experiment with 24 participants, in which a human operator teleoperated a simulated High Mobility Multipurpose Wheeled Vehicle (HMMWV) with the help from an autonomy while performing a surveillance task simultaneously. Participants' eye-related features, including gaze trajectory and pupil size change, were used as the physiological input to the proposed Bayesian inference model. Results show that the Bayesian inference model achieves a 0.823 F_1 score, 0.824 precision, and 0.821 recall, outperforming the single models.

Keyword: Human-automation interaction, Human-autonomy interaction, Bayesian inference, workload estimation

Introduction

Despite significant research and development efforts, it has been established that fully driverless vehicles are still at least several decades away (Leonard, Mindell, & Stayton, 2020). As such, there has been a growing focus on alternative approaches to leverage the potential benefits of advanced driving automation. One such approach is adaptive shared control, in which the human driver and the vehicle autonomy control the vehicle concurrently. Recent studies have suggested that adaptive shared control, when implemented in a manner that is sensitive to the human driver's workload, can lead to improved driving performance, higher levels of trust, and reduced control effort (Luo et al., 2021; Weng et al., 2020). The adaptive shared control system allocates higher control authority to the vehicle autonomy if the driver is overloaded; This scheme is especially beneficial when the driver needs to handle multiple tasks simultaneously and the cumulative workload can become excessive.

The success of such a adaptive shared control system hinges on the accurate estimation of a driver's cumulative workload. Human workload can be measured offline or online. Offline measures are assessed after a human operator finishes a task, typically using a questionnaire (e.g., NASA Task Load Index (Hart & Staveland, 1988)). However, offline measures are not applicable for designing real-time adaptive systems. To measure workload online, prior studies have used primary task performance (Hicks & Wierwille, 1979; Lansdown, Brook-Carter, & Kersloot, 2004; Liu, 2019), secondary task performance (Chang, Boyle, Lee, & Jenness, 2017; Lu, Zhang, Ersal, & Yang, 2019; Miller, Boyle, Jenness, & Lee, 2018; Owen, McMillan, Laird, & Bullmore, 2005), and physiological measures (Moacdieh, Devlin, Jundi, & Riggs, 2020; Sun et al., 2021). Primary and secondary task measures require task performance data, which may not be available in real time. Therefore, assessing workload using physiological measures has received substantial research attention.

Physiological measures of workload rely on changes in human physiological signals. Prior research has looked into various types of physiological signals for workload estimation, including electroencephalogram (EEG), functional near-infrared spectroscopy (fNIRS), galvanic skin response (GSR), heart rate indices, and eye-related signals (see Heard, Harriott, and Adams (2018) and Skaramagkas, Giannakakis, et al. (2021) for detailed reviews). Eye-related signals, due to their non-intrusiveness and robustness to movement (Moacdieh et al., 2020; Zhou, Yang, & de Winter, 2022), have been increasingly used to assess operators' workload, and are the focus of the present study. Various types of eye-related signals have been examined in previous literature. They can be broadly categorized into three groups: 1) pupil-related features, 2) blink-related features, and 3) gaze-related features. Table 1 illustrate some metrics in each category. Please note that Table 1 should not be treated as an exhaustive list.

1) Pupil-related features. Pupil diameters are widely used to assess human workload (Demberg, 2013; Palinko, Kun, Shyrokov, & Heeman, 2010; van der Wel & van Steenbergen, 2018). Prior research reveals that pupil diameter, pupil diameter change, and pupil diameter change rate increase under high workload (Palinko et al., 2010; van der Wel & van Steenbergen, 2018). Ahlstrom and Friedman-Berg (2006) found that the human operators' mean pupil diameter was significantly larger when using a static storm forecast tool (i.e., high workload) than when using a dynamic storm forecast tool. Pupil diameter change is the difference between people's pupil diameter and the baseline pupil diameter, whereas pupil diameter change rate is the first order derivative of pupil diameter over time. Klingner, Kumar, and Hanrahan (2008) found that the pupil diameter change increased under high workload during three 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). Palinko et al. (2010) found that the mean pupil diameter change rate was sensitive to cognitive load during driving. The rate increased when drivers experienced high workload.

TABLE 1: *The list of eye-related features that has been used to indicate or estimate workload.*

Metric	References
Pupil diameter	Ahlstrom and Friedman-Berg (2006); M. A. Recarte and Nunes (2000, 2003) Vogels, Demberg, and Kray (2018) M. Á. Recarte, Pérez, Conchillo, and Nunes (2008)
Pupil diameter change	Ahern and Beatty (1979); Backs and Walrath (1992) Klingner et al. (2008) Kun, Palinko, Medenica, and Heeman (2013) Benedetto et al. (2011); Palinko and Kun (2011) Palinko et al. (2010); Skaramagkas, Ktistakis, et al. (2021)
Pupil diameter change rate	Palinko et al. (2010)
ICA	Marshall (2000, 2002) Demberg (2013); Vogels et al. (2018) Rerhaye, Blaser, and Alexander (2018)
Blink duration	De Waard (1996); Van Orden, Limbert, Makeig, and Jung (2001) Ahlstrom and Friedman-Berg (2006); Benedetto et al. (2011) Skaramagkas, Ktistakis, et al. (2021)
Blink rate	De Waard (1996); Van Orden et al. (2001) Benedetto et al. (2011); M. Á. Recarte et al. (2008) Skaramagkas, Ktistakis, et al. (2021); Tsai, Viirre, Strychacz, Chase, and Jung (2007)
Blink latency	Carmody (1994); Eggemeier et al. (1990)
Fixation frequency	Backs and Walrath (1992); Skaramagkas, Ktistakis, et al. (2021); Van Orden et al. (2001)
Fixation duration	Backs and Walrath (1992); Rayner and Morris (1990) M. A. Recarte and Nunes (2000); Skaramagkas, Ktistakis, et al. (2021) Li, Chiu, and Wu (2012); Marquart, Cabrall, and de Winter (2015)
Variability of fixation duration	M. A. Recarte and Nunes (2000)
Variability of fixation position	M. A. Recarte and Nunes (2000); Reimer (2009)
Percentage of fixations in area of interest (AOI)	M. A. Recarte and Nunes (2000)
Saccadic extent	May, Kennedy, Williams, Dunlap, and Brannan (1990) M. A. Recarte and Nunes (2000); Van Orden et al. (2001) Skaramagkas, Ktistakis, et al. (2021)
Saccadic amplitude	Moacdieh et al. (2020); Skaramagkas, Ktistakis, et al. (2021)
Saccadic velocity	Mallick, Slayback, Touryan, Ries, and Lance (2016) Bodala, Ke, Mir, Thakor, and Al-Nashash (2014); He, Wang, Gao, and Chen (2012)
Saccadic rate	Menekse Dalveren and Cagiltay (2018); Y. Yang, McDonald, and Zheng (2012) Gao, Yan, and Sun (2015)
NNI (Nearest Neighbor Index)	Di Nocera, Camilli, and Terenzi (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)

Furthermore, instead of directly using pupil diameter, pupil diameter change, and pupil diameter change rate, researchers defined 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, where higher ICA values indicate higher cognitive workload (Demberg, 2013; Rerhaye et al., 2018; Vogels et al., 2018).

2) Blink-related features. Various blink-related features have been investigated in the previous literature, such as blink duration, blink rate, and blink latency (De Waard, 1996; Heard et al., 2018; Marquart et al., 2015). Blink duration is the length of a blink, and it decreases under high workload (Ahlstrom & Friedman-Berg, 2006). Blink rate, also called blink frequency, is the number of blinks per minute. M. Á. 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). Their results showed 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. In addition, 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. Prior studies showed that blink latency increases as cognitive and visual workload increases (Carmody, 1994; Eggemeier et al., 1990).

3) Gaze-related features. Gaze-related features are based on fixations and saccades, the two phases of 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), whereas saccades are the rapid eye movements between fixations (Jacob, 1995; Jacob & Karn, 2003; Salvucci & Goldberg, 2000). The metrics computed from fixations and saccades can be broadly categorized into two groups: temporal information and spatial information (Marquart et al., 2015). Temporal information 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 high workload (Backs & Walrath, 1992; Marquart et al., 2015; Rayner & Morris, 1990; M. A. Recarte & Nunes, 2000; Van Orden et al., 2001). Spatial information includes various measures to describe gaze distribution. For example, M. A. Recarte and Nunes (2000) investigated a number of fixation-related features when drivers perform mental tasks (verbal or spatial imagery) while driving on highways and on regular 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 the distribution is expected to be random.

Previous studies looking into eye-related features have largely focused on uncovering the relationships between physiological features and workload (Demberg, 2013; Kun et al., 2013; Palinko et al., 2010); for example, pupil diameter increases as workload increases. Recently, researchers started to use machine learning techniques for workload estimation by formulating it as a supervised classification problem (Heard et al., 2018) (Table 2).

Kosch, Hassib, Buschek, and Schmidt (2018) applied Support Vector Machines (SVMs) with a linear kernel to human operators' pupil dilation data for workload classification and achieved a 0.79 accuracy on average. Instead of using pupil diameters in a time domain, Yokoyama, Eihata, Muramatsu, and Fujiwara (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, Estepp, Christensen, and Monnin (2012) used SVMs with various 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 numerous features they studied (i.e., blink duration, blink frequency, closure, fixation duration, NNI, percentage of eye closure [PERCLOS], pupil diameter, saccade duration, saccade frequency, and saccade velocity), they found that pupil diameter from a five-second time window with a linear kernel achieved the best performance.

TABLE 2: *Machine learning studies for workload estimation using eye-related features. "Within-participants" means that the the training data and testing data are from the same participant. "Cross-participants" means that the training data and testing data are from different participants.*

Reference	**Model**	**Feature**	**Evaluation** Method
Chen and Epps (2013)	Gaussian Mixture Models (GMMs)	Pupil diameter, saccadic amplitude, fixation duration	Within-participants
Liang, Reyes, and Lee (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, saccade frequency, saccade velocity, NNI, percentage eye closure	Within-participants
Yokoyama et al. (2018)	SVM (linear kernel)	High and low Frequency power of pupil size variation	Within-participants
Kosch, Hassib, Buschek, and Schmidt (2018)	SVM (linear kernel)	Pupil dilation	Within-participants
Kosch, Hassib, Woźniak, Buschek, and Alt (2018)	SVM	Gaze deviation from reference track	Within- and cross-participants
Zhang, Owechko, and Zhang (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, Reimer, Mehler, and Freeman (2018)	HMM, Convolutional neural network (CNN)	Gaze trajectory, eye image	Cross-participants
Hogervorst, Brouwer, and Van Erp (2014)	SVM (linear kernel), Elastic net	Pupil size, blink rate, blink duration, other EEG and ECG features	Within-participants

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 several features into one feature vector (Chen & Epps, 2013; Liang et al., 2007; Zhang et al., 2008). 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 the Gaussian Mixture Model (GMM) 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.

As Table 2 shows, 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. These methods have two major limitations: First, the single machine learning model for a single feature method lacks robustness and is susceptible for changes in contextual and environmental factors. For example, SVM works well with pupil diameter data under constant lighting conditions, however, it is not suitable for outside environment where dramatic lighting changes might happen. Second, the single machine learning model for multiple features method may have difficulty analyzing all available features due to the inherent property of a machine learning model. For example, SVM requires a constant feature size (i. e., the length of the feature vector should be the same), and therefore cannot be used when both pupil diameter and fixation trajectory data are available, because the feature size for the fixation trajectory data varies among people (i. e., During a period, a person may fixate his/her eye on one spot or several spots, leading to different feature sizes.)

To over come the above-mentioned limitations, in the present study, we propose a Bayesian inference model to estimate human workload. The Bayesian inference approach can leverage different machine learning models, each of which may work best for particular features. For example, prior literature shows the SVM model has superior results in analyzing pupil size change and the HMM model has superior results in analyzing gaze trajectory. The proposed Bayesian inference approach is able to "merge" the two machine learning models, each of which has been proven to work well with a particular feature. In the present study, we used the Bayesian inference approach to merge four different machine learning models for four different features, i.e., SVMs for pupil size change, HMM for gaze trajectory, SVMs for fixation feature, and GMMs for fixation trajectory.

Methods

This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board at the University of Michigan (Application #: HUM00154094).

Participants

A total of 25 university students participated in the experiment. Data from one participant were discarded due to equipment malfunction. The remaining 24 participants were on average 25.9 years old (*SD* = 3.4 years) and had an average of 6.5 years of driving experience (*SD* = 3.9 years). There were 10 females and 14 males in the remaining 24 participants.

Participants in the study met the following inclusion criteria: (1) be 18 years old and above; (2) be in possession of a valid driving license; (3) have normal or corrected to normal vision; and (4) have normal or corrected-to-normal hearing.

Experimental apparatus and stimuli

The study employed a dual-task shared control simulation platform for teleoperation of a simulated notional High Mobility Multipurpose Wheeled Vehicle (HMMWV). In this testbed, participants performed two tasks simultaneously: a driving task and a surveillance task, as shown in Figure 1. In the driving task, a participant and an autonomy shared the control of the steering of the HMMWV at a fixed cruising speed of 15 m/s (around 34 mph) to drive as close to the centerline as possible. To simulate perception failures of the autonomy, an offset was introduced such that the autonomy tracked a line which deviated from the centerline by one meter. During the experiment, the positions of the monitors and the steering wheel were fixed. The screen of the driving task was approximately 95 cm in front of the participant. The experiment was under the normal room lighting condition.

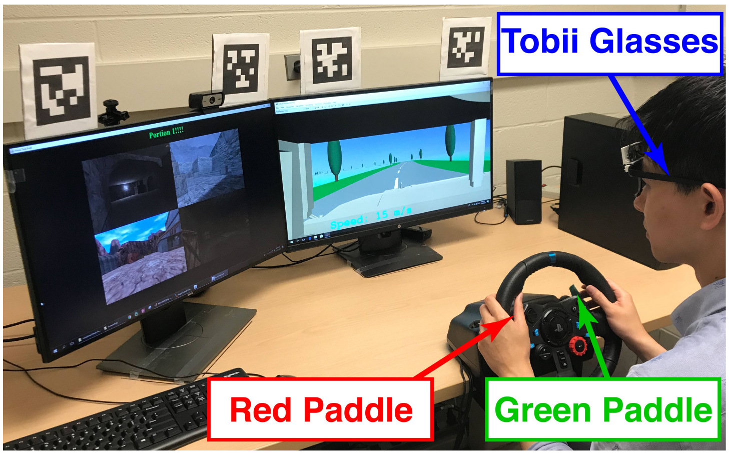


Figure 1. Illustration of the dual-task shared control simulation platform. The left screen shows the surveillance task and the right screen shows the driving task.

In the surveillance task, the participant received image feeds and identified potential threats (Figure 2). If the participant identified a threat, s/he needed to press the red paddle on the steering wheel to report “danger." Otherwise, the participant pressed the green paddle to report “clear" (see Figure 1). The potential threat appeared in only one of the four images in a given set with threat. The screenshots were selected with the same difficulty benchmarking prior studies (Du, Huang, & Yang, 2020; Guo & Yang, 2021; X. J. Yang, Unhelkar, Li, & Shah, 2017).

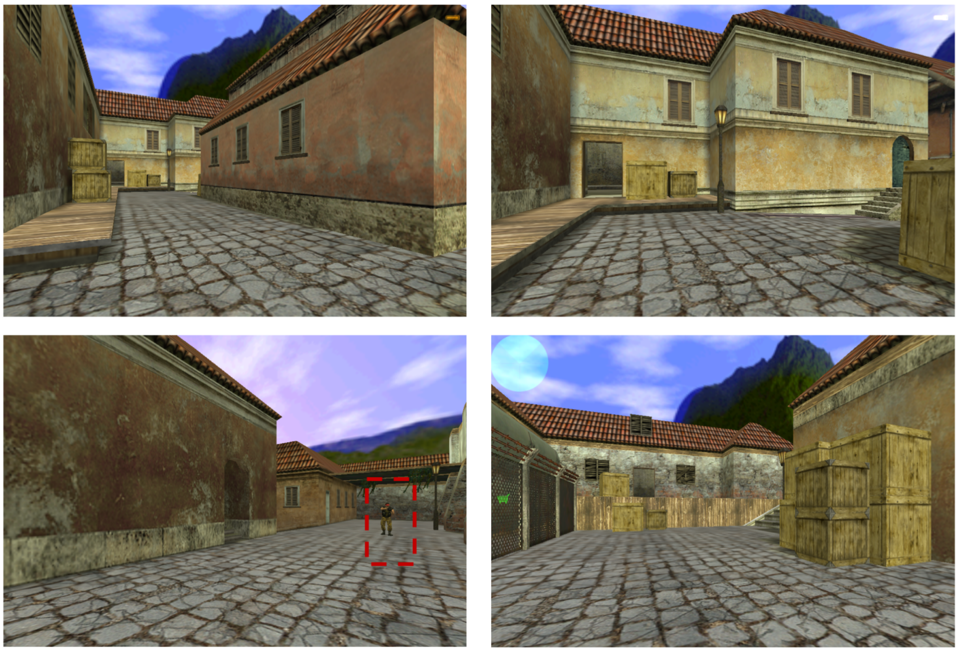


Figure 2. Illustration of the surveillance task. A threat appears in the lower left.

Figure 3 shows the flow of 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 budget, which was varied to manipulate the workload level (See Appendix B for more details on the selection of time budgets used in the present study).

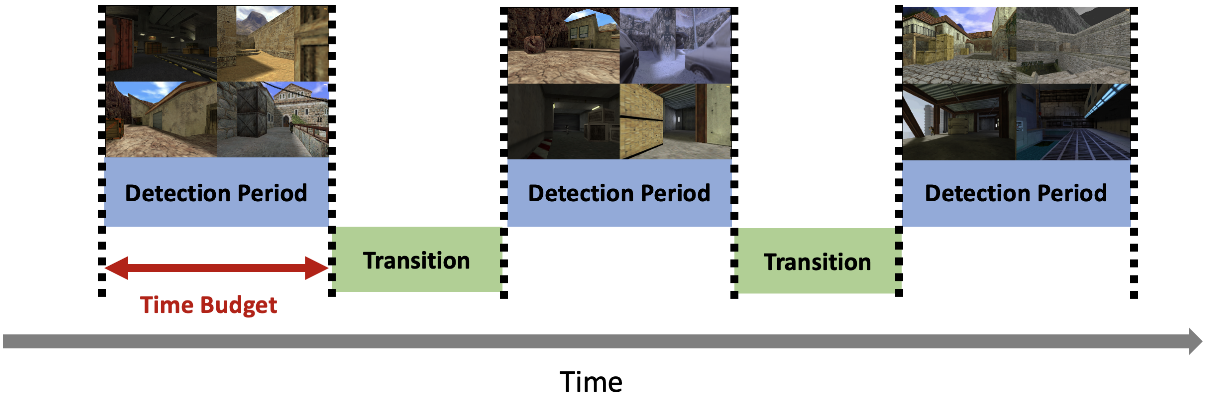


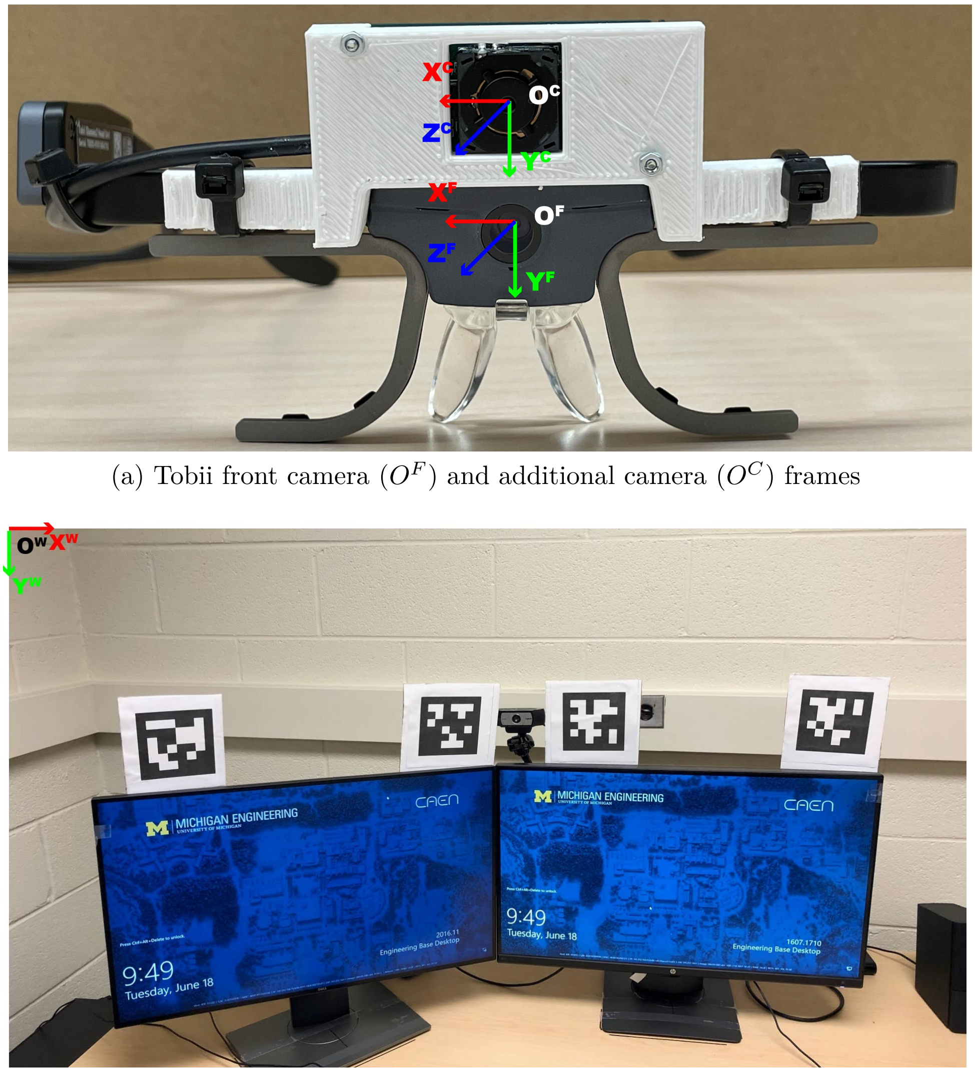
Figure 3. Pipeline for surveillance task. Participants receive image feeds and identify potential threats within the time budget. There is a transition period two sets of image feeds using a white screen. The transition period lasts for one second.

Experiment design

We manipulated the workload of the task by varying the time budgets for detection in the surveillance task. During the experiment, the participants drove on six different tracks, each lasting for approximately three minutes (see Appendix A for more details on the selection of tracks). Every track was equally segmented into three portions, and each portion had a different time budget for detection, 1.5, or 2.5, or 6.5 s. The order of presentation for the time budgets on each track was balanced by two 3 × 3 Latin squares.

Measures

Participants wore a pair of the Tobii Pro Glasses 2 (Tobii Pro AB, 2014), to measure human pupil sizes and gaze points in real time. In our study, we required human gaze points in the world frame, indicated as O^W in Figure 4b (i. e., the coordinates of the gaze point should be with respect to the world) to estimate human workload. However, the frame of reference of the Tobii Pro Glasses 2 is with respect to the Tobii front camera, indicated as O^F in Figure 4a). To convert the frame of reference from the Tobii front camera to the world frame in real time, we built one additional camera on top of the Tobii Pro Glasses 2, indicated as O^c and performed homography transformation. To perform this conversion, the additional camera detects the locations of the AprilTags (Wang & Olson, 2016) attached on top of the monitors (see Figure 4b). Although the Tobii Pro Glasses 2 provides gaze points and pupil sizes at 50 Hz, we down-sampled the Tobii Pro Glasses 2 to 30 Hz due to the computation limitation, i.e., the additional camera can only capture and process images at 30 Hz.



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

(b) World image frame (O^W)

Figure 4. Coordinate systems for the Tobii front camera (O^F), the additional camera (O^C), and the world frame (O^W).

Experiment procedure

Participants provided a signed informed consent and filled in a demographic survey. After that, they received a training session, in which they first performed a driving-only task for approximately 1.5 minutes to become familiar with the haptic shared control with autonomy and then performed three 1-minute trials of the surveillance task with the $6.5-, 2.5-, 1.5-$second time budget for detection. After that, the participants performed the driving and surveillance tasks together on three different tracks, each lasting 1.5 minutes.

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 three minutes.

Bayesian Inference Model for Workload Estimation

As mentioned in Table 2, researchers have investigated different machine learning models for different eye-related features for workload estimation. For example, previous studies showed that SVMs could be used with human pupil dilation (Kosch, Hassib, Buschek, & Schmidt, 2018) and fixation features (i.e., fixation duration) (Liang et al., 2007) to estimate human workload. In addition, different kernels have been used for different features (i.e., the linear kernel for pupil dilation (Kosch, Hassib, Buschek, & Schmidt, 2018) and the RBF kernel for fixation duration (Liang et al., 2007)). In the present study, we propose a Bayesian inference model that can leverage the different machine learning models that work best for different features.

Figure 5 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:

$$\begin{aligned}
& p(W_L|X_1, X_2, ..., X_n) \\
\propto\ & p(X_1, X_2, ..., X_n|W_L)p(W_L) \\
=\ & p(W_L) \sum_{M_1, M_2, ..., M_n} p(X_1, X_2, ..., X_n, M_1, M_2, ..., M_n|W_L) \\
=\ & p(W_L) \sum_{M_1, M_2, ..., M_n} p(X_1, X_2, ..., X_n|M_1, M_2, ..., M_n, W_L)P(M_1, M_2, ..., M_n|W_L) \\
=\ & p(W_L) \sum_{M_1, M_2, ..., M_n} \{\prod_{M_i} p(M_i|W_L)p(X_i|M_i)\} \\
=\ & p(W_L) \prod_{M_i}\{\sum_{M_i} p(M_i|W_L)p(X_i|M_i)\}
\end{aligned} \tag{1}$$

$p(W_L)$ 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 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 the present study, we investigated four eye-related features. We selected three features from the literature, including the 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 of the four features, we used a machine learning model that works well for a feature.

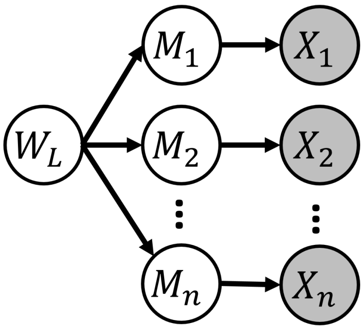


Figure 5. A graphical representation of the Bayesian inference model. W_L is the human's workload. M_i represents the workload estimated by different machine learning models. X_i is the feature for the different machine learning models.

Support-vector machines (SVMs) for pupil size change

In the experiment, we used the Tobii Pro Glasses 2 to measure the 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, Hassib, Buschek, & Schmidt, 2018). The SVM is a supervised learning algorithm that aims to find the optimal hyperplane that separates data points into clusters. We found that using an RBF kernel can achieve better performance than a linear kernel for the pupil size change feature. We used pairwise coupling to estimate probability $p(X_1|M_1)$ for a multi-class classification problem, where each class represents a workload level (Wu, Lin, & Weng, 2004).

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).

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 (g_x^t, g_y^t) shown as the magenta dots in Figure 6. 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 6. Thus, $B = \{\mu_j, \Sigma_j\}$.

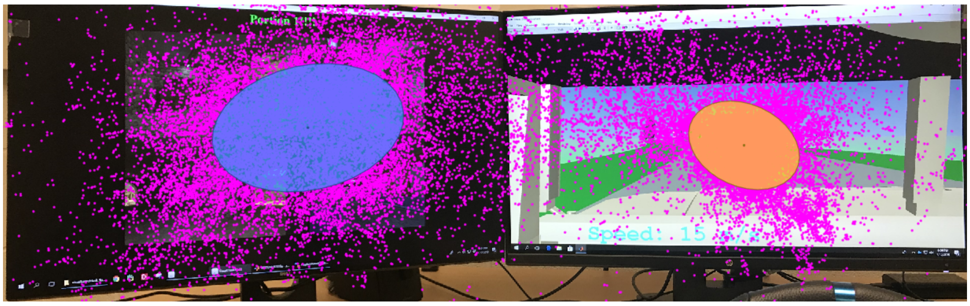


Figure 6. Example of using the Hidden Markov Model to model gaze trajectory to estimate workload. Magenta dots: gaze points. Ellipsoids: Multivariate normal distributions.

We trained multiple HMMs, each for a different workload level w. 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 learned sets of HMM parameters $\Theta_w = (\pi, A, B)$, one set 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, Silverio, Calinon, and Caldwell (2016) and Calinon (2016). To learn the parameters, we defined four probabilities:

$$
\begin{aligned}
\alpha_i^l(t)^k &= P(O_1 = \mathbf{o}_1^l, ..., O_t = \mathbf{o}_t^l, S_t = i | \Theta^k) \\
\beta_i^l(t)^k &= P(O_{t+1} = \mathbf{o}_{t+1}^l, ..., O_T = \mathbf{o}_T^l | S_t = i, \Theta^k) \\
\gamma_i^l(t)^k &= P(S_t = i | \mathcal{O}_l, \Theta^k) \\
\xi_{i,j}^l(t)^k &= P(S_t = i, S_{t+1} = j | \mathcal{O}_l, \Theta^k)
\end{aligned}
\tag{2}
$$

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:

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

The two steps were iterated until convergence. The number of hidden states was determined by the Bayesian Information Criterion (BIC) (Calinon & 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 different workload levels w. The forward algorithm is similar to the recursive update of α in the E-step of the EM algorithm. We have $P(\mathcal{O}|\tilde{\Theta}_w) = \sum_{i=1}^N \tilde{\alpha}_i(T)$. 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, we used a geometric mean of the probability density of a trajectory

(Luo, Hayne, & Berenson, 2018), shown as follows:

$$\hat{p}(X_2|M_2 = w) = \sqrt[T]{P(\mathcal{O}|\tilde{\Theta}_w)} \tag{3}$$

Support-vector machines (SVMs) for fixation feature

Human eye movement can be broken into 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 examining (Jacob, 1995; Rayner, 1995, 2009). Saccades are the rapid eye movements between fixations (Jacob, 1995; Jacob & Karn, 2003; Salvucci & Goldberg, 2000). Given a sequence of gaze points, researchers have proposed various criteria to determine a fixation. The center of a fixation is typically within $2-3°$ (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). Figure 7 illustrates a set of fixations 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.

Researchers have found that a number of measurements related to fixations and saccades can indicate human workload (Moacdieh et al., 2020; M. A. Recarte & Nunes, 2000). 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 than an RBF kernel for the fixation feature. Similar to pupil size change,

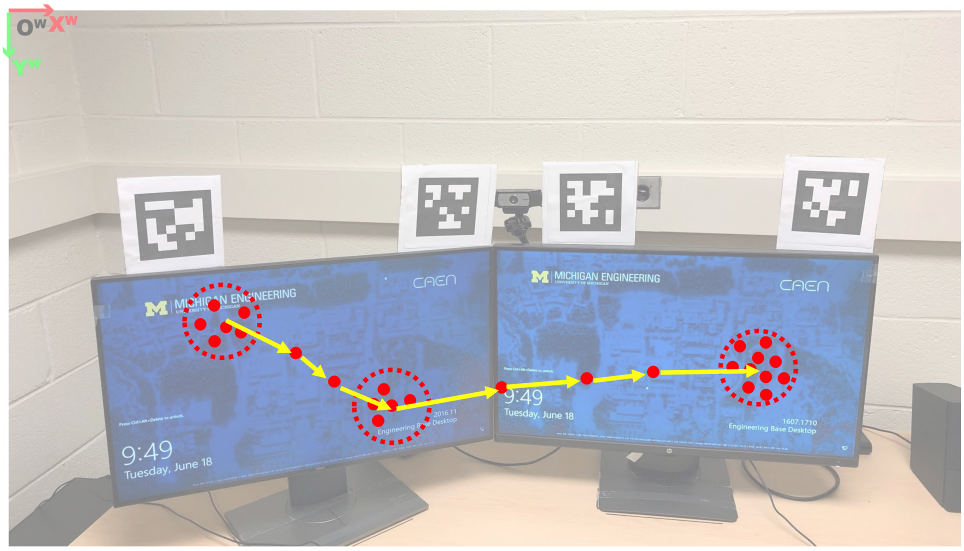


Figure 7. Illustration of fixations and saccades mapped on the world image. Red dots are gaze points. Red dashed circles are fixations. Yellow arrows are saccades.

we can use the pairwise coupling method to estimate $p(X_3|M_3)$.

Gaussian mixture models (GMMs) for fixation trajectory

The fixation feature X_3 ignores the spatial information of the fixations. Therefore, we developed 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. The order of the fixations does not matter. Therefore, we used GMMs to model the fixation trajectory. Similar to the HMM, we learned a GMM for each level of workload M_4^w, where w represents different workload levels. Given an observation X_4, the output of a GMM was the likelihood $p(X_4|M_4^w)$.

Each GMM 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 GMM M_4^w represented by K multivariate Gaussians is given by:

$$p(\xi^l|M_4^w) = \sum_{k=1}^{K} p(gc_k|M_4^w)p(\xi^l|gc_k, M_4^w) \tag{4}$$

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:

$$\begin{aligned} p(\xi^l|gc_k, M_4^w) &= \mathcal{N}(\mu_k, \Sigma_k) \\ &= \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} e^{-\frac{1}{2}(\xi^l-\mu_k)^T\Sigma_k^{-1}(\xi^l-\mu_k)} \end{aligned} \tag{5}$$

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:

$$\hat{p}(X_4|M_4^w) = \prod_{l=1}^{L} p(\xi^l|M_4^w) \tag{6}$$

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 used the geometric mean of the probability density of a trajectory (Luo et al., 2018), shown as follows:

$$p(X_4|M_4^w) = \sqrt[L]{\prod_{l=1}^{L} p(\xi^l|M_4^w)} \tag{7}$$

Similar to the HMM, we used the BIC (Calinon & Billard, 2005; Schwarz et al., 1978) 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 EM algorithm.

Results

Data processing

Participants drove on six different tracks. Each track was segmented into three portions, and each portion had a different time budget for detecting potential threats. We treated the portion with 1.5 s time budget as the high workload portion and the portion with 6.5 s time budget as the moderate workload portion (see Appendix A for details).

We evaluated our proposed Bayesian inference model against other single models

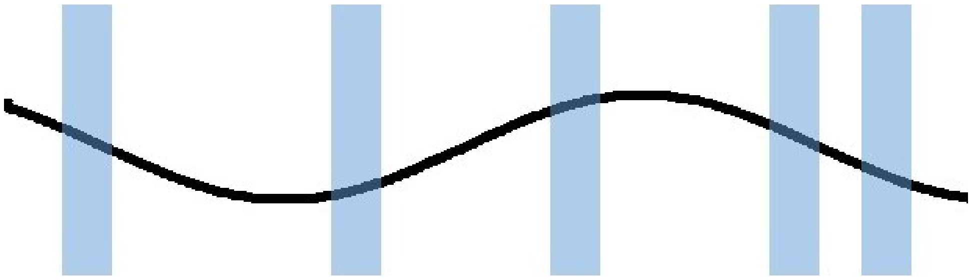


Figure 8. An example for 5 sequences of data selected from a portion for cross-participants evaluation. Blue boxes represents the randomly selected sequences of data, each lasting 4 s.

in two 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 as shown in Figure 8. For the within-participants evaluation, we randomly selected 20 sequences of data from each portion in each trial as shown in Figure 9. For each 4-second sequence of data, we extracted the four features $X_1, ..., X_4$.

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 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 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

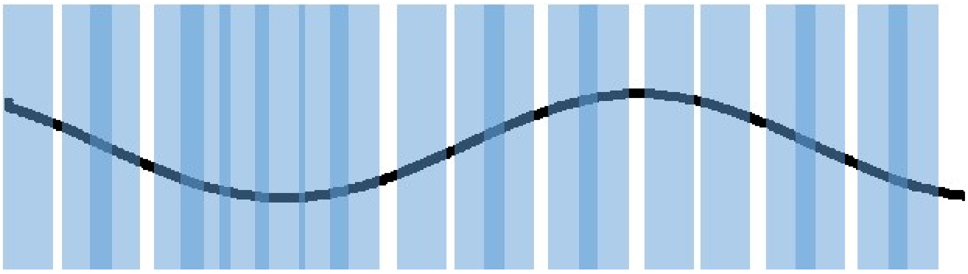


Figure 9. An example for 20 sequences of data selected from a portion for within-participants evaluation. Blue boxes represents the randomly selected sequences of data, each lasting 4 s. The shades indicates potential overlaps between two adjacent sequences.

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 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 3 is the mean and standard error over the 50 rounds of holdouts. Table 4 shows the pairwise t-test results for the overall performance between our proposed Bayesian inference model.

TABLE 3: *Overall performance of the Bayesian inference (BI) model and other single models for cross-participants evaluation.*

	Bayesian inference (BI)	**SVMs pupil size change**	**HMM gaze trajectory**	**SVMs fixation feature**	**GMMs fixation trajectory**
$\mathbf{F_1}$ 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: *Pairwise t-tests between the 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**
$\mathbf{F_1}$ 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	$t(49) = 10.34, p < .001$	$t(49) = 39.38, p < .001$	$t(49) = 24.12, p < .001$	$t(49) = 32.70, p < .001$

The results indicate that our proposed Bayesian inference model significantly outperforms the single models alone using cross-participants evaluation. Our proposed Bayesian inference model achieved a 0.823 ± 0.004 F_1 score, 0.824 ± 0.004 precision, and

0.821 ± 0.004 recall using cross-participants.

Within-participants evaluation

The within-participants evaluation 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), and can be considered a personalized model. We used the k-fold cross validation for the within-participants evaluation, where k was 6, as there were 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 5 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 reveal that our proposed Bayesian inference model achieved a 0.85 ± 0.01 F_1 score, 0.86 ± 0.01 precision, and 0.85 ± 0.01 recall on average using within-participants evaluation.

TABLE 5: *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

Continued on next page

Table 5 – continued from previous page

	Bayesian	SVMs	HMM	SVMs	GMMs
	inference	pupil size change	gaze trajectory	fixation feature	fixation trajectory
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
	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

Continued on next page

Table 5 – continued from previous page

	Bayesian	SVMs	HMM	SVMs	GMMs
	inference	pupil size change	gaze trajectory	fixation feature	fixation trajectory
	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

Discussion and Conclusion

In the present study, we proposed a Bayesian inference model for workload estimation that can leverage different machine learning models for different features. By merging four different machine learning models for four different features, i.e., SVMs for pupil size change, HMM for gaze trajectory, SVMs for fixation feature, and GMMs for fixation trajectory, our proposed Bayesian inference model can achieve an average F_1 score of 0.823 ± 0.004 using cross-participants evaluation and an average F_1 score of 0.85 ± 0.01 using within-participants evaluation for workload estimation. As shown in Figure 5, the proposed Bayesian inference model can integrate workload estimation results of any machine learning models including deep learning models. In the present study, the base models were SVM, HMM and GMM, all of which are considered traditional machine learning models. They were chosen because that they required less data and offered interpretale results. Furthermore, this Bayesian inference model model can be applied to any scenario where a non-intrusive measure of workload is needed, including adaptive shared control (Luo et al., 2021).

The cross-participants evaluation and the within-participants evaluation have their advantages and disadvantages, and therefore are particularly suitable for certain contexts. The cross-participants evaluation can be considered a *population-based model*, which is generalizable to any human operator. This approach is convenient to use once developed. However, in order to build it, a set of training data is required. 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 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 training dataset. However, this approach requires more trials for each participant and hence much longer experiment time.

We notice several limitations and directions for future research. In the present study, the different levels of human workload were induced by manipulating the surveillance task urgency. The results indicate that our proposed Bayesian inference model distinguishes the different workload levels caused by different surveillance task urgency. However, it is unknown if the proposed Bayesian inference model is able to classify different workload levels caused by other factors. Also, the population of participants in our experiments were young adults. Different age groups may have different patterns for certain physiological signals under different workload conditions. Future research should investigate the generalizability of the proposed method to other contexts where varying workload is caused by other factors such as driving speed, road curvature, surrounding traffic, weather, and etc., and to other populations.

In addition, in the present study, we treated the workload estimation problem as a classification problem and segmented the time series of 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 extracted feature vectors for classifiers. Future work could take into account the workload dynamics (i. e., patterns of workload changes over time) to improve the workload estimation performance. As our proposed Bayesian inference model is based on the graphical model, it can be naturally extended to a 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.

Acknowledgement

We acknowledge the technical and financial support of the University of Michigan Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-19-2-0001 U.S. Army CCDC Ground Vehicle Systems Center (GVSC), Warren, MI.

Appendix A

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.

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).

Experimental apparatus and stimuli: Pilot study 1 used the same platform as in the experiment except that only driving task was involved. 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 A1). The presentation of tracks followed a 10 × 10 Latin square design to eliminate potential order effects.

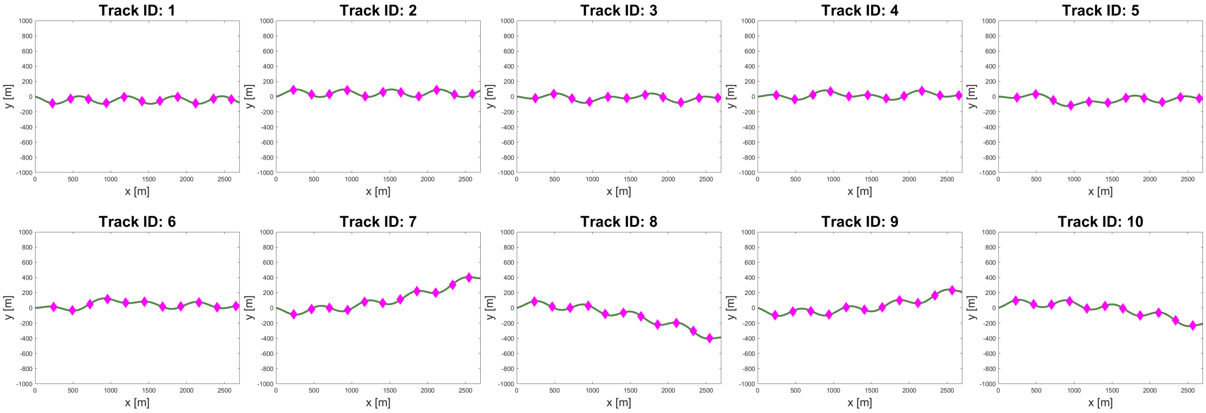


Figure A1. Candidate tracks. Magenta dots indicate the locations where the participants reported the difficulty of driving.

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 A1 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.

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 the experiment.

Appendix B

Pilot Study 2 – Design of Surveillance Task

In the present study, 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 budget for the detection period of the surveillance task so that the difficulty and workload of the dual-task mission could be manipulated.

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 same platform as in the experiment.

Experimental design: Pilot Study 2 used a within-subjects design with six different time budgets for the detection period of the surveillance task: 1.5, 2.5, 3.5, 4.5, 5.5, and 6.5 seconds (i.e., participants had to complete the detection task within the given time budget). The six time budgets were selected based on the results from our previous study (Luo et al., 2019). Participants performed both the driving task and the surveillance task on six different tracks, each with a different constant time budget for the detection period. The presentation of surveillance task conditions followed a 6 × 6 Latin square design to eliminate potential order effects.

Measures: Participants reported their workload of the dual-task mission using the NASA TLX survey (Hart & 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 budget 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 budgets. 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.

Results: We first conducted an omnibus test to see if the time budget affected participants' difficulty and workload. A Wilcoxon Signed Ranks showed that the 1.5 s condition is significantly more difficult than the 6.5 s condition ($Z = -2.214, p = .027$), and the 1.5 s condition has a significantly higher workload than the 6.5 s condition ($Z = -3.066, p = .002$). We then performed paired-sample t tests to select two time-budgets to be used in the experiment. We expected to see large differences in difficulty and workload between the selected time budgets and therefore focused on the *difference* between the 1.5 s and the 6.5 s conditions. The Shapiro–Wilk test showed normality for the difference of difficulty and of workload (difficulty: $D = .215, p = .2$; workload: $D = .174, p = .2$). A paired-sample t-test showed that the 1.5 s condition is significantly more difficult than the 6.5 s condition ($t = 8.306, p < .001$, Cohen's $d = 3.39$, large effect), and the 1.5s condition has a significantly higher workload than the 6.5s condition ($t(11) = 7.45, p < .001$, Cohen's $d = 2.15$, large effect).

Based on the results, we selected 1.5 s and 6.5 s time budgets to be used in the experiment to induce varying levels of workload. Note that in the experiment, we also included the 2.5 s time budget, as we were interested in exploring participants' performance with a slightly larger time budget compared to the 1.5 s time budget.

References

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

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636.

Backs, R. W., & 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.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(3), 199–208.

Bodala, I. P., Ke, Y., Mir, H., Thakor, N. V., & Al-Nashash, H. (2014). Cognitive workload estimation due to vague visual stimuli using saccadic eye movements. In *2014 36th annual international conference of the ieee engineering in medicine and biology society* (pp. 2993–2996).

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

Calinon, S., & 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* (pp. 105–112).

Carmody, M. A. (1994). *Current issues in the measurement of military aircrew performance: A consideration of the relationship between available metrics and operational concerns.* (Tech. Rep.). Naval Air Warfare Center.

Chang, C.-C., Boyle, L. N., Lee, J. D., & 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.

Chen, S., & 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.

De Waard, D. (1996). *The measurement of drivers' mental workload* (Unpublished doctoral dissertation). Netherlands: University of Groningen.

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* (Vol. 35).

Di Nocera, F., Camilli, M., & 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.

Du, N., Huang, K. Y., & Yang, X. J. (2020). 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.

Eggemeier, F., Biers, D., Wickens, C., Andre, A., Vreuls, D., Billman, E., & Schueren, J. (1990). *Performance assessment and workload evaluation systems: Analysis of candidate measures* (Tech. Rep.). Human Systems Division, Air Force Systems Command.

Fridman, L., Reimer, B., Mehler, B., & Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 652).

Gao, X., Yan, H., & Sun, H.-j. (2015). Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons. *Journal of vision*, *15*(3), 3–3.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.

Guo, Y., & Yang, X. J. (2021). Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, *13*, 1899–1909.

Halverson, T., Estepp, J., Christensen, J., & 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.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

He, X., Wang, L., Gao, X., & Chen, Y. (2012). The eye activity measurement of mental workload based on basic flight task. In *Ieee 10th international conference on industrial informatics* (pp. 502–507).

Heard, J., Harriott, C. E., & Adams, J. A. (2018). A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems*, *48*(5), 434–451.

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

Hogervorst, M. A., Brouwer, A.-M., & 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.

Jacob, R. J. (1995). Eye tracking in advanced interface design. *Virtual Environments and Advanced Interface Design*, *258*, 288.

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

Klingner, J., Kumar, R., & 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* (pp. 69–72).

Kosch, T., Hassib, M., Buschek, D., & Schmidt, A. (2018). 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* (p. 1–6). New York, NY, USA: Association for Computing Machinery.

Kosch, T., Hassib, M., Woźniak, P. W., Buschek, D., & Alt, F. (2018). Your eyes tell:

Leveraging smooth pursuit for assessing cognitive workload. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).

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

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

Leonard, J. J., Mindell, D. A., & Stayton, E. L. (2020). Autonomous vehicles, mobility, and employment policy: The
roads ahead. *MIT Task Force on Work of the Future Research Brief*. Retrieved from `https://workofthefuture.mit.edu/research-post/autonomous-vehicles-mobility-and-em`

Li, W.-C., Chiu, F.-C., & Wu, K.-J. (2012). The evaluation of pilots performance and mental workload by eye movement. In *Proceeding of the 30th european association for aviation psychology conference.*

Liang, Y., Reyes, M. L., & 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.

Liu, K. (2019). *Measuring and quantifying driver workload on limited access roads* (Unpublished doctoral dissertation). University of Michigan.

Lu, S., Zhang, M. Y., Ersal, T., & 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*, 1–11.

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

Luo, R., Wang, Y., Weng, Y., Paul, V., Brudnak, M. J., Jayakumar, P., … 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.*

Luo, R., Weng, Y., Wang, Y., Jayakumar, P., Brudnak, M. J., Paul, V., . . . Yang, X. J. (2021). A workload adaptive haptic shared control scheme for semi-autonomous driving. *Accident Analysis & Prevention*, *152*, 105968.

Mallick, R., Slayback, D., Touryan, J., Ries, A. J., & Lance, B. J. (2016). The use of eye metrics to index cognitive workload in video games. In *2016 ieee second workshop on eye tracking and visualization (etvis)* (pp. 60–64).

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

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

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* (pp. 7–7).

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

Menekse Dalveren, G. G., & Cagiltay, N. E. (2018). Insights from surgeons' eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions. *Behaviour & Information Technology*, *37*(5), 517–537.

Miller, E. E., Boyle, L. N., Jenness, J. W., & 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.

Moacdieh, N. M., Devlin, S. P., Jundi, H., & 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*, 1555343419892184.

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

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

Palinko, O., Kun, A. L., Shyrokov, A., & 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* (pp. 141–144).

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

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.

Rayner, K., & 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.* (pp. 179–190). Oxford, England: North-Holland.

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, *6*(1), 31.

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

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

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

Rerhaye, L., Blaser, T., & 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* (pp. 350–359).

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.

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

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

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*.

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., … Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*.

Skaramagkas, V., Ktistakis, E., Manousos, D., Tachos, N. S., Kazantzaki, E., Tripoliti, E. E., … Tsiknakis, M. (2021). Cognitive workload level estimation based on eye tracking: A machine learning approach. In *2021 ieee 21st international conference on bioinformatics and bioengineering (bibe)* (pp. 1–5).

Sun, Z., Li, B., Duan, F., Jia, H., Wang, S., Liu, Y., … Solé-Casals, J. (2021). Wlnet: Towards an approach for robust workload estimation based on shallow neural networks. *IEEE Access*, *9*, 3165-3173. doi: 10.1109/ACCESS.2020.3044732

Tobii Pro AB. (2014). *Tobii pro lab.* Computer software. Danderyd, Stockholm. Retrieved from `http://www.tobiipro.com/`

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

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

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

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

Wang, J., & Olson, E. (2016, October). AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS).*

Weng, Y., Luo, R., Jayakumar, P., Brudnak, M. J., Paul, V., Desaraju, V. R., . . . Ersal, T. (2020). Design and evaluation of a workload-adaptive haptic shared control framework for semi-autonomous driving. In *2020 american control conference (acc)* (pp. 4369–4374).

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

Yang, X. J., Unhelkar, V. V., Li, K., & 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)* (pp. 408–416).

Yang, Y., McDonald, M., & Zheng, P. (2012). Can drivers' eye movements be used to monitor their performance? a case study. *IET Intelligent Transport Systems*, *6*(4), 444–452.

Yokoyama, H., Eihata, K., Muramatsu, J., & Fujiwara, Y. (2018). Prediction of driver's workload from slow fluctuations of pupil diameter. In *2018 21st international conference on intelligent transportation systems (itsc)* (pp. 1775–1780).

Zhang, Y., Owechko, Y., & Zhang, J. (2008). Learning-based driver workload estimation. In *Computational intelligence in automotive applications* (pp. 1–17).

Springer.

Zhou, F., Yang, X. J., & de Winter, J. C. F. (2022). Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE Transactions on Intelligent Transportation Systems*, *23*(3), 2284-2295. doi: 10.1109/TITS.2021.3069776