Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving

AUTHOR(S):

Chen, Weiya; Sawaragi, Tetsuo; Hiraoka, Toshihiro

CITATION:

Chen, Weiya ...[et al]. Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving. Transportation Research Part F: Traffic Psychology and Behaviour 2022, 89: 109-128

ISSUE DATE:

2022-08

URL:

http://hdl.handle.net/2433/285054

Transportation Research Part F: Psychology and Behaviour 89 (2022) 109–128

Contents lists available at ScienceDirect

Transportation Research Part F: Psychology and Behaviour

journal homepage: www.elsevier.com/locate/trf

Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving

Weiya Chen [a,*], Tetsuo Sawaragi [a], Toshihiro Hiraoka [b]

[a] Deaprtment of Mechanical Engineering and Science, Kyoto University, Kyoto, Japan
[b] Institute of Industrial Science, University of Tokyo, Tokyo, Japan

ARTICLE INFO

Keywords:
Eye-tracking
Autonomous driving
Mental workload
Multitasking

ABSTRACT

The objective of this study was to verify the effectiveness of eye-tacking metrics in indicating driver's mental workload in semi-autonomous driving when the driver is engaged in different non-driving related tasks (NDRTs). A driving simulator was developed for three scenarios (high-, medium-, and low-mental workload presented by SAE (Society of Automotive Engineers) Levels 0, 1, and 2) and three uni-modality secondary tasks. Thirty-six individuals participated in the driving simulation experiment. NASA-TLX (Task Load Index), secondary task performance, and eye-tracking metrics were used as indicators of mental workload. The subjective rating using the NASA-TLX showed a main effect of autonomous level on mental workload in both visual and auditory tasks. Correlation-matrix calculation and principal-component extraction indicated that pupil diameter change, number of saccades, saccade duration, fixation duration, and 3D gaze entropy were effective indicators of a driver's mental workload in the visual and auditory multi-tasking situations of semi-autonomous driving. The accuracy of predicting the mental-workload level using the K-Nearest Neighbor (KNN) classifier was 88.9% with bootstrapped data. These results can be used to develop an adaptive multi-modal interface that issues efficient and safe takeover requests.

1. Introduction

"Society 5.0" for Japanese society is a concept in which economic development and the resolution of social issues become compatible through a system that highly integrates cyberspace (virtual space) and the physical space (real space) (Cabinet Office, Government of Japan). In Society 5.0, the relationship between humans and automated systems will become increasingly closer, and human–machine cooperation will occur on a daily basis. One of the most affected areas may be mobility and transportation, in which autonomous driving is expected to dramatically change the modes of movement.

1.1. Take-Over-Request

Not only is autonomous-driving technology based on computer vision and machine learning rapidly developing, human–machine interaction is also drawing increasing attention. At SAE (Society of Automotive Engineers) Level 3, the driver does not have to monitor the driving environment and is free to engage in non-driving related tasks (NDRTs), such as reading and using smartphones. However,

* Corresponding author at: 615-8246 C3 Building, Kyoto University, Kyoto, Japan.
E-mail address: weiyachen15@gmail.com (W. Chen).

https://doi.org/10.1016/j.trf.2022.05.004
Received 30 August 2021; Received in revised form 21 November 2021; Accepted 5 May 2022
Available online 17 June 2022

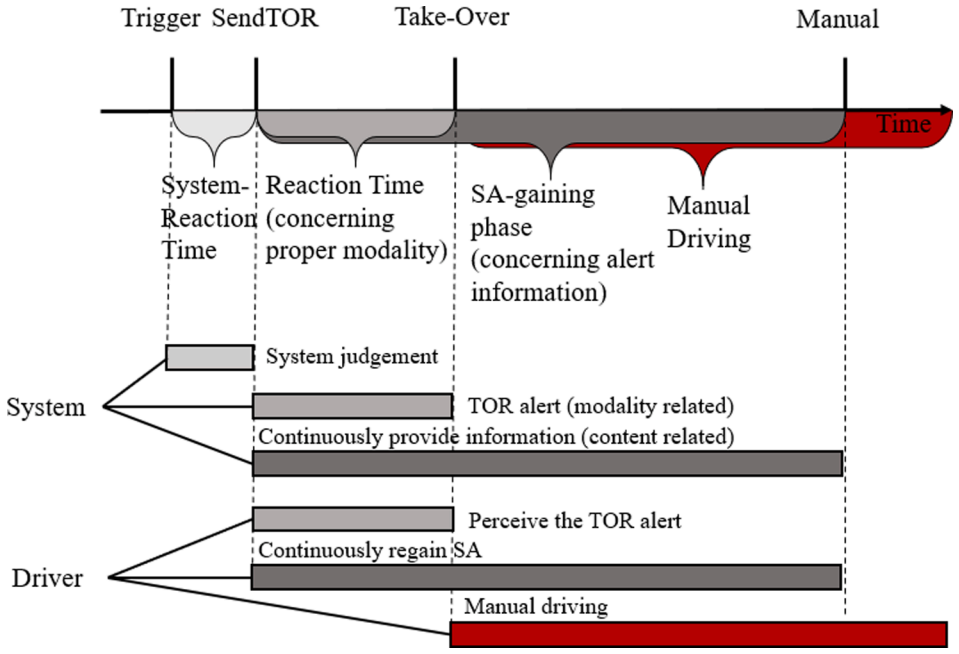


Fig. 1. TOR process as time sequence.

when the autonomous system determines that the concurrent driving situation is too complicated to handle properly, it issues a takeover request (TOR) and passes control of the vehicle back to the driver. One indicator that can be used to evaluate human–machine interaction is mental workload (MWL). MWL is an important design concept for exploring interactions between people and technological devices (Longo, 2015). To issue TORs in an effective manner, i.e., to get the driver's attention and help him/her rebuild situation awareness (SA) as soon as possible, each TOR should place as low an MWL on the driver as possible.

The MWL allocation to the driver should be known before a TOR is issued so that the system can issue the TOR on non-occupied cognitive channels for a lower MWL in accordance with the adaptive multi-modal interface model proposed by Chen et al. (Chen, Sawaragi, & Hiraoka, 2021). In this study, we defined four time points in the TOR process: the driving automation system of the vehicle detects a forthcoming difficult situation (Trigger); the vehicle system informs the driver of potential danger (SendTOR); the driver takes control (Takeover); and the vehicle becomes stable and safe after the driver has sufficient SA and properly operates the vehicle (Manual). The TOR process is illustrated as a sequence of phases in Fig. 1.

Multi-modal interfaces have been reported to be more effective than visual-only or auditory-only interfaces with respect to aircraft (Sklar & Sarter, 1999; Levulis, DeLucia, & Kim, 2018; Riggs & Sarter, 2019) and autonomous driving (Bazilinskyy, Petermeijer, Petrovych, Dodou, & de Winter, 2018; Tang, Guo, Zhang, Zhang, & Wu, 2020; Geitner, Biondi, Skrypchuk, Jennings, & Birrell, 2019; Huang, Steele, Zhang, & Pitts, 2019; Kalb, Streit, & Bengler, 2018). In the autonomous driving, a crowdsourcing study (Bazilinskyy et al., 2018) found that multi-modal TORs are the most preferred option in high-urgency situations, while auditory TORs are the most preferred in low-urgency scenarios and visual-only TORs are preferred over tactile-only TORs. Tang et al. (2020) compared implementations of auditory TOR and tactile TOR in a driving simulator. The auditory modality was found to have a more positive effect on response time compared with tactile TOR. In an autonomous driving simulation study, Geitner et al. (2019) compared three in-vehicle warnings (auditory, tactile, and auditory-tactile), and their results showed that the reaction times to the tactile warning (for the emergency braking event) were significantly slower compared with the auditory and auditory-tactile (i.e., multi-modal) warnings. In addition, the multi-modal warning led to fewer missed warnings and fewer false responses. Huang et al. (2019) investigated the effects of uni-, bi- and tri-modal combinations of visual, auditory, and tactile cues on response times to TOR and found that the tri-modal combinations had the shortest response time. Kalb et al. (2018) studied the differences among visual, auditory, and tactile modalities in TOR in a driving simulator experiment; they did not observed a main effect of modality type on task load but recommended to use auditory information via content rather than the position of the stimulus.

The adaptive multi-modal interface proposed by Chen et al. (2021) can be used to monitor drivers engaged in NDRTs and acquire the corresponding MWL allocations of NDRTs (phase 1 of TOR, i.e., system-reaction time). Although the fact that inconsistent TORs could be confusing to the driver at the early stage, with the limited number of TOR combinations, the driver can get used to all TOR forms after the training period. The adaptive multi-modal interface reduces the MWL of a driver when he/she is already familiar with autonomous driving and TORs. When there is a TOR, the interface will use the least occupied cognitive channel to alert the driver (phase 2, i.e., reaction time). When the MWL allocations of manual driving, i.e., SAE Level 0, are determined (phase 4, i.e., manual driving), alert-information content (phase 3, i.e., SA-gaining phase) enabling the driver to regain her/his SA can be sent on non-occupied cognitive channels.

1.2. Eye tracking metrics

We focused on phase 1 (system-reaction time) and examined whether the driver's monitoring system can determine the MWL placed on the driver. In most autonomous driving vehicles on the market or under development, the driver is monitored using a camera. The driver's behavior can be analyzed in real time through computer vision. Therefore, which cognitive channel is occupied can be determined from the camera image. However, video alone is insufficient to determine the MWL the driver is under. To develop an adaptive multi-modal interface, besides cognitive channels, we need to also understand the qualitative relations among the cognitive channels.

Eye-tracking metrics have been proven to be convincing indicators of multi-tasking performance in semi-autonomous driving scenarios (Yang & Kuo, 2020; Matton, Paubel, & Puma, 2020; Wang, Reimer, Dobres, & Mehler, 2014; Li, Schroeter, Rakotonirainy, Kuo, & Lenńe, 2020; Niezgoda, Tarnowski, Kruszewski, & Kamiński, 2015; Faure, Lobjois, & Benguigui, 2016; Jeong & Liu, 2018; Biswas & Prabhakar, 2018; Yoon & Ji, 2018; Huang, Bian, Zhao, Xu, & Rong, 2019; Chihara, Kobayashi, & Sakamoto, 2020; Wilkie, Mole, Giles, Merat, Romano, & Markkula, 2018) and other research fields such as surgical training or general psychological research (Wu, Cha, Sulek, Zhou, Sundaram, Wachs, & Yu, 2020; Benedetto et al., 2010; Kosch & Hassib, 2018). Yang et al. (2020) measured the off-road glance duration under different levels of distraction and found that high distraction caused extremely long glance durations of more than 30 s. Among the various eye-tracking metrics, pupil diameter change (Matton et al., 2020; Li et al., 2020; Niezgoda et al., 2015), blink duration (Benedetto et al., 2010), horizontal gaze dispersion (Wang et al., 2014), blink frequency (Faure et al., 2016; Chihara et al., 2020), gaze duration, saccade number (Huang et al., 2019), the standard deviation (SD) of horizontal eyeball rotation (Chihara et al., 2020), and deviations of gaze points (Kosch & Hassib, 2018) have been suggested as strong indicators of MWL. Niezgoda et al.'s study (Faure et al., 2016) involved an auditory secondary task, but only in full manual driving. An auditory secondary task was performed after a TOR was issued in Wilkie et al.'s experiments (Wilkie et al., 2018). They concluded that the driver state (cognitive load and gaze direction) during automatic driving may have important consequences on whether manual takeover of vehicle control is successful. Jeong et al. (Jeong & Liu, 2018) used a visual secondary task to study driving performance and safety on horizontal curves. Their experimental results indicated that drivers performing non-driving-related tasks involving visual stimuli or manual responses on curved roads fixated less frequently and with shorter durations on the road and showed poorer lane-keeping performance compared with other modalities. Similarly, Biswa et al. (Biswas & Prabhakar, 2018) used visual secondary tasks in a driving simulation and found that saccadic intrusion increases with cognitive load. Relations among different NDRTs and MWL were examined by Yoon et al. (Yoon & Ji, 2018). They found that the NDRT type has a significant effect as well as a positive correlation between the performance dimension and takeover.

However, MWL concerning only auditory tasks before TOR has not been examined with eye-tracking metrics. In normal driving situations, drivers mentally process auditory information, such as when they are listening to a podcast or judging a neighboring vehicle's location after hearing a car horn. Therefore, not only visual secondary tasks but also auditory secondary tasks should be considered when the autonomous driving system monitors the driver's condition at SAE Level 3. The main monitoring methods used by autonomous driving systems are camera-based and do not require special equipment for measuring physiological indicators. Although universal physiological indicators, such as heart rate or skin galvanic response, are not affected by the task type, eye-tracking metrics are affected by visual task content. We examined whether current driver monitoring systems can determine the driver's MWL state when he/she is engaged in visual and auditory NDRTs. We used an eye tracker instead of a camera. Although other researchers have successfully assessed the MWL of drivers by using computer-vision analysis of camera images (Fridman, Reimer, Mehler, & Freeman, 2018), we believe that an eye tracker provides data with higher accuracy.

2. Hypotheses

When considering a visual task, it is instinctive to conclude that such a task places different MWL on the driver at different autonomous driving levels because the monitoring task varies at different levels, so at a high autonomous driving level there should be more visual cognition resource remaining. The visual multi-tasking situation at a high autonomous level should have lower MWL. According to previous studies, the auditory tasks' influence a driver's MWL during semi-autonomous driving has not been verified. Therefore, we devised two hypotheses.

(1) H1: Different semi-autonomous levels place different MWL on the driver in both visual and auditory multi-tasking situations
(2) H2: Eye-tracking metrics are effective indicators for MWL-level prediction in both visual and auditory multi-tasking situations

The first hypothesis focuses on whether the auditory task places different MWL at different autonomous driving levels, while the second hypothesis examines the possibility of using eye-tracking metrics to predict the MWL level no matter if a visual or auditory multi-tasking situation occurs.

3. General methods

3.1. Driving simulation

We modified a semi-autonomous driving simulator that is based on the Dash Self-Driving Simulator (mattbradley). One autonomous driving level, SAE Level 1, was added to the simulator, so the driving simulation included three autonomous driving levels: SAE

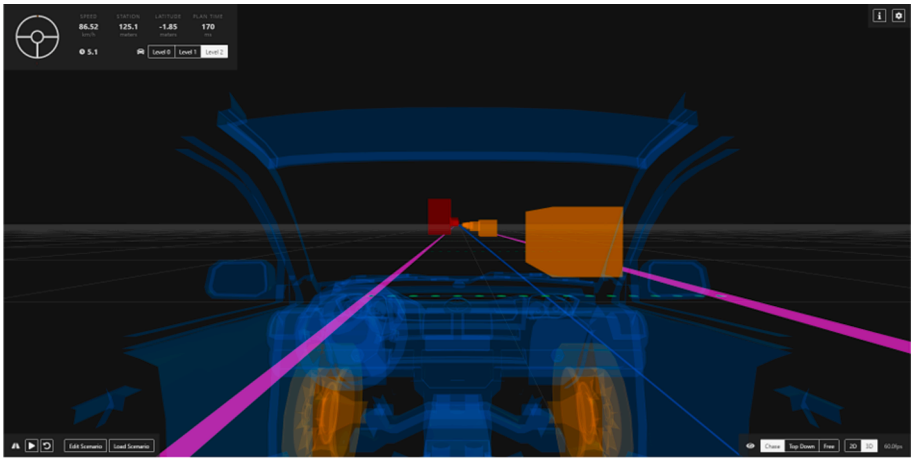


Fig. 2. Simulation display.

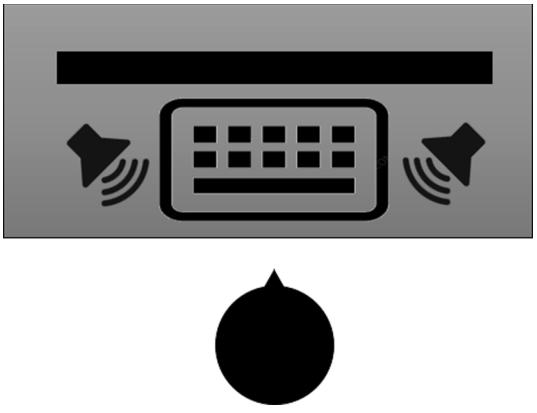


Fig. 3. Simulation layout.

Levels 0, 1 and 2. Although the final goal of this research project is to construct an adaptive multi-modal interface for TORs at Level 3, we designed the experiment only up to Level 2 for two reasons. First, drivers occasionally do not satisfy driving-environment monitoring requirements in the real world, e.g. a Tesla driver watching YouTube in autopilot mode when he/she is supposed to be monitoring the situation. Second, this experiment involved no TOR simulation. If we conduct the experiment at Level 3, the participants only need to focus on the secondary tasks instead of dealing with a multi-tasking situation. We plan to conduct Level 3 simulations in future TOR simulation experiments.

At SAE Level 1, the autonomous system is responsible for speed control while the driver should control the wheel. The simulation display is shown in Fig. 2. The yellow blocks represent dynamic obstacles and red blocks represent static obstacles. The participants used the keyboard to control the simulation. This driving simulator is less realistic compared with those using steering-wheel and pedal controls. Future experiments will be conducted with a more realistic driving simulator. The autonomous driving level can be selected by mouse-clicking the top-left buttons. A simple user interface (UI) is in the top-left corner, where the user can see the speed, acceleration and turning angle at any autonomous driving level. At Levels 1 and 2, the UI information can help the user see whether the system is accelerating/decelerating or steering automatically. The participants were told that the autonomous driving might be invalid during the experiment and were requested to monitor the driving environment all the time. The headway time and lane-keeping status can be continuously checked by combing the UI information and simulation vehicle dynamics.

3.2. Secondary tasks

The basis of the secondary tasks was the 2-back task. The reason for choosing the 2-back task is that working memory is crucial in the TOR process. To handle a TOR, the driver has to access long-term memory about manual driving. However, working memory of the driving environment is more important for the driver to take proper action because it provides situation information from the recent past. In contrast, long-term memory about manual driving has been turned into driving skill, which can be conducted autonomously. Insufficient working memory should be the main cause for TOR failure. We chose three secondary tasks; each one involved one

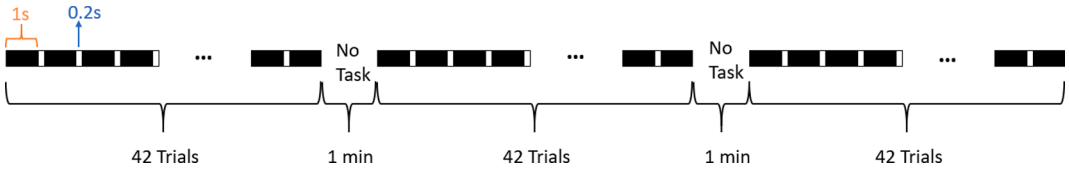


Fig. 4. Secondary-task time sequence.

perceptual modality and one cognitive channel in accordance with Chen et al.'s study (Chen et al., 2021). The visual-verbal task involved presenting a number from 0 to 9 on a tablet close to the simulator screen. The auditory-spatial task was to identify the location of a resource from the sound of a beep: left, middle and right (refer to Fig. 3: only left speaker, both speakers, only right speaker). The auditory-verbal task involved presenting audio of a number (from 0 to 9) spoken in English. The three types were determined on the basis of the possibility of them using suitable cognitive channels for the TOR alert. In Chen et al.'s study (Chen et al., 2021), the visual-verbal task imposed the least amount of MWL on the driver among the visual tasks. Since the auditory-visual cognitive channel is barely used in driving, we decided that the remaining two auditory channels (auditory-spatial and auditory- verbal) would be occupied with the other two secondary tasks.

The experiment had a within-subject design, which means each participant performed three secondary tasks at three autonomous levels, so there were 3*3 = 9 con– ditions. The tasks under each condition had three sets; each set had 42 stimuli. The stimuli were presented for 1 s and the time interval between two consecutive stimuli was 0.2 s. The sequence of the secondary tasks is shown in Fig. 4.

3.3. Participants

The experiment included three autonomous driving levels and three secondary tasks; considering counter-balancing, the number of participants should at least be a multiple of 6 and ideally a multiple of 36 (6 for levels and 6 for tasks). To determine the sample size with a significance level larger than 0.05, a priori F test power analysis (ANCOVA) was conducted using G*Power (Cohen; Faul et al., 2009). The parameters of the sample-size calculation are listed in Table B1. The G*Power calculation showed that more than 251 samples would be necessary because there were three levels and three tasks, so more than 251/9 = 28 participants should make for a powerful enough test. Accordingly, we recruited 36 students of Kyoto University (M = 18, F = 18, mean age = 26, age SD = 2.7971). Every participant received 2300 JPY as a reward.

4. Data collection

4.1. NASA-TLx

The NASA-TLX (Task Load Index) (Hart and Staveland, 1988) is one of the most commonly used subjective rating techniques for MWL. The NASA-TLX has six sub-scales: mental demand, physical demand, temporal demand, performance, effort, and frustration. After the individual weighting for these six sub-scales, the participants needed to score every sub-scale from 0 to 10. The total NASA-TLX was calculated as a weighted sum. The higher the NASA-TLX was, the higher MWL the participant felt. The NASA-TLX is an indicator of MWL and physical workload. When the physical workload is kept at similar level, the NASA-TLX can be regarded as an effective indicator for MWL.

4.2. Secondary-task performance

Another important indicator of MWL is the level of performance of secondary tasks. The better the participant performed a secondary task, the lower the MWL that task was supposed to place on the participant. The performance evaluation for the 2-back task follows Kane's method (Kane, Conway, Miura, & Colflesh, 2017): Hit for target items (8 items), FA (false answer) for lure items (8 misleading items), and FA for other foils (other 24 items).

4.3. Eye-Tracking metrics

A portable eye tracker, Tobii Pro Glasses 2.0 (Tobii Technology AB, Danderyd, Sweden), was used to binocularly sample eye movements at 50 Hz. The entire experiment was recorded. Recordings were annotated using Tobii Pro Lab Software (Tobii Technology AB) and extracted for further analysis. Nine eye-tracking metrics were analyzed: pupil diameter change, number of saccades, saccade duration, number of fixations, fixation duration, number of blinks, blink duration, 2D entropy, and 3D gaze entropy. The pupil diameter change was measured as the pupil diameter during the secondary task minus that in the driving-only situation at Level 0. Saccade and fixation were classified using Tobii Pro Lab's built-in algorithm, while the blink parameters and gaze entropies were calculated using raw data. We defined a blink as a 50- to 500-ms blank in the raw data in accordance with Wang's study (Wang, Toor, Gautam, & Henson, 2011). Two-dimensional gaze and 3D gaze entropies were calculated on the basis of the Shannon entropy (Shannon, 2001) using the gaze-point information:

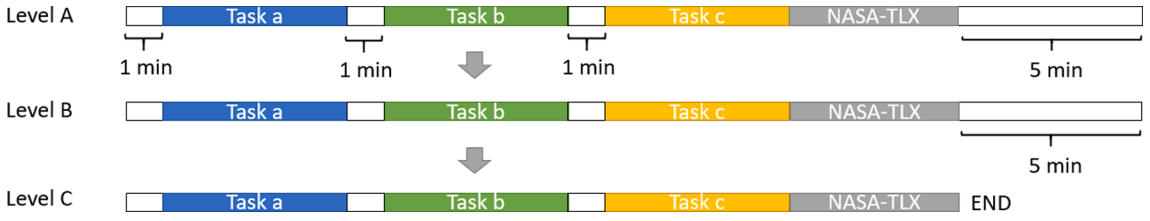


Fig. 5. Experimental process.

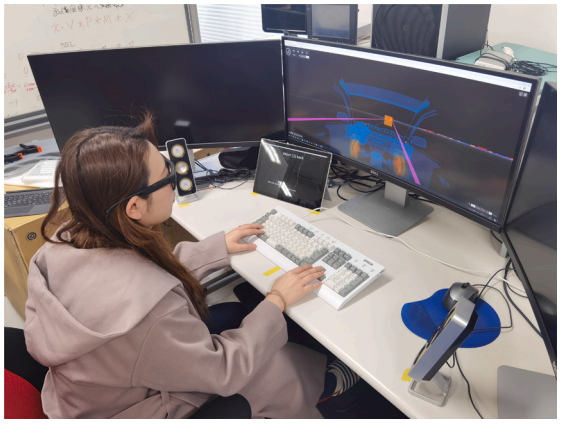


Fig. 6. Photo of experiment.

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log P(x_i)$$

where P means probability. Specifically, the entropy equations for 2D and 3D gaze entropies are as follows:

$$H_{2D}(X) = -\sum_{i=1}^{n} p(x,y)\log_2 p(x,y)$$

$$H_{3D}(X) = -\sum_{i=1}^{n} p(x,y,z)\log_3 p(x,y,z)$$

where p(x,y) and p(x,y,z) is the probability of the gaze falling on (x,y) or (x,y,z), respectively. The coordinates of the gaze point are those captured by the eye tracker. The 2D and 3D gaze entropies were calculated (in Python) using kernel-density estimation with a Gaussian kernel to estimate the probability.

5. Experiment procedure

We examined the validity of eye-tracking metrics as measures of MWL caused by visual and auditory NDRTs in semi-autonomous driving scenarios. The results can be regarded as a basis for a driver monitoring system that works before issuing a TOR. The experiment was designed as a driving simulation with three different secondary tasks. Each participant was required to perform the driving task and secondary task simultaneously.

The participants consented to and signed a participation consent form (Appendix a). After a brief introduction to the equipment and explanation of each secondary task (Appendix b), each participant wore an eye tracker (Tobii Pro Glasses 2). A participant had 10 min to get used to the simulation at all autonomous levels. Although the training time was not so long and the simulator was new to the participants, they were believed to access long-term memory about driving games or manual driving during the simulation. Without the long-term memory gathered from previous experience, the participants would not be able to successfully complete the driving task after 10-min training. The red static obstacles and orange dynamic obstacles reflect the driving scenario in the physical world, the predictions about the movement involves long-term memory. The eye tracker then started recording and the formal experiment began. The participants used oral responses"Yes!" when they thought a 2-back match appeared, while no motor response was required for the secondary tasks. The level sequence (Levels 0, 1, and 2) and secondary task sequence (visual-verbal, auditory-spatial, and auditory-verbal) were counter balanced to eliminate the effect of fatigue and learning effect on the experiment results. The participants were required to perform all three secondary tasks at one autonomous level then change to the next level after a 5-min break. Before

Table 1
G Power sample-size-calculation variables.

Effect size f	α err prob	Power	Numerator df	Number of groups	Total sample size
0.25	0.05	0.95	2	3	251

Table 2
ANOVA results of NASA-TLX in three secondary task.

Task Type	p-value	Effective size	Mean with confidence interval					
			L0		L1		L2	
Visual-Verbal	6.51E-07	0.23	72.42	[67.47, 77.37]	67.15	[62.19, 72.1]	52.82	[47.87, 57.78]
Auditory–Spatial	6.4E-07	0.23	72.16	[67.56, 76.75]	63.29	[58.7, 67.89]	53.34	[48.75, 57.94]
Auditory–Verbal	5.44E-05	0.17	65.49	[60.05, 70.93]	67.15	[52.55, 63.43]	52.82	[42.07, 52.95]

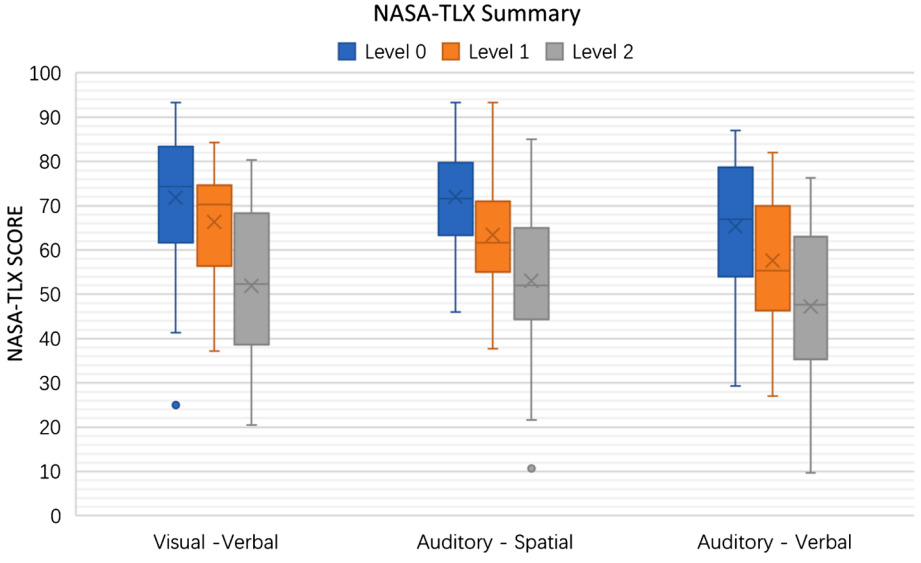


Fig. 7. NASA-TLX results.

the break, each participant was asked to stop the simulation and finish the three NASA-TLX surveys about the three secondary tasks at the level they just experienced on a tablet. Three NASA-TLX data sets were also collected before the break: one for Task a, one for Task b, and one for Task c.The experimental procedure is shown in Fig. 5 and a photo of one of the participants doing the experiment is shown in Fig. 6. The location of the tablet, which displayed the number in the visual-verbal task, is also shown.Table 1.

6. Results

6.1. H1 Examination: ANOVA

Because the physical workload of the experiment was manipulated at similar levels: the motor response was limited to keyboard operation, and the NASA-TLX was regarded as an indicator of MWL in the following part of this paper. To examine H1: different semi-autonomous levels place different MWL on the driver in both visual and auditory multi-tasking situations, a one-way repeated-measures analysis of variance (ANOVA) was applied to the NASA-TLX scores and secondary-task performance indicators. The difficulties of the tasks were determined by autonomous driving level: level 0 (L0) should be the hardest, while level 2 (L2) should impose the least MWL on the participants. The NASA-TLX score showed a significant main effect of autonomous levels on MWL ($\alpha = 0.05$, $p < 0.01$). Table 2 shows the statistical data of ANOVA for each of the three secondary tasks. The effective size was determined using Omaga-sq as an indicator. According to a rule of thumb of effective size (d 0.10 is a small effect, 0.25 is a medium effect, and 0.40 or more is a large effect), the visual-verbal and auditory- spatial tasks had a medium effect while the auditory-verbal task had a small effect. Fig. 7 illustrates the NASA-TLX scores of the three secondary tasks at three levels. Three indicators (Hit for target items, FA for lure items, and FA for other foils) for the secondary-task performance did not show any significant main effect of autonomous driving levels on MWL. Therefore, according to the subjective ratings, the H1 was proven to be correct.

Table 3
Means and standard deviations of all eye-tracking metrics.

Levels		Pupil Diameter Change			Saccade Number			Saccade Duration		
		L0	L1	L2	L0	L1	L2	L0	L1	L2
Visual-Verbal	M	0.11	0.01	−0.21	1.09×10^3	1.12×10^3	1.14×10^3	0.05	0.02	−0.01
	SD	0.10	0.08	0.11	1.33×10^5	1.24×10^5	0.93×10^5	0.00	0.02	0.01
Auditory-Spatial	M	0.11	−0.12	−0.28	6.65×10^2	6.90×10^2	7.61×10^2	0.01	−0.00	0.04
	SD	0.06	0.06	0.42	1.15×10^5	0.95×10^5	0.80×10^5	0.01	0.01	0.03
Auditory-Verbal	M	0.12	−0.11	−0.45	7.96×10^2	8.22×10^2	1.00×10^3	0.01	−0.01	0.03
	SD	0.08	0.09	0.41	1.58×10^5	1.47×10^5	1.55×10^5	0.01	0.00	0.02
		Fixation Duration			Fixation Number			Blink Number		
		L0	L1	L2	L0	L1	L2	L0	L1	L2
Visual-Verbal	M	0.06	−0.02	−0.03	7.15×10^2	7.17×10^2	7.04×10^2	1.89×10^2	2.17×10^2	2.77×10^2
	SD	0.04	0.01	0.03	2.77×10^4	2.10×10^4	2.06×10^4	1.78×10^4	2.18×10^4	3.31×10^4
Auditory-Spatial	M	0.06	−0.03	−0.18	3.87×10^2	3.98×10^2	4.41×10^2	1.73×10^2	1.92×10^2	2.36×10^2
	SD	0.02	0.01	0.06	3.46×10^4	2.42×10^4	1.73×10^4	1.33×10^4	1.87×10^4	1.82×10^4
Auditory-Verbal	M	0.05	−0.04	−0.23	4.65×10^2	4.89×10^2	5.77×10^2	2.01×10^2	2.46×10^2	2.87×10^2
	SD	0.01	0.01	0.03	3.88×10^4	3.96×10^4	2.72×10^4	2.00×10^4	2.77×10^4	2.53×10^4
		Blink Duration			2D Gaze Entropy			3D Gaze Entropy		
		L0	L1	L2	L0	L1	L2	L0	L1	L2
Visual-Verbal	M	−0.02	−0.04	0.10	4.55×10	4.50×10	4.48×10	6.56×10	6.58×10	6.68×10
	SD	0.05	0.03	0.07	4.80×10	8.20×10	6.10×10	1.80×10	1.50×10	4.20×10
Auditory-Spatial	M	−0.05	0.02	0.22	5.44×10	5.18×10	4.98×10	6.23×10	6.26×10	6.26×10
	SD	0.02	0.02	0.08	2.38×10^2	1.31×10^2	1.26×10^2	2.0×10	1.5×10	1.8×10
Auditory-Verbal	M	−0.04	0.02	0.03	5.64×10	5.37×10	5.22×10	6.40×10	6.41×10	6.38×10
	SD	0.02	0.02	0.06	1.70×10^2	1.42×10^2	1.33×10^2	1.0×10	2.5×10	9.6×10

6.2. H2 Examination: Data normalization

To remove the effects of individual differences from the experimental results, normalization was applied to the indicators with temporal variations: pupil diameter change, saccade duration, fixation duration, and blink duration. We selected z-score normalization because the pre-judgement about the data outliers is unclear:

$$x' = (x - \mu)/\sigma$$

The data of other indicators without temporal variations (NASA-TLX score, correct rate for secondary tasks, number of saccades, number of fixations, number of blinks, 2D gaze entropy, and 3D gaze entropy) were not normalized. The means and standard deviations (SDs) of all indicators in all nine situations are summarized in Table 3.

6.3. H2 Examination: MANOVA

A two-way MANOVA (autonomous level and secondary-task type as the independent variables) was applied to all nine eye-tracking metrics. The MANOVA results suggest a significant main effect ($p = 0.0002$ with effective size part eta-sq $= 0.065$, statistical power $1\text{-}\beta = 0.99$) of the eye metrics among the autonomous levels and secondary task- types. The effect size represents a medium effect, and the nine eye-tracking metrics had strong enough statistical power (99%) to reject the null hypothesis when considering H2 as the alternative hypothesis.

The results are not surprising because whether the participant processed the visual information affected all the eye-tracking metrics. The difference between the visual and auditory tasks is directly reflected in the eye-related indicators. However, nine eye-tracking metrics with 99% statistical power to predict MWL is unnecessarily accurate. Relatively high enough accuracy with less

Table 4
Correlation matrix (significant correlations from zero in red).

Task Types		CR	PD	Sac*n*	Sac_d	Fix*n*	Fix_d	B*n*	B_d	En_{2D}	En_{3D}
Visual	NALX	**−0.25**	**0.25**	0.02	0.14	0.06	−0.08	−0.08	−0.17	−0.06	**−0.31**
	p	**0.01**	**0.01**	0.83	0.18	0.57	0.45	0.43	0.10	0.56	**0.000**
Auditory	NALX	**−0.33**	**0.24**	0.01	0.04	0.01	**0.22**	−0.08	−0.07	**−0.14**	**−0.22**
	p	**0.00**	**0.00**	0.89	0.58	0.84	**0.00**	0.23	0.30	**0.04**	**0.01**
All	NALX	**−0.30**	**0.25**	0.05	0.07	0.08	**0.15**	−0.08	−0.11	**−0.14**	**−0.13**
	p	**0.00**	**0.00**	0.38	0.24	0.18	**0.01**	0.14	0.053	**0.01**	**0.02**

Table 5
Validity of eye-tracking metrics.

Eye-tracking Metric	Visual Task	Auditory Tasks	All Tasks
Pupil Diameter Change	+	+	+
Number of Saccades			
Saccade Duration			
Number of Fixations			
Fixation Duration		+	+
Number of Blinks			
Blink Duration			
2D Gaze Entropy		–	–
3D Gaze Entropy	–	–	–

metrics is preferred in human–machine systems. Therefore, effective metrics extraction should be conducted to find the most representative eye-tracking metrics with decent accuracy to predict the MWL in both visual and auditory multi-tasking situations.

6.4. H2 Examination: Metrics extraction

To extract the effective eye-tracking metrics in visual and auditory multi-tasking situations, two analysis methods were used: correlation calculation and principal-component extraction. Pearson's linear matrix was first calculated for all indicators (correct rate plus all nine eye-tracking metrics) for the visual task, auditory tasks, and all three tasks (Table 4). The NASA-TLX score was regarded as the indicator for MWL.

As for the auditory tasks (the combination of the data from auditory-spatial and auditory-verbal tasks), secondary-task performance, pupil diameter change, which were significant indicators in the visual tasks, fixation duration, and 2D gaze entropy significantly correlated with MWL.

When all the data from the three tasks were used to calculate the matrix, secondary- task performance, pupil diameter change, fixation duration, 2D gaze entropy, and 3D gaze entropy showed significant correlations, hence could be considered effective indicators of MWL.

Comparing the correlation matrixes of the visual and auditory tasks, we found that the number of saccades and number of fixations tended to be more correlated with each other when auditory tasks were performed. If no visual modality is used, the eye-tracking metrics seem to be more constant by reducing the intended attention switch.

The significant indicators for the visual task, auditory tasks, and all tasks are listed in Table 5. Positive correlations are marked with"+" and negative correlations are marked with"-".

As we can see from Table 5, pupil diameter change and 3D gaze entropy were universally valid indicators of MWL in the visual task and auditory tasks. Pupil diameter change was significantly positively correlated with the NASA-TLX, and 3D gaze entropy was significantly negatively correlated with the NASA-TLX. The Peason's co– efficient (Fisher) of pupil diameter change had a p-value of 1.23E-05 with confidence interval [0.1375, 0.348] and the Peason's coefficient (Fisher) of 3D gaze entropy had a p-value of 0.0148 with confidence interval [-0.247, −0.0274].

As MWL increased, pupil diameter change increased; this phenomenon has also been observed in other studies (Marinescu, Sharples, & Campbell Ritchie, 2018; Matton et al., 2020; Wu et al., 2020). Three-dimensional gaze entropy decreased as MWL increased, which is contrary to Wu et al.'s results (Wu et al., 2020), in which they concluded that gaze entropy is positively correlated with perceived workload in robotic surgical-skills training. Our experiment was a driving simulation in a lab, whereas they focused on surgical-skills training; the different application scenarios might be the reason for the opposite results. The object with our experiment was multi-tasking rather than only one task. The attention switching tended to occur often at L0, at which MWL should be the highest. As the autonomous driving level increased, participants did not necessarily switch their attention so often, so their gaze in that situation had less entropy compared with a situation with frequent view changes.

We found that the fixation duration for the auditory task was significantly positively correlated with MWL, while 3D gaze entropy increased as MWL decreased.

As well as the correlation-matrix calculation, principal-component extraction was applied in the data-analysis process. Because we had nine eye-tracking metrics, con– sidering the factor-analysis model, our preference was to use fewer than (9–1)/2 = 4 factors if

Table 6
Eigenvalues of nine eye- tracking metrics.

	eValue	%	Cum%
1	2.8546	31.72%	31.72%
2	1.5552	17.28%	49.00%
3	1.1106	12.34%	61.34%
4	1.0589	11.77%	73.10%
5	0.8411	9.35%	82.45%
6	0.5986	6.65%	89.10%
7	0.5175	5.75%	94.85%
8	0.3859	4.29%	99.14%
9	0.0776	0.86%	100.00%

Table 7
Factor-score matrix - Bartlett's method.

	PC1	PC2	PC3	PC4
Pupil Diameter Change	0.000244	−0.34831	−0.01672	−0.29792
Number of Saccades	**0.470877**	−0.03163	−0.02045	−0.03235
Saccade Duration	0.009135	0.178882	0.008633	**−0.92664**
Number of Fixations	0.399317	−0.10206	0.134683	0.000165
Fixation Duration	0.0059	**−0.61572**	−0.02799	−0.01858
Number of Blinks	0.122223	0.124212	−0.07703	0.050523
Blink Duration	−0.00467	0.359904	0.001844	−0.12388
2D Gaze Entropy	−0.16914	0.037015	0.086876	0.032428
3D Gaze Entropy	−0.19188	0.066112	**0.963636**	0.005812

Table 8
Rough KNN confusion matrix.

	Predicted High	Class Medium	Low	TPR	FNR
Actual Class					
High	100.0%			100.0%	
Medium	19.0%	75.0%	6.0%	75.0%	25.0%
Low			100.0%	100.0%	

possible. The top four eigenvalues accounted for 73.1% of the variance. The eigenvalues in decreasing order are listed in Table 6 and show the percentage of the total variance accounted for by that eigenvalue.

By using Bartlett's method, the number of saccade, saccade duration, fixation duration, and 3D gaze entropy were found to be the four variables highly correlated (a threshold of 0.4) with only one principal component. The factor-score matrix is shown in Table 7.

Combing the results of the correlation-matrix calculation and principal-component extraction, pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy were the five extracted effective eye-tracking metrics.

6.5. H2 Examination: Workload classification

To examine whether the extracted five eye-tracking metrics were efficient enough to predict MWL, the K-Nearest Neighbor (KNN) classifier was used. Because of the small sample size, the data were first bootstrapped (mean-based) to 5000 datasets (NASA-TLX, pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy) in one autonomous driving level. Based on the original NASA-TLX distribution, the dataset were labeled as'High mental workload' with a NASA-TLX score over 69;'Low mental workload' with a NASA-TLX lower than 53; and the rest were labeled as'Medium mental workload'. Using the KNN classifier, five features were included to classify the MWL level: pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy. The average classification accuracy was 88.9%. The confusion matrix for the rough KNN classifier (number of neighbors: 100; distance metrics: Euclidean; distance weight: equal) is presented in Table 8.

Through the classifier examination, we concluded that H2 is true. Eye-tracking metrics are effective indicators for MWL-level prediction in both visual and auditory multi-tasking situations. The five eye-tracking metrics (pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy) are efficient enough to predict the MWL level.

7. Discussion

7.1. NDRT types and mental workload

The difficulty of multi-tasking is controlled by the autonomous driving level because all three secondary tasks are essentially the

same: a 2-back task with an oral response. Level 0 is supposed to be the most difficult while L2 should impose the least MWL on the driver. The one-way ANOVA indicated that the NASA-TLX showed a significant effect of the three autonomous levels in all three secondary tasks. Despite the fact that vision is the most important information resource during driving and the visual cognitive resource can be saved with the increase of autonomous driving level, our study showed that the auditory cognitive channel has the same phenomenon. The visual multi-tasking situation before TOR should be considered when designing the interface for TOR as well as the auditory multi-tasking situation.

7.2. Eye-tracking metrics and mental workload

Pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy were proven to be five effective eye-tracking metrics for predicting the MWL level. Although the exact prediction model remains unclear, it is crucial to know that it is possible to predict MWL through eye-tracking metrics. With computer vision technology, the in-vehicle driver monitoring system, normally a camera, has the ability to not only monitor driver behavior but also driver MWL in multi-tasking situations.

7.3. Limitations

There were several limitations to this study. The first limitation is the effect of fatigue. The experiment took a relatively long time to finish: 20 mins for each autonomous driving level, 5 mins for answering the NASA-TLX survey for each level, plus the time for simulator practice and rests between levels; the total time for the experiment was about 95 mins. Furthermore, the multi-tasking design was quite difficult and tired the participants very quickly. Although the effect of fatigue on the level order and task order was removed through counter balancing, the long experimental time might not be reflective of normal situations.

The second limitation is the simplified driving simulator. The driving simulator in the experiment involved using a keyboard instead of a steering wheel and pedal to control the simulated vehicle. The motor workload was lower than a more realistic simulator and actual driving. Future experiments are planned with a more realistic driving simulator for more convincing results.

The third limitation is the short training time. The training time in the experiment was 10 min, which seems to be relatively short. Although the simulation operation is quite simple and after 10-min training all participants reported as being confident, longer training time will be beneficial to remove the learning effect on the results. This study used counter balancing to remove the learning effect, but a longer training time can make the learning effect less impactful.

The fourth limitation is individual differences in spatial-detection ability. The auditory-spatial task is a 2-back task about the location of a sound source: front left, front, and front right. Several participants gave feedback that it was difficult for them to distinguish the locations. Such individual difference was not considered in the experimental design phase, and no pilot experiment on personal spatial-detection ability was designed. Individual differences may affect MWL in terms of the subjective NASA-TLX score or eye-tracking metric.

The fifth limitation is that the auditory-spatial task was reported to be more difficult than the other two tasks. Although the three secondary tasks were designed to have similar difficulty by using the same type of cognitive task, the auditory-spatial task might have essentially imposed a higher MWL on participants who do not use much spatial cognition in daily life. All the participants were university students; thus, we assumed that they would be more familiar with numbers than with making judgements about directions.

The sixth limitation is the representativeness of the secondary tasks in the experiment. The secondary tasks were designed in consideration of the mental resources in a uni-cognitive channel. In real situations, NDRTs cover multiple cognitive channels; even NDRTs with only one perception modality, such as reading the news on a smart phone, usually involves multiple cognitive processing codes (visual, spatial, and verbal). Thus, whether the simulation results on the visual-verbal task can be generalized to all visual tasks may need further discussion. However, the auditory-task results are plausible enough because the not tested auditory-visual channel is rarely used in driving situations. One example of an auditory-verbal cognitive channel is playing audio of the names of different animals (e.g. dog and elephant) and asking the participant which animal has a larger body.

The seventh limitation is the 3D-gaze point measurement. In the driving simulation, all driving related information was visually presented on a display, so if the participants kept their eyes on the real-world coordinate system, the view depth would not change so much. Although the participants were free to move their heads and look at other places than the display in the experiment, the depth distribution differs from a real driving situation, which shows a much smaller variation.

The eighth limitation is the different strategy in the auditory-spatial secondary task. When the participant heard the beep and judged its location, he or she may have changed from spatial cognition to verbal cognition for memorizing the sequence of locations. Some may have memorized the locations by using spatial imagination but others may have used a verbal sequence. This all depends on the participants' strategies for completing the secondary task, and no instructions about how specifically the participants should conduct the experiment was given beforehand. Because the dual-tasking situation is quite difficult, there is still a possibility that the participants used the cognitive channel that they felt comfortable with (e.g. engineering students may tend to memorize a sequence of numbers while drama or design students may prefer imagining a spatial arrangement).

7.4. Application

Our experiments proved that the eye-tracking metrics of pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy are valid indicators for both visual and auditory tasks in semi-autonomous driving. This finding can be applied to a driver monitoring system for semi-autonomous vehicles. The current driver monitoring systems use computer-vision

algorithms that analyze camera images in real time. This means that eye-tracking metrics are also available in real time and can be used for estimating MWL in real time. NDRTs can also be recorded with a camera and can be classified using machine-learning technology. Therefore, the NDRT type determines the MWL allocation and the eye-tracking metrics determine the MWL degree in that MWL-allocation frame. Supposing that a driver's MWL is continuously monitored; when the autonomous driving vehicle faces an emergency or traffic situation beyond its capacity, it can issue a TOR in accordance with the driver's MWL allocation. By using an unoccupied cognitive channel, the driver will be spared from too high an MWL that might delay his/her response and increase the possibility of poor judgement.

Note that this application needs more attention concerning the limitations of the experiment. The auditory-spatial task was evaluated as more difficult and showed large individual differences, which indicates that spatial cognition should have a lower weight in a TOR alert even when MWL is reported to be lower in an auditory-spatial task than in a visual-verbal task. In an emergency, the same information presented in a spatial cognitive channel might cause a longer reaction time. For example, a"Left!" auditory instruction might have a swifter effect than a seat vibration on the left side when the direction of an on-coming vehicle has to be indicated. However, tactile interfaces have been reported to be efficient for TORs. Thus, further experiments involving the tactile cognitive channel should be conducted.

8. Conclusion

We examined the validity of eye-tracking metrics to indicate the mental workload of visual and auditory tasks in semi-autonomous driving situations. We found that when the NASA-TLX was used as an indicator to represent mental workload, pupil diameter change, number of saccades, saccade duration, number of fixations, and 3D gaze entropy were efficient indicators of mental workload. These findings can be applied to camera in-vehicle driver monitoring systems to predict the driver's mental-workload in real time, which would be beneficial for adaptive multimodal interface designs for takeover requests.

CRediT authorship contribution statement

Weiya Chen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Tetsuo Sawaragi:** Conceptualization, Validation, Writing – review & editing, Supervision. **Toshihiro Hiraoka:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Consent to participate in a research study

Introduction

- You are being asked to be in a research study on driver assistance technology. Please read this form and ask any questions that you may have before agreeing to be in the study.

Purpose of Study

- The purpose of the study is to determine how driver's mental workload concerning different cognitive channels changes depending on the level of autonomous driving.
- Ultimately, this research may be presented as a journal article.

Description of the Study Procedures

- If you agree to be in this study, you will be asked to do the following things: Wear an eye tracker when doing the driving simulation. Finish various secondary task during simulations of different levels of autonomous driving. Fill in NASA- TLX scales using the NASA-TLX APP after each secondary task.
- Before the experiment, you can have a 5-min test drive to become familiar with the simulator (all autonomous levels included) and experience the three secondary tasks.
- Three secondary tasks (2-back task, 40 * 3 trials for each task) need to be conducted:
  - Visual modality – Verbal cognition: 2-back task with presentation of numbers on a tablet
  - Auditory modality – Spatial cognition: 2-back task on speaker locations Auditory modality – Verbal cognition: 2-back task with audio presentation of numbers
- Experimental schedule:
- Throughout the whole process, views of the participant will be recorded by a camera on the eye tracker. (See Table A1)

Table A1
Validity of eye-tracking metrics.

	Estimated time
Explanation of the experiment (including explanation of NASA-TLX)	5 mins
Driving simulator practice (all levels)	5 mins
Eye tracker adjustment	5 mins
Task a: trial	5 mins
Task a: simulation (all levels)	15 mins
Task a: NASA-TLX scoring	5 mins
Break	5 mins
Task b: trial	5 mins
Task b: simulation (all levels)	15 mins
Task b: NASA-TLX scoring	5 mins
Break	5 mins
Task c: trial	5 mins
Task c: simulation (all levels)	15 mins
Task c: NASA-TLX scoring	5 mins
In total	95 mins

Risks/Discomforts of Being in this Study

- There are no reasonable foreseeable (or expected) risks. There may be unknown risks.

Confidentiality

- Your identity might be disclosed in the material that is published. However, you will be given the opportunity to review and approve any material that is published about you.

Right to Refuse or Withdraw

- The decision to participate in this study is entirely up to you. You may refuse to take part in the study at any time without affecting your relationship with the investigator. Your decision will not result in any loss or benefits to which you are otherwise entitled. You have the right not to answer any single question, as well as to withdraw completely from the interview at any point during the process; additionally, you have the right to request that the interviewer not use any of your interview material.

Right to Ask Questions and Report Concerns

- You have the right to ask questions about this research study and to have those questions answered by me before, during or after the research. If you have any further questions about the study, at any time feel free to contact me, Weiya Chen at weiyachen15@gmail.com. If you like, a summary of the results of the study will be sent to you.

Consent

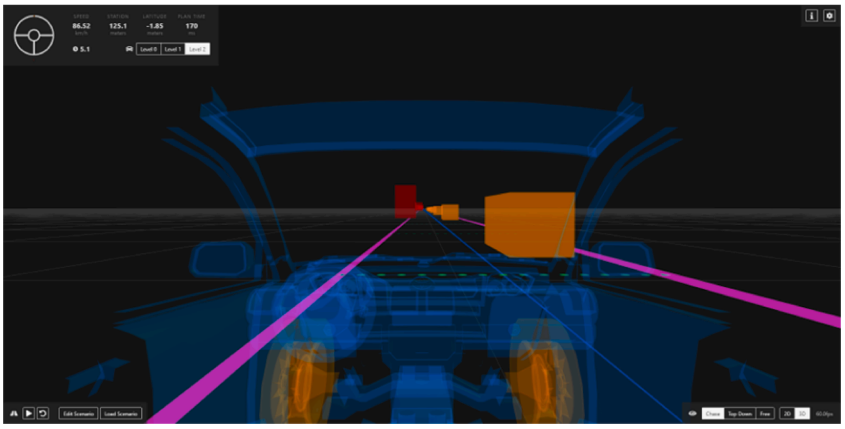
- Your signature below indicates that you have decided to volunteer as a research participant for this study and that you have read and understood the information provided above. You will be given a signed and dated copy of this form to
- keep, along with any other printed materials deemed necessary by the study investigators.

Subject's Name (print): _____
Subject's Signature: _____ Date: _____
Investigator's Signature: Weiya Chen Date: _____

Appendix B. . Instruction for participants in a research study

Driving Simulator

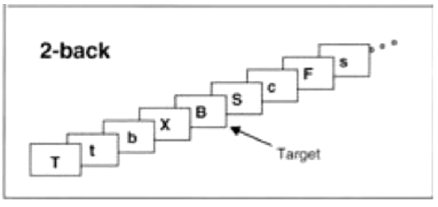
- The driving simulator is modified from a real-time, on-road, lattice-based autonomous vehicle motion planner in the browser.
- The simulator uses 'a' and 'd' keys for steering control; pressing 'a' means turning left and pressing 'd' means turning right. Pressing the 'w' key activates the accelerator and the 's' key means reversing gear. The 'space' key is the brake. The basic interface is shown in the following diagram: red blocks represent static obstacles, while orange blocks represent dynamic obstacles:



- The experiment will be conducted at three autonomous driving levels: level 0, level 1 and level 2. As shown in the left upper corner, these three levels can be chosen by mouse click.
  - o Level 0: the participant controls both speed and steering.
  - o Level 1: the participant controls the steering, and the simulator controls the vehicle's speed. But in emergency situations (e.g. the speed control is too slow and the vehicle seems collide with the front car), the participant
  - o needs to change the autonomous driving level to level 0. After avoiding the crisis, the autonomous driving level should be changed back to level 1.
  - o Level 2: Both speed and steering are controlled by the automatic system. But the participant is required to monitor the driving environment, and when necessary (e.g. the car stops in front of a static obstacle), the participant needs to change the autonomous driving level to level 0 and drive the car to a suitable location and reengage level 2.
- During the driving simulation, safety is the primary principle; please try to avoid crashes and stay in the lane.

Secondary Tasks

- 2-back tasks:



- Visual modality – Verbal cognition: 2-back task with numbers presented on a tablet
- Auditory modality – Spatial cognition: 2-back task on speaker locations Auditory modality – Verbal cognition: 2-back task with audio presentation of numbers

NASA-TLX

- The NASA Task Load Index (NASA-TLX) is a widely used, subjective, multi-dimensional assessment tool that rates perceived workload in order to assess a task, system, or team's effectiveness or other aspects of performance.
- NASA-TLX has six subjective subscales:
  - o Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
  - o Physical Demand: How much physical activity was required? Was the task easy or demanding, slack or strenuous?
  - o Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid? Performance: How successful were you in performing the task? How satisfied were you with your performance?
  - o Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Table B1
G Power sample-size-calculation variables.

Effect size f	α err prob	Power	Numerator df	Number of groups	Total sample size
0.25	0.05	0.95	2	3	251

Table B2
ANOVA results of NASA-TLX in three secondary task.

Task Type	p-value	Effective size	Mean with confidence interval					
			L0		L1		L2	
Visual-Verbal	6.51E-07	0.23	72.42	[67.47, 77.37]	67.15	[62.19, 72.1]	52.82	[47.87, 57.78]
Auditory–Spatial	6.4E-07	0.23	72.16	[67.56, 76.75]	63.29	[58.7, 67.89]	53.34	[48.75, 57.94]
Auditory–Verbal	5.44E-05	0.17	65.49	[60.05, 70.93]	67.15	[52.55, 63.43]	52.82	[42.07, 52.95]

Table B3
Means and standard deviations of all eye-tracking metrics.

Levels		Pupil Diameter Change			Saccade Number			Saccade Duration		
		L0	L1	L2	L0	L1	L2	L0	L1	L2
Visual-Verbal	M	0.11	0.01	−0.21	1.09×10^3	1.12×10^3	1.14×10^3	0.05	0.02	−0.01
	SD	0.10	0.08	0.11	1.33×10^5	1.24×10^5	0.93×10^5	0.00	0.02	0.01
Auditory-Spatial	M	0.11	−0.12	−0.28	6.65×10^2	6.90×10^2	7.61×10^2	0.01	−0.00	0.04
	SD	0.06	0.06	0.42	1.15×10^5	0.95×10^5	0.80×10^5	0.01	0.01	0.03
Auditory-Verbal	M	0.12	−0.11	−0.45	7.96×10^2	8.22×10^2	1.00×10^3	0.01	−0.01	0.03
	SD	0.08	0.09	0.41	1.58×10^5	1.47×10^5	1.55×10^5	0.01	0.00	0.02
		Fixation Duration			Fixation Number			Blink Number		
		L0	L1	L2	L0	L1	L2	L0	L1	L2
Visual-Verbal	M	0.06	−0.02	−0.03	7.15×10^2	7.17×10^2	7.04×10^2	1.89×10^2	2.17×10^2	2.77×10^2
	SD	0.04	0.01	0.03	2.77×10^4	2.10×10^4	2.06×10^4	1.78×10^4	2.18×10^4	3.31×10^4
Auditory-Spatial	M	0.06	−0.03	−0.18	3.87×10^2	3.98×10^2	4.41×10^2	1.73×10^2	1.92×10^2	2.36×10^2
	SD	0.02	0.01	0.06	3.46×10^4	2.42×10^4	1.73×10^4	1.33×10^4	1.87×10^4	1.82×10^4
Auditory-Verbal	M	0.05	−0.04	−0.23	4.65×10^2	4.89×10^2	5.77×10^2	2.01×10^2	2.46×10^2	2.87×10^2
	SD	0.01	0.01	0.03	3.88×10^4	3.96×10^4	2.72×10^4	2.00×10^4	2.77×10^4	2.53×10^4
		Blink Duration			2D Gaze Entropy			3D Gaze Entropy		
		L0	L1	L2	L0	L1	L2	L0	L1	L2
Visual-Verbal	M	−0.02	−0.04	0.10	4.55×10	4.50×10	4.48×10	6.56×10	6.58×10	6.68×10
	SD	0.05	0.03	0.07	4.80×10	8.20×10	6.10×10	1.80×10	1.50×10	4.20×10
Auditory-Spatial	M	−0.05	0.02	0.22	5.44×10	5.18×10	4.98×10	6.23×10	6.26×10	6.26×10
	SD	0.02	0.02	0.08	2.38×10^2	1.31×10^2	1.26×10^2	2.0×10	1.5×10	1.8×10
Auditory-Verbal	M	−0.04	0.02	0.03	5.64×10	5.37×10	5.22×10	6.40×10	6.41×10	6.38×10
	SD	0.02	0.02	0.06	1.70×10^2	1.42×10^2	1.33×10^2	1.0×10	2.5×10	9.6×10

- o Frustration: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?
- First, 15 pairwise comparisons between these six subscales will be made based on the influence on the general mental workload of a task.
- Then, the participant rates each task within a 100-point range with 5-point steps.

Table B4
Correlation matrix (significant correlations from zero in red).

Task Types		CR	PD	Sac*n*	Sac_d	Fix*n*	Fix_d	B*n*	B_d	En_{2D}	En_{3D}
Visual	NALX	**−0.25**	**0.25**	0.02	0.14	0.06	−0.08	−0.08	−0.17	−0.06	**−0.31**
	p	**0.01**	**0.01**	0.83	0.18	0.57	0.45	0.43	0.10	0.56	**0.000**
Auditory	NALX	**−0.33**	**0.24**	0.01	0.04	0.01	**0.22**	−0.08	−0.07	**−0.14**	**−0.22**
	p	**0.00**	**0.00**	0.89	0.58	0.84	**0.00**	0.23	0.30	**0.04**	**0.01**
All	NALX	**−0.30**	**0.25**	0.05	0.07	0.08	**0.15**	−0.08	−0.11	**−0.14**	**−0.13**
	p	**0.00**	**0.00**	0.38	0.24	0.18	**0.01**	0.14	0.053	**0.01**	**0.02**

Table B5
Validity of eye-tracking metrics.

Eye-tracking Metric	Visual Task	Auditory Tasks	All Tasks
Pupil Diameter Change	+	+	+
Number of Saccades			
Saccade Duration			
Number of Fixations			
Fixation Duration		+	+
Number of Blinks			
Blink Duration			
2D Gaze Entropy		–	–
3D Gaze Entropy	–	–	–

Table B6
Eigenvalues of nine eye- tracking metrics.

	eValue	%	Cum%
1	2.8546	31.72%	31.72%
2	1.5552	17.28%	49.00%
3	1.1106	12.34%	61.34%
4	1.0589	11.77%	73.10%
5	0.8411	9.35%	82.45%
6	0.5986	6.65%	89.10%
7	0.5175	5.75%	94.85%
8	0.3859	4.29%	99.14%
9	0.0776	0.86%	100.00%

Table B7
Factor-score matrix - Bartlett's method.

	PC1	PC2	PC3	PC4
Pupil Diameter Change	0.000244	−0.34831	−0.01672	−0.29792
Number of Saccades	**0.470877**	−0.03163	−0.02045	−0.03235
Saccade Duration	0.009135	0.178882	0.008633	**−0.92664**
Number of Fixations	0.399317	−0.10206	0.134683	0.000165
Fixation Duration	0.0059	**−0.61572**	−0.02799	−0.01858
Number of Blinks	0.122223	0.124212	−0.07703	0.050523
Blink Duration	−0.00467	0.359904	0.001844	−0.12388
2D Gaze Entropy	−0.16914	0.037015	0.086876	0.032428
3D Gaze Entropy	−0.19188	0.066112	**0.963636**	0.005812

Others

- Please switch your mobile phone to silent mode and put it on a desk away from the experiment desk.
- If you have any questions, please ask the investigator.

See Tables B1-B8 and Figs. B1-B7.

Table B8
Rough KNN confusion matrix.

	Predicted High	Class Medium	Low	TPR	FNR
Actual Class					
High	100.0%			100.0%	
Medium	19.0%	6.0%	75.0%	75.0%	25.0%
Low	100.0%			100.0%	

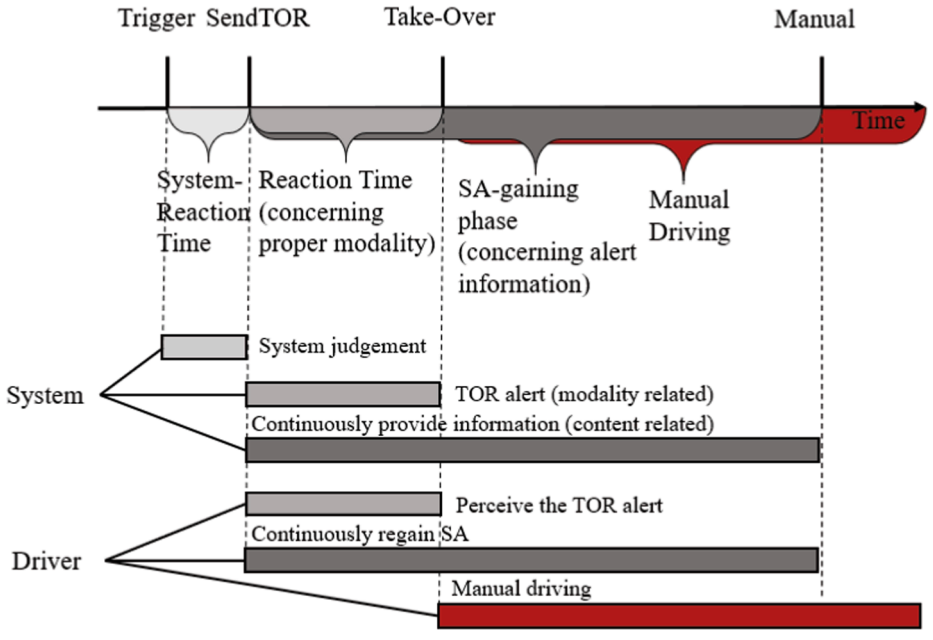


Fig. B1. TOR process as time sequence.

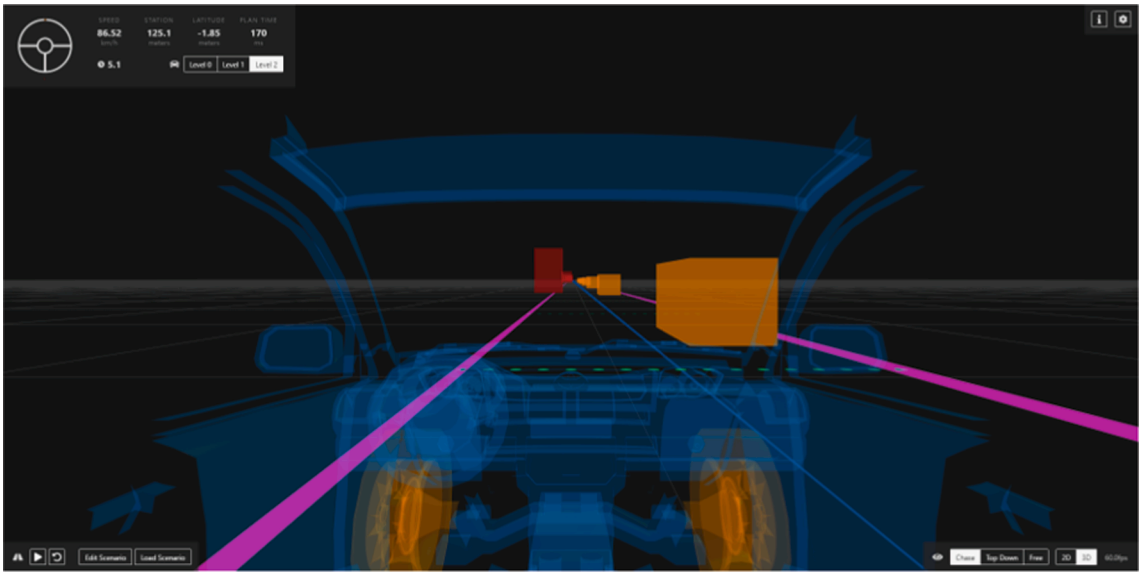


Fig. B2. Simulation display.

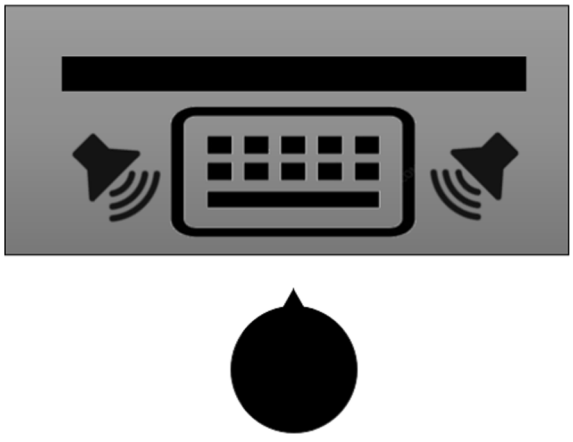


Fig. B3. Simulation layout.

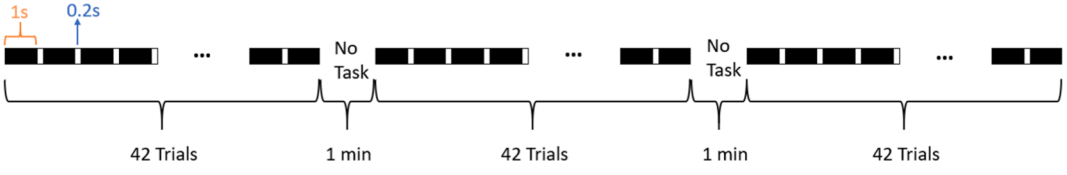


Fig. B4. Secondary-task time sequence.

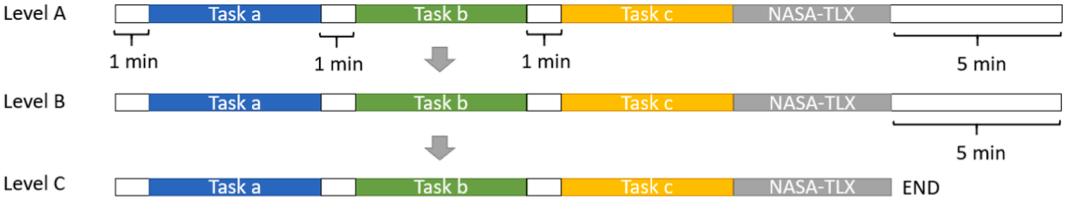


Fig. B5. Experimental process.

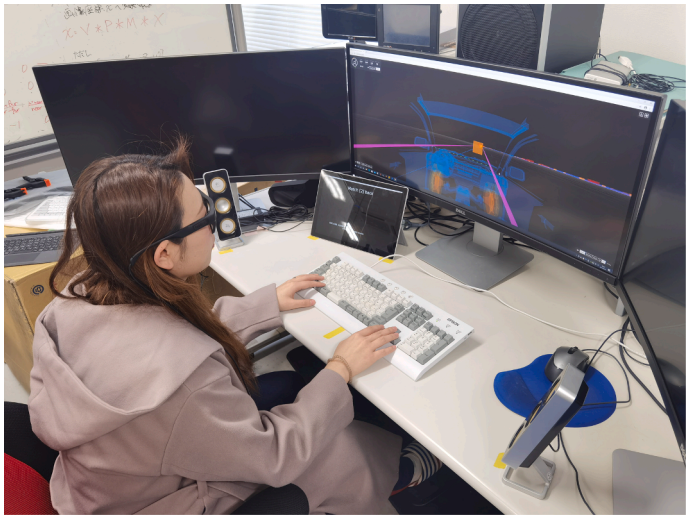


Fig. B6. Photo of experiment.

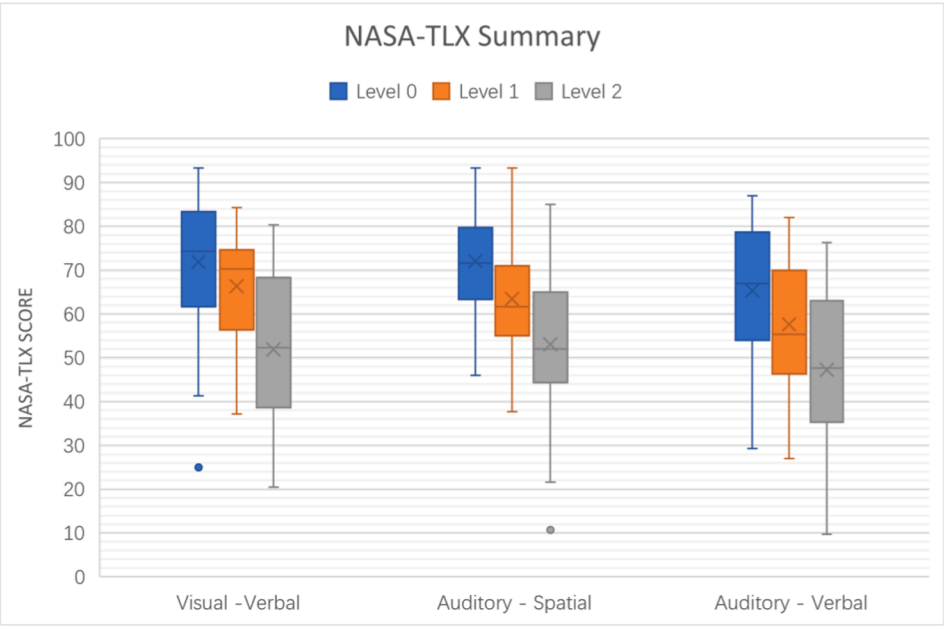


Fig. B7. NASA-TLX results.

References

Bazilinskyy, P., Petermeijer, S. M., Petrovych, V., Dodou, D., & de Winter, J. C. F. (2018). Take- over requests in highly automated driving: A crowdsourcing survey on auditory, vibrotactile, and visual displays. *Transportation Research Part F, 56*(2018), 82–98.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2010). Driver workload and eye blink duration. *Transportation Research, Part F, 14*(2011), 199–208.

Biswas, P., & Prabhakar, G. (2018). Detecting drivers' cognitive load from saccadic intrusion. *Transportation Research Part F, 54*(2018), 63–78.

Cabinet Office, Government of Japan. *Society 5.0*. Retrieved from https://www8.cao.go.jp/cstp/society5 0.

Chen, W., Sawaragi, T., & Hiraoka, T. (2021). Adaptive Multi-Modal Interface Model Concerning Mental Workload in Take-over Request during Semi-Autonomous Driving. *SICE Journal of Control, Measurement, and System Integration, 14*(2021), 10–21.

Chihara, T., Kobayashi, F., & Sakamoto, J. (2020). Estimation of mental workload during automobile driving based on eye-movement measurement with a visible light camera. *Transactions of the JSME (in Japanese), 86*(881), 2020.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses, *ehavior Research Methods*, *41*, 1149–1160.

Faure, V., Lobjois, R., & Benguigui, N. (2016). The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F, 40*(2016), 78–90.

Fridman, L., Reimer, B., Mehler, B., Freeman, W. T. (2018). *Cognitive Load Estimation in the Wild*, CHI 2018, April 21–26, 2018, Montŕeal, QC, Canada.

Geitner, C., Biondi, F., Skrypchuk, L., Jennings, P., & Birrell, S. (2019). The comparison of auditory, tactile, and multimodal warnings for the effective communication of unexpected events during an automated driving scenario. *Transportation Research Part F, 65*(2019), 23–33.

Hart, S. G., Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *Advances in Psychology*, *52*, 139-–183.

Huang, L., Bian, Y., Zhao, X., Xu, Y., & Rong, J. (2019). Quantitatively exploring the relationship between eye movement and driving behavior under the effect of different complex diagrammatic guide signs. *Cognition, Technology Work, 22*(2020), 373–388.

Huang, G., Steele, C., Zhang, X., Pitts, B. J. (2019). Multimodal Cue Combinations: A Possible Approach to Designing In-Vehicle Takeover Requests for Semi-autonomous Driving, *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting*, 1739– 1743.

Jeong, H., & Liu, Y. (2018). Effects of non-driving-related-task modality and road geometry on eye movements, lane-keeping performance, and workload while driving. *Transportation Research Part F, 60*(2019), 157–171.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2017). Working Memory, Attention Control, and the N-Back Task: A Question of Construct Validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615–622.

Kalb, L., Streit, L., Bengler, K. (2018). Multimodal Priming of Drivers for a Cooperative Take- Over, *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, p. 1029-1034, Maui, Hawaii, USA, November 4-7, 2018.

Kosch, T., Hassib, M., Wózniak, P. W., Buschek, D., Alt, F. (2018). Your Eyes Tell: Lever- aging Smooth Pursuit for Assessing Cognitive Workload, *CHI 2018*, April 21–26, 2018, Montreal, QC, Canada.

Levulis, S. J., DeLucia, P. R., Kim, S. Y. (2018). Effects of Touch, Voice, and Multimodal Input, and Task Load on Multiple-UAV Monitoring Performance During Simulated Manned- Unmanned Teaming in a Military Helicopter, *Human Factors*, *60* (8), 2018, 1117—1129.

Li, X., Schroeter, R., Rakotonirainy, A., Kuo, J., & Lenńe, M. G. (2020). Effects of different non-driving-related-task display modes on drivers' eye-movement patterns during take-over in an automated vehicle. *Transportation Research Part F, 70*(2020), 135–148.

Longo. L. (2015). Designing Medical Interactive Systems Via Assessment of Human Mental Workload, *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, 364-–365.

Marinescu, A. C., Sharples, S., Campbell Ritchie, A., Śanchez Ĺopez, T., McDowell, M., Morvan, H. P. (2018). Physiological Parameter Response to Variation of Mental Workload *Human Factors*, *60* (1):31–56.

Matton, N., Paubel, P-V., Puma, S. (2020). Toward the Use of Pupillary Responses for Pilot Selection, *Human Factors*, August 2020. doi:10.1177/0018720820945163.

MIT License, https://github.com/mattbradley/dash.

Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015). Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye-movement measures in the assessment of cognitive workload. *Transportation Research Part F, 32*(2015), 23–34.

Riggs, S. L., Sarter, N. (2019). Tactile, Visual, and Crossmodal Visual-Tactile Change Blindness: The Effect of Transient Type and Task Demands, *Human Factors*, *61* (1), February 2019, 5—24.

Shannon, C. E. *A mathematical theory of communication*. ACM SIGMOBILE Mobile Computing and Communications Review, 5, 3—55.

Sklar, A. E., Sarter, N. B. (1999). Good Vibrations: Tactile Feedback in Support of Attention Allocation and Human-Automation Coordination in Event-Driven Domains, *Human Factors*, *41* (4), December 1999, 543—552.

Tang, Q., Guo, G., Zhang, Z., Zhang, B., Wu, Y. (2020). Olfactory Facilitation of Takeover Performance in Highly Automated Driving, *Human Factors*, January 2020. doi:10.1177/0018720819893137.

Wang, Y., Toor, S. S., Gautam, R., & Henson, D. B. (2011). Blink Frequency and Duration during Perimetry and Their Relationship to Test-Retest Threshold Variability. *Investigative Ophthalmology Visual Science, 52*, 4546–4550.

Wang, Y., Reimer, B., Dobres, J., & Mehler, B. (2014). The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. *Transportation Research Part F, 26*(2014), 227–237.

Wilkie, R., Mole, C., Giles, O, Merat, N., Romano. R., Markkula, G. (2018). Cognitive Load during Automation Affects Gaze Behaviours and Transitions to Manual Steering Control, *The Proceedings of the 10th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 24-27 Jun 2019, Santa Fe, New Mexico.

Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., Yu, D. (2020). Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training, *Human Factors*, *62* (8), December 2020, 1365–1386.

Yang, S., Kuo, J., Lenńe, M. G. (2020). Effects of Distraction in On-Road Level 2 Automated Driving: Impacts on Glance Behavior and Takeover Performance, *Human Factors*, July 2020. doi:10.1177/0018720820936793.

Yoon, S. H., & Ji, Y. G. (2018). Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts. *Transportation Research Part F, 60* (2019), 620–631.