Transportation Engineering 2 (2020) 100008

Contents lists available at ScienceDirect

Transportation Engineering

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

Cognitive load estimation using ocular parameters in automotive

Gowdham Prabhakar[a], Abhishek Mukhopadhyay[a], Lrd Murthy[a], Madan Modiksha[b], Deshmukh Sachin[b], Pradipta Biswas[a]

[a] Centre for Product Design and Manufacturing, Indian Institute of Science, Bangalore, India
[b] Faurecia, Pune, India

ARTICLE INFO

Keywords:
Automotive
Distraction
Cognitive load
Pupil dilation
N-back

ABSTRACT

In automotive, usage of electronic devices increased visual inattention of drivers while driving and might lead to accidents. It is often challenging to detect if a driver experienced a change in cognitive state requiring new technology that can best estimate driver's cognitive load. In this paper, we investigated the efficacy of various ocular parameters to estimate cognitive load and detect cognitive state of driver. We derived gaze and pupil-based metrics and evaluated their efficacy in classifying different levels of cognitive states while performing psychometric tests in varying light conditions. We validated the performance of our metrics in simulation as well as in-car environments. We compared the accuracy (from confusion matrix) of detecting cognitive state while performing secondary task using our proposed metrics and Machine Learning models. It was found that a Neural Network model combining multiple ocular metrics showed better accuracy (75%) than individual ocular metrics. Finally, we demonstrated the potential of our system to alert drivers in real-time under critical distractions.

1. Introduction

In recent time, a plethora of sophisticated IVIS (In-Vehicle-Infotainment-System) features increases the amount of non-driving activities of drivers. Engaging in these non-driving activities can cause significant distraction to drivers and might lead to the risk of accidents. NHTSA (National Highway Traffic Safety Administration) reported that 17% of car crashes involved distracted drivers and 5% of distraction-related crashes involved electronic devices. It recommended that the operation of any secondary task should not take drivers' eyes-off-road time greater than 2 s [32]. Systems that detect distraction in real-time will be of great importance for alerting drivers and ensure their safety. Distraction in automotive can be categorised into three types viz., visual, manual and cognitive [32].

1 Visual distraction is caused due to performing tasks that require drivers to look away from road.
2 Manual distraction is caused due to performing tasks that require drivers to take hands off the steering wheel.
3 Cognitive distraction is caused due to performing tasks that require drivers to take his/her mental attention away from driving.

Detecting cognitive distraction is challenging as it is often not explicitly expressed by drivers and can only be detected through estimation of cognitive load. Cognitive load is referred to user's mental effort to solve a given problem [43]. It is the amount of information stored and processed in the working memory. Traditionally, driver monitoring systems use driving behaviour, which is often estimated from telemetry and not from drivers' physiological parameters. There is plethora of physiological monitoring systems available in consumer market but so far there are not many commercial systems available in automotive domain for drivers' behaviour monitoring through physiological parameters. In this paper, we used non-invasive eye gaze trackers to estimate cognitive load from ocular parameters. While earlier research [37] already used eye gaze tracker to estimate cognitive load, validating such a system in a real car driven through usual traffic is challenging. Our study did not interfere with driving task and thus limit the experiment design in terms of collecting data in different affective states or distractions of drivers. Our work went beyond the earlier NHTSA study on measuring glance duration [32] and used Machine Learning (ML) models to estimate and discriminate differences in cognitive load due to undertaking secondary tasks. We evaluated the accuracy of individual ocular metrics in classifying driver's cognitive state. The proposed cognitive load estimation system (patent application number: 201941052358) defined a threshold (baseline value) for all ocular metrics while drivers were only engaged in driving without any secondary task or upcoming road hazard. As finding a global threshold from individual metrics for all drivers was challenging, we developed a ML model for binary classification ('No Task', 'Task') using different ocular metrics that worked for all drivers. We found that the ML based classifier outperformed individual metrics-based classification models in terms of accuracy. In summary, this paper's main contributions are as follows

E-mail address: gowdhamp@iisc.ac.in (G. Prabhakar).

https://doi.org/10.1016/j.treng.2020.100008
Received 27 March 2020; Received in revised form 1 June 2020; Accepted 26 June 2020
Available online xxx

- We proposed new gaze and pupil-based metrics and evaluated their efficacy in varying lighting conditions through standard psychometric tests.
- We evaluated performance of these metrics in real driving environment with professional drivers.
- We proposed a Neural Network (NN) model with an accuracy of 75% to classify driver's cognitive state into performing driving task and performing secondary task while driving.

This paper is organized as follows. We discussed the related work in Section 2 followed by descriptions of algorithms of proposed metrics in Section 3. We discussed psychometric study, driving simulator study and in-car study in Sections 4, 5, and 6 respectively. We discussed about ML based cognitive load estimation module in Section 7 followed by discussion and conclusion in Sections 8 and 9 respectively.

2. Related work

We can detect if a driver is taking eyes off the road or performing any secondary task by monitoring his/her eye gaze movement in a car. There are situations where drivers do not take their eyes off the road, but their thoughts or perceiving road hazard change their cognitive state. Cognitive load can be categorised into three types [44] viz., intrinsic, extraneous and germane. Intrinsic load refers to the inherent level of difficulty in a presented information/task. Extraneous load refers to the load due to the way the information or task is presented. Germane load refers to the work done to create permanent knowledge or schemas. When the cognitive load is estimated as a single entity, there is no distinction between these three types [24]. Cognitive state is the level of cognitive processing of an individual and refers to binary conditions of engaging and disengaging in a single or a set of mental tasks [31]. Secondary tasks affect driver's attention and increase cognitive load, in turn distracting the driver [40]. Though designing a probe for estimating cognitive load with high accuracy is challenging, researchers investigated different techniques for measuring cognitive load using physiological parameters like eye metrics [15, 17, 19, 30, 31, 39, 46, 47], heart rate (skin response) [21], acoustic voice features [9], affective states [2, 41] and electroencephalogram (EEG) [23]. These methods have shortcomings like the heart rate and skin response systems require intrusive methods that causes discomfort to drivers and the acoustic features are effective only when the driver is talking. In some cases, detecting facial feature points becomes challenging due to different conditions of illuminance and field of view of camera inside the car. In addition to that, facial expressions of each individual fail to correspond to the mapped emotion as there is considerable differences in facial expression across individuals and culture. Despite problems of occlusion, lighting and pose variation, researchers gave evidences on affective computing [49] for cognitive state detection. Braun [10] reviewed literature corresponding to state-of-the-art approaches in finding affective state and emotion regulation of drivers in automotive. He depicted that emotion regulation is highly dependent on factors like cognitive load and situational context of the driver. Recent work [27] reported high (99%) accuracy of binary classification for detecting drowsiness of drivers using heart rate from wearable technology like commercially available fitness wristband. Though the trials were conducted in low and high-fidelity simulated environments, the performance of the system in a real driving condition is challenging due to the influence of uncontrolled parameters like ambient light, road condition and traffic participants. While a detailed literature survey on various physiological sensors can be found in other papers [3, 7, 37], the following literature survey only considers measuring cognitive states through ocular parameters. As we focused on eye tracking technologies for interacting with IVIS in our early works [36], we chose ocular metrics to estimate cognitive load such that a single eye tracking device can be used for interaction as well as behaviour monitoring.

2.1. Pupil-based metrics

Redlich [39] and Westphal [47] reported a positive relation between physical task demand and pupil dilation. Hess [22] reported that change in pupil dilation is related to change in the viewing of angles of the photograph. Psychologists [35] reported strong association between cognitive load and pupil dilation of eyes. Recent researchers used frequency of pupil dilation to estimate cognitive load. Gavas [19] and Duchowski [15] measured cognitive load from pupil dilation but used chin rest for their experiments to restrict head movement which is impractical in real driving. Though Prabhakar [37] reported significant performance of pupil-based metrics in estimating cognitive load in a driving simulation study, he did not report EEG values in terms of band power which might have affected the relation between pupil-based metrics and EEG. Marshall [30, 31] reported that a hike in pupil dilation corresponds to increase in cognitive load. This hike is identified by processing the pupil dilation signal for its coefficients of wavelet transform and calculating a metric called Index of Cognitive Activity (ICA). She [30] evaluated this method using mental tasks (questioning the participant to answer verbally) for estimating cognitive load of the participant in automotive as well as aviation [31]. Though she [30] evaluated the robustness of her metric in two conditions of lighting viz., dark and constantly lit room, she did not consider varying light conditions. Babu [3] reported a more detailed literature survey on cognitive load estimation in aviation sector.

2.2. Gaze-based metrics

As there were limited eye gaze trackers that can detect pupil diameter, researchers investigated other ocular metrics like variance in Saccadic Intrusion (SI), change in fixation duration and blink count [7, 28, 29, 35, 48] for estimating cognitive load. Abadi [1]fined a set of characteristics of Monophasic Square Wave Intrusions (MSWI) which is a type of SI. Toyota [4] patented a system for detecting if the driver looked away from the road by detecting his eyelid movements. Biswas [7] measured SI from low cost eye gaze tracker and reported results on simulation studies involving secondary tasks and road hazards. Average velocities of SI found to be higher while drivers undertook secondary tasks and perceive road hazard. He also reported the limitation of choice of ocular parameters and suggested to evaluate different combinations of eye metrics like SI with eye blinks for better performance. Tokuda [46] conducted a dual task study with N-back test and a free viewing task and reported a strong evidence of increase in velocity of SI with respect to difficulty of task. Since the task performance during free viewing task was not discussed, it is unclear whether performance had any impact on experienced cognitive load. Siegenthaler [42] found decrease in microsaccade rate with increase in task difficulty. Their study of arithmetic task involved increasing load on working memory. Gao [18] reported reduction of microsaccade rate with respect to increase in difficulty of arithmetic task for non-visual cognitive processing. Dalmaso [13] reported that microsaccade rate drops with high demand task. Krzysztof [26] used pupil diameter and microsaccades as indicators of cognitive load. He reported a mild evidence of decrease in microsaccade rate with increase in difficulty of task. He also reported a strong evidence of increase in magnitude of microsaccade with increase in difficulty of task. These researchers used a chin rest to ensure minimal head movement which limits the application of such technology to be used in real world systems.

2.3. Summary

From the literature, saccade-based metrics showed mixed relations with difficulty of tasks in different studies whereas pupil-based metrics predominantly showed positive correlation with difficulty of tasks. Since pupil-based metrics were not evaluated in dynamically varying lighting conditions and their measurement involves restricting head movement,

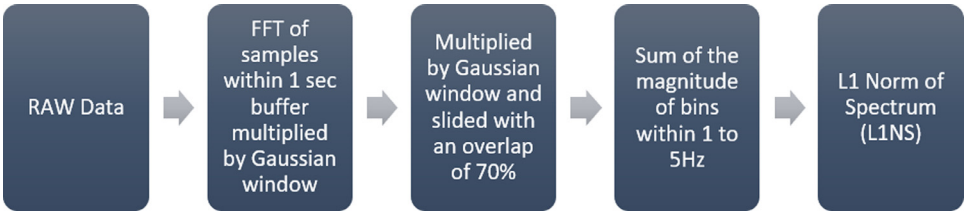

Fig. 1. L1NS algorithm.

Fig. 2. STDP algorithm.

it questions validity of such results in real world conditions. Since we use thresholding to classify an event as distraction using ocular metrics, it is challenging to find a global threshold for all drivers. Our work focuses on addressing these limitations by investigating our proposed ocular metrics to estimate cognitive load of driver and classify cognitive states. As previous works restricted the head movement due to the limitation of eye trackers, they evaluated in a simulation environment and it was challenging to set up the tracker to estimate cognitive load in real cars. In this paper, we evaluated performance of our proposed metrics in laboratory as well as real driving environments.

3. Methodology

This section described different algorithms used to estimate cognitive load from ocular parameters. We calculated saccade rate, fixation rate and median SI velocity from eye gaze points. We also calculated L1 Norm of Spectrum (L1NS), STandard Deviation of Pupil (STDP) and Low Pass Filter of pupil (LPF) from pupil dilation. We calculated metrics corresponding to different tasks (say C1, C2, C3). We designed our experiments such that the task performed by each participant demands high cognitive load for C3 than C2 and C1. The algorithms to calculate these metrics are described in the following sections.

3.1. L1 norm of spectrum (L1NS)

An FFT (Fast Fourier Transform) was applied over the raw data of pupil dilation in 1 s running gaussian window with an overlap of 70%. We calculated L1 norm of bins corresponding to 1 Hz to 5 Hz [34] in the single-sided spectrum as described in Fig. 1. This algorithm uses frequency domain calculation.

3.2. Standard deviation of pupil dilation (STDP)

We calculated the standard deviation of pupil dilation data in a running gaussian window of 1 s with an overlap of 70% as described in the Fig. 2. This algorithm uses time domain calculation and can be deployed using a microcontroller instead of a general-purpose microprocessor.

3.3. Low pass filter of pupil (LPF)

We divided the pupil dilation data into sections of 100 samples. We removed DC offset from the raw data by subtracting its mean. We used a Butterworth lowpass filter with a cut off frequency of 5 Hz and summed up magnitude of filtered data in a running window of 1 s with 70% overlap as described in Fig. 3. This algorithm uses a conventional filtering technique in Digital Signal Processing (DSP) which uses time domain difference equations to filter the signal. It may be noted that the L1NS algorithm uses only frequency domain processing while the STDP uses only time domain processing. The LPF algorithm combines both time and frequency domain processing.

3.4. Saccade rate and fixation rate

We calculated saccade rate (Fig. 4) and fixation rate (Fig. 5) by detecting saccades/fixations from gaze direction data following the method used in [33] as described in. We detected a saccade when gaze velocity was greater than threshold and detected a fixation when gaze velocity was less than threshold. We calculated saccade/fixation rate as number of saccades/fixations per second. We chose the gaze velocity threshold as 100°/s. We illustrated the manual calculation of saccade/fixation rate in [sup 2].

3.5. Median velocity of saccadic intrusions (SI)

We extracted 2D gaze positions (x, y) and their corresponding timestamps from the data file. We detected the occurrences of SI and calculated its velocity [7]. We took the median over the period to get median SI velocity as described in Fig. 6. We illustrated the manual calculation of median SI velocity in [sup 1].

We tested the performance of these ocular metrics in psychometric tasks, driving tasks and secondary tasks. We investigated the accuracy of these metrics in classifying cognitive states by thresholding methods which are discussed in later sections. As finding an optimal threshold that can best classify for all drivers was challenging, we used ML models to train with these ocular metrics and classify best for all drivers.

3.6. Machine learning models

We used Support Vector Classifier (SVC) with polynomial and Radial Basis Function (RBF). RBF kernel on two samples x and x', represented as feature vectors in input space, is defined as $K(x,\ x') = \exp(-\gamma||x,\ x'||^2)$, where $||x, x'||^2$ is squared Euclidean distance between two feature vectors. SVC classifier using RBF kernel has two parameters, γ and C. If we change value of γ from low to high, the curve of the decision boundary also changes from low to high. Correspondingly decision region also changes from broad area to small islands around data points. C is the penalty for misclassifying a data point. Polynomial kernel is formulated as $K(x,\ x') = (x^T x' + 1)^d$, where x, x' represent feature vector in input space with degree of d.

We also used a feedforward Neural Network (NN) model structured as: 8 – 160 – 80 – 1 (input layer – hidden layer – hidden layer – output layer). We used ReLU *[$(f(x) = \max(0,\ x))$, x is the input feature]* activation function in hidden layers. We used Sigmoid *[$(f(x) = 1/(1 + e^{-x}))$, x is the input feature]* activation function in output layer as our problem is binary classification. We used 'Adam' optimization algorithm to overcome slow convergence and high variance in the parameter updates. We

Fig. 3. LPF algorithm.

Fig. 4. Saccade rate algorithm.

Fig. 5. Fixation rate algorithm.

Fig. 6. Saccadic Intrusion (SI) velocity algorithm.

used 'binary cross entropy' loss function. We discussed our user studies in the following sections.

4. User study on psychometric tests

Our first study investigated differences in values of ocular metrics for standard psychometric tests in laboratory. In this section, we described a user study which was conducted to validate if L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity can distinguish between different cognitive loads of participants caused by task difficulty. We used psychometric tests like N-back test and arithmetic questions to assess increase in cognitive load of participants with increase in task difficulty. We chose these tests as they were associated with working memory load [31, 46]. Since the pupil dilation is sensitive [5] to ambient light variation, we evaluated both N-back test and arithmetic test in dark room as well as varying light conditions in same room. While we evaluated N-back test in both auditory and visual presentations, arithmetic test was conducted only in auditory presentation.

We hypothesised that L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity

1 are robust to ambient light variations
2 can be used to distinguish different levels of cognitive load with respect to change in task difficulty of visual and auditory tasks.

4.1. Participants

We collected data from 21 participants (16 Male and 5 female) with an average age of 26 years from our university. We chose participants randomly such that the group had a mixture of people wearing and not wearing prescription lenses. The participants wearing lenses had either spherical or cylindrical or both type of powers.

4.2. Materials

We collected data using Tobii Pro-Glasses 2 [45]. We affixed two ambient light sensor modules, one sensor on either side of the glass frame to capture illumination variations on both eyes independently (Fig. 7). We used a Dell 17″ monitor to display numbers for visual N-Back and a Logitech keyboard to press space bar for responding to N-back test. We also used a Bose SoundLink speaker for auditory cue in auditory N-back test.

4.3. Design

We undertook the following three tests:

1 Auditory N-back Test
2 Visual N-back Test
3 Auditory Arithmetic Test

The auditory tests were carried out in dark as well as dynamically varying light conditions. The room illuminance was varied from 0 to 150 lx by turning ON and OFF the set of lights. The variation of illuminance was randomised.

4.4. N-back test

The N-back test had three levels of difficulties viz., 1-back, 2-back, and 3-back. Participants were shown /spelled one stimulus (sequence of one-digit numbers from 1 to 9) in intervals of 2 s and had to press space bar if current stimulus matches the previous one (1-back), or second previous (2-back), or third previous (3-back). The N-back test levels were randomised to avoid the order effect. We developed a software [8] to spell out/ visually display numbers in N-back and to log the response from participants with a local timestamp.

4.5. Arithmetic test

Arithmetic test had three levels viz., easy, medium, and difficult. We developed a tool using python to read out questions using Text-to-Speech engine in arithmetic test. We recorded participants' response using following steps:

1 Software read out all questions loudly.
2 Participant answered to questions loudly.
3 Instructor checked the answer and pressed right/wrong key to log the event.

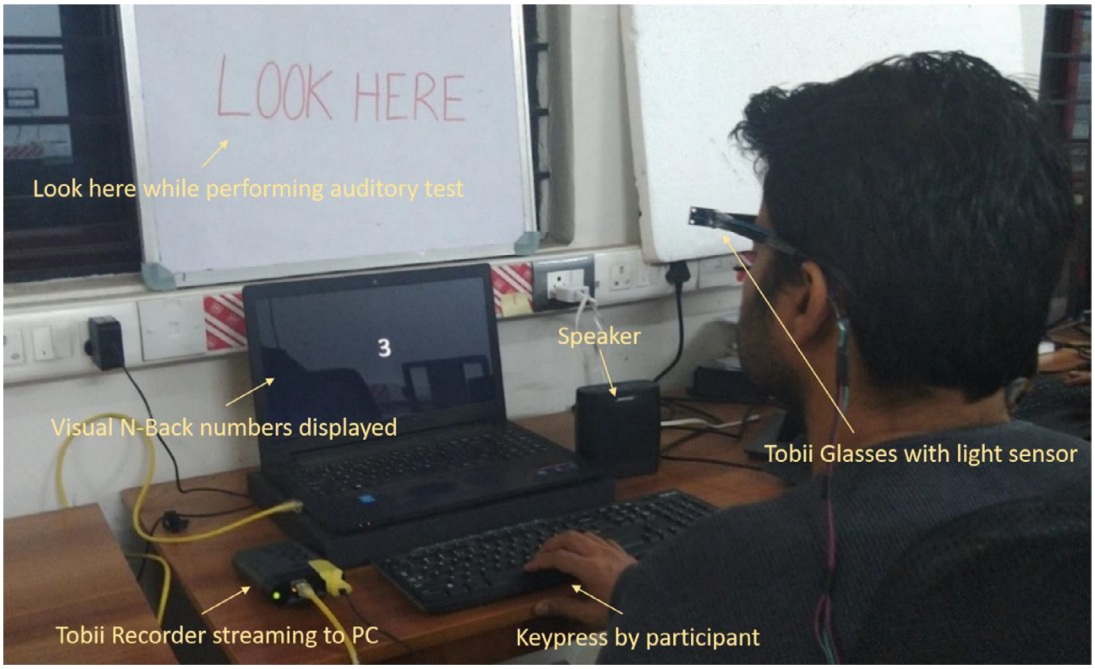

Fig. 7. Participant performing visual N-back test.

Table 1
Performance of N-back test.

	1-Back/Easy	2-Back/Medium	3-Back/Difficult
Auditory N-back dark room	0.961 (0.073)	0.962 (0.059)	0.874 (0.122)
Auditory N-back dynamic light room	0.972 (0.084)	0.948 (0.090)	0.889 (0.121)
Visual N-back	0.985 (0.041)	0.942 (0.071)	0.891 (0.132)
Auditory Arithmetic dark room	0.992 (0.036)	0.905 (0.135)	0.770 (0.207)
Auditory Arithmetic dynamic light room	0.968 (0.067)	0.937 (0.134)	0.730 (0.318)

The difficulty levels were randomised to avoid the order effect.

4.6. Procedure

The participants were asked to wear the Tobii glass affixed with light sensor modules. They were instructed to look at a poster pasted on the wall in front of them and to concentrate on the auditory task given to them. They were asked neither to close their eyes and nor to look around during answering the questions such that the tracker always detected eyes. The participants were explained about the N-back task and arithmetic task and could practice 1-back test before the actual trial to avoid learning effect. The timestamps from logged events were used to synchronize the pupil/gaze data corresponding to start and stop of N-back tests and arithmetic tests. We calculated L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity corresponding to events. We checked if these metrics were high for 3-back than 2-back and 1-back. We also checked if these metrics were high for difficult than medium than easy arithmetic levels.

4.7. Result

4.7.1. Performance of tests

We measured performance of the tests as accuracy calculated from confusion matrix as described in Table 1. The accuracy of N-Back test is calculated as

$$\text{Accuracy} = \frac{correct + avoid}{correct + wrong + avoid + missed}$$

and accuracy of Arithmetic test is calculated as

$$\text{Accuracy} = \frac{correct}{correct + wrong}$$

Table 2
Repeated measure one-way ANOVA for each metric with effect size.

L1NS Right eye	$F(2,19)=6.419, p<0.05, \eta^2 = 0.403$
L1NS Left eye	$F(2,19)=33.964, p<0.05, \eta^2 = 0.781$
STDP Right eye	$F(2,19)=7.849, p<0.05, \eta^2 = 0.452$
STDP Left eye	$F(2,19)=29.408, p<0.05, \eta^2 = 0.756$
LPF Left eye	$F(2,19)=30.718, p<0.05, \eta^2 = 0.764$

As the groups did not follow normality, we performed signed rank test for each pair and found accuracy of 3-Back/Difficult was significantly ($p<0.05$) less than that of 1-Back/Easy for all the tests. The accuracy of 3-back/Difficult was significantly ($p<0.05$) less than 2-Back/Medium for auditory N-back dark room and both arithmetic tests. Accuracy of 2-back/Medium was significantly ($p<0.05$) less than 1-back/Easy for visual N-back and auditory arithmetic dark room.

4.7.2. Visual N-Back

A repeated measure one-way ANOVA for metrics in Visual N-back is described in Table 2.

We found that L1NS and STDP of both eyes were significantly (t-test: $p<0.05$) higher for 3-back than 1-back. Similarly, 3-back was significantly (t-test: $p<0.05$) higher than 2-back. We also found that LPF of left eye was significantly (t-test: $p<0.05$) higher for 3-back than 1-back and higher for 3-back than 2-back. We found LPF right (3-back > 1-back) to be tending to significance ($p<0.1$).We did not find any significance difference for saccade rate, fixation rate and median SI velocity. We showed comparison graph of L1NS for visual N-back in Fig. 8a.

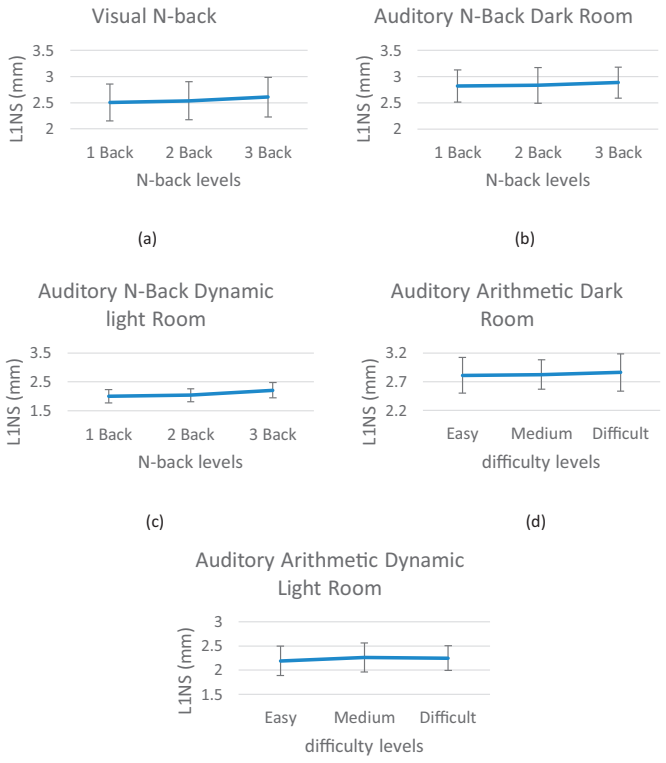

Fig. 8. L1NS of right eye for (From Top left) (a)visual N-back, (b)auditory N-Back dark room, (c)auditory N-back dynamic light room, (d)auditory arithmetic dark room, (e)auditory arithmetic dynamic light room.

Table 3
Repeated measure one-way ANOVA for each metric with effect size.

L1NS Right eye	$F(2,19)=8.155, p<0.05, \eta^2 = 0.462$
L1NS Left eye	$F(2,19)=7.813, p<0.05, \eta^2 = 0.451$
STDP Right eye	$F(2,19)=5.91, p<0.05, \eta^2 = 0.384$
STDP Left eye	$F(2,19)=15.842, p<0.05, \eta^2 = 0.625$
LPF Left eye	$F(2,19)=18.088, p<0.05, \eta^2 = 0.656$

4.7.3. Auditory N-Back dark room

A repeated measure one-way ANOVA for metrics in Auditory N-back dark room is described in Table 3.

We found that L1NS and STDP of both eyes as well as LPF of Left eye were significantly (*t*-test: $p<0.05$) higher for 3-back than 1-back. We found L1NS left (2-back > 1-back), LPF left (2-back > 1-back), saccade rate left (3-back > 1-back), fixation rate left (3-back > 1-back) and median SI velocity (2-back > 1-back, 3-back > 2-back) to be tending to significance ($p<0.1$). We showed comparison graph of L1NS for auditory N-back in dark room in Fig. 8b.

4.7.4. Auditory N-Back dynamic light room

A repeated measure one-way ANOVA for metrics in Auditory N-back dynamic light room is described in Table 4.

We found that L1NS, STDP, and LPF of both eyes were significantly (*t*-test: $p<0.05$) higher for 3-back than 1-back. Similarly, 3-back was significantly (*t*-test: $p<0.05$) higher than 2-back. We found that saccade rate and fixation rate of left eye were significantly higher for 3-back than 1-back as well as for 2-back than 1-back. We found saccade rate right (2-back > 1-back) and fixation rate right (2-back > 1-back) to be tending to significance ($p<0.1$). We showed comparison graph of L1NS for auditory N-back in dynamic light room in Fig. 8c.

Table 4
Repeated measure one-way ANOVA for each metric with effect size.

L1NS Right eye	$F(2,19)=24.961, p<0.05, \eta^2 = 0.724$
L1NS Left eye	$F(2,19)=43.017, p<0.05, \eta^2 = 0.819$
STDP Right eye	$F(2,19)=29.461, p<0.05, \eta^2 = 0.756$
STDP Left eye	$F(2,19)=39.767, p<0.05, \eta^2 = 0.807$
LPF Left eye	$F(2,19)=38.847, p<0.05, \eta^2 = 0.804$
LPF Right eye	$F(2,19)=28.797, p<0.05, \eta^2 = 0.752$
Fixation rate left eye	$F(2,19)=5.139, p<0.05, \eta 2 = 0.351$
Saccade rate right eye	$F(2,19)=5.139, p<0.05, \eta 2 = 0.351$

Table 5
Repeated measure one-way ANOVA for each metric with effect size.

LPF Left eye	$F(2,19)=7.657, p<0.05, \eta^2 = 0.446$
LPF Right eye	$F(2,19)=6.280, p<0.05, \eta^2 = 0.398$

Table 6
Repeated measure one-way ANOVA for each metric with effect size.

L1NS Right eye	$F(2,18)=4.928, p<0.05, \eta^2 = 0.354$
L1NS Left eye	$F(2,19)=5.966, p<0.05, \eta^2 = 0.386$
STDP Right eye	$F(2,18)=4.790, p<0.05, \eta^2 = 0.347$
STDP Left eye	$F(2,19)=4.595, p<0.05, \eta^2 = 0.326$
LPF Left eye	$F(2,18)=7.662, p<0.05, \eta^2 = 0.460$
LPF Right eye	$F(2,18)=6.648, p<0.05, \eta^2 = 0.425$

4.7.5. Auditory arithmetic dark room

A repeated measure one-way ANOVA for metrics in Arithmetic dark room is described in Table 5.

We found no significant differences for L1NS and STDP of both eyes. LPF of both eyes were significantly (*t*-test: $p<0.05$) higher for 3-back than 1-back. Similarly, 3-back was significantly (*t*-test: $p<0.05$) higher than 2-back. We found L1NS left (3-back > 2-back), L1NS right (3-back > 2-back, 3-back > 1-back), STDP left (3-back > 2-back), STDP right (3-back > 2-back, 3-back > 1-back), LPF left (2-back > 1-back) to be tending to significance ($p<0.1$). We showed comparison graph of L1NS for auditory arithmetic test in dark room in Fig. 8d.

4.7.6. Auditory arithmetic dynamic light room

A repeated measure one-way ANOVA for metrics in Arithmetic dynamic light room is described in Table 6.

We found that L1NS and STDP of both eyes were significantly (*t*-test: $p<0.05$) higher for 2-back than 1-back. We also found that LPF of both eyes were significantly (*t*-test: $p<0.05$) higher for 3-back than 1-back. We found L1NS left (3-back > 1-back), L1NS right (3-back > 1-back), STDP left (3-back > 1-back), STDP right (3-back > 1-back) to be tending to significance ($p<0.1$). We showed comparison graph of L1NS for auditory arithmetic test in dynamic light room in Fig. 8e.

4.7.7. Interaction effect

We performed a repeated measure two-way ANOVA on metric values for factors like light, presentation, task type and task difficulty and reported the metrics which showed significant interaction effect between respective factors in Table 7a (tests of within-subjects effects) and Table 8a (multivariate tests). The factors and their levels are listed below.

1 Dark room versus dynamic light room (factors: light and task difficulty)
   a Dark room (Auditory N-back) versus dynamic light room (Auditory N-back)
   b Dark room (Auditory Arithmetic) versus dynamic light room (Auditory Arithmetic)

Table 7a
List of Interaction Effects.

Interacting factors	Levels of factors	Metrics	Tests of Within-Subjects Effects (sphericity assumed)
Light versus Task- Difficulty	Dark room (Auditory N-back) versus dynamic light room (Auditory N-back)	STDP Left	$F(2,40)=19.369, p<0.05, \eta 2 = 0.492$
		STDP Right	$F(2,40)=28.784, p<0.05, \eta 2 = 0.309$
		L1NS Left	$F(2,40)=20.654, p<0.05, \eta 2 = 0.508$
		L1NS Right	$F(2,40)=8.215, p<0.05, \eta 2 = 0.291$
		LPF Left	$F(2,40)=18.881, p<0.05, \eta 2 = 0.486$
		LPF Right	$F(2,40)=8.303, p<0.05, \eta 2 = 0.293$
	Dark room (Auditory Arithmetic) versus dynamic light room (Auditory Arithmetic)	STDP Left	$F(2,40)=3.708, p<0.05, \eta 2 = 0.156$
		L1NS Left	$F(2,40)=3.394, p<0.05, \eta 2 = 0.145$
		STDP Left	$F(2,40)=18.229, p<0.05, \eta 2 = 0.477$
		STDP Right	$F(2,38)=13.501, p<0.05, \eta 2 = 0.415$
Task Type Versus Task- Difficulty	Auditory Arithmetic (Dynamic light room) versus Auditory N-back (Dynamic light room)	L1NS Left	$F(2,40)=20.832, p<0.05, \eta 2 = 0.51$
		L1NS Right	$F(2,38)=15.238, p<0.05, \eta 2 = 0.445$
		LPF Left	$F(2,38)=17.889, p<0.05, \eta 2 = 0.485$
		LPF Right	$F(2,38)=12.496, p<0.05, \eta 2 = 0.397$
Presentation versus Task- Difficulty	Auditory N-back (Dark room) Versus Visual N-back	STDP Left	$F(2,40)=8.348, p<0.05, \eta 2 = 0.294$
		L1NS Left	$F(2,40)=9.381, p<0.05, \eta 2 = 0.319$
		LPF Left	$F(2,40)=8.855, p<0.05, \eta 2 = 0.307$
	Auditory N-back (Dynamic light room) versus Visual N-back	STDP Left	$F(2,40)=3.719, p<0.05, \eta 2 = 0.157$
		STDP Right	$F(2,40)=4.979, p<0.05, \eta 2 = 0.199$
		L1NS Left	$F(2,40)=3.444, p<0.05, \eta 2 = 0.147$
		L1NS Right	$F(2,40)=4.807, p<0.05, \eta 2 = 0.194$
		LPF Left	$F(2,40)=4.154, p<0.05, \eta 2 = 0.172$

For Auditory N-back (Dynamic light room) versus Visual N-back, LPF Right violated sphericity assumption and we found significant interaction between the factors using Greenhouse-Geisser as $F(1.453,29.059)=5.094, p<0.05, \eta 2 = 0.203$.

Table 7b
Friedman test with Kendall's-W as effect size.

Alpha	Chi square(2)=11.091, $p<0.05$, Kendall's $W = 0.504$
Low Beta	Chi square(2)=14.364, $p<0.05$, Kendall's $W = 0.653$
Theta	Chi square(2)=11.091, $p<0.05$, Kendall's $W = 0.504$

2 Auditory Arithmetic versus Auditory N-back (factors: task type and task difficulty)
   a Auditory Arithmetic (Dark room) versus Auditory N-back (Dark room)
   b Auditory Arithmetic (Dynamic light room) versus Auditory N-back (Dynamic light room)
3 Auditory N-back versus Visual N-back (factors: presentation and task difficulty)
   a Auditory N-back (Dark room) versus Visual N-back
   b Auditory N-back (Dynamic light room) versus Visual N-back

We did not find any significant interaction effect between factors for gaze-based metrics.

4.8. Discussion

We confirmed the decrease in performance with increase in task difficulty [20, 46]. We observed that L1NS, STDP and LPF increased with increase in task difficulty consistent with the study reported by Coulacoglou [11]. In all the cases, we observed that the parameter corresponding to difficult task (3-Back and difficult arithmetic) was significantly higher than that corresponding to easy task (1-Back and easy arithmetic). The intermediate task difficulty did not have significant effect with respect to all parameters. This might be because of the overlapping region of cognitive load present in 2-back test due to the transition of difficulty levels from 1-back to 3-back tests. Some participants would have found 2-back level easy and some would have found it difficult. Similarly, overlapping region might be present in medium level arithmetic questions. We found relatively largest effect sizes in L1NS left eye for Visual N-Back, LPF left eye for Auditory N-Back Dark room, L1NS left eye for Auditory N-Back Dynamic light room, LPF left eye for Auditory Arithmetic Dark room, LPF left eye for Auditory Arithmetic Dynamic light room. This infers that each metric performed significantly in each test condition. We also observed that the trend of increase in metric values with respect to increase in task difficulty is same for changes in light conditions for visual and auditory presentations. Though we found interaction effect between task difficulty and lighting conditions for pupil-based metrics, the *t*-test result showed that our pupil-based metrics were able to significantly distinguish between task difficulties in different lighting conditions. Similarly, a set of pupil-based metrics were able to significantly distinguish between task difficulties in different task type and presentation conditions despite significant interaction between the factors.

5. User study on driving simulator

Our second study investigated differences in values of ocular metrics for a standard lane changing task and performing secondary tasks while driving in simulated automotive environment in laboratory. In this section, we discussed the user study which we conducted in a driving simulator to validate our proposed metrics with respect to EEG. In this study, our motive was to find whether these metrics can classify cognitive state of participants while driving and performing secondary task.

5.1. Participants

We recruited 11 participants (10 male and 1 female) with an average age of 26 years from our university to engage in this study. We conducted free trials for participants to get used to driving simulator. We started the actual trial after sufficiently training participants to make sure that increase in cognitive load correspond only to the task and not to the driving experience of the simulator.

5.2. Materials

Our driving simulator consist of a Logitech G29 steering wheel with pedals and ISO 26022 lane changing task software. We used an Emotive Insight 5 channel wireless EEG tracker for recording EEG and Tobii Pro-Glasses 2 for recording eye gaze and pupil dilation of participants. We used Lenovo Yoga 500 laptop as IVIS (In-Vehicle Infotainment System) display.

Table 8a
Multivariate tests.

Interacting factors	Levels of factors	Metrics	Multivariate test (Pillai's trace)
Light	Dark room (Auditory N-back)	STDP Left	$F(2,19)=18.634, p<0.05, \eta 2 = 0.662$
versus	versus	STDP Right	$F(2,19)=8.398, p<0.05, \eta 2 = 0.469$
Task- Difficulty	dynamic light room (Auditory N-back)	L1NS Left	$F(2,19)=20.084, p<0.05, \eta 2 = 0.679$
		L1NS Right	$F(2,19)=7.375, p<0.05, \eta 2 = 0.437$
		LPF Left	$F(2,19)=17.182, p<0.05, \eta 2 = 0.644$
		LPF Right	$F(2,19)=6.747, p<0.05, \eta 2 = 0.415$
	Dark room (Auditory Arithmetic)	STDP Left	$F(2,19)=4.073, p<0.05, \eta 2 = 0.3$
	versus		$F(2,19)=4.813, p<0.05, \eta 2 = 0.336$
	dynamic light room (Auditory Arithmetic)	L1NS Left	
Task Type	Auditory Arithmetic (Dynamic light	STDP Left	$F(2,19)=21.436, p<0.05, \eta 2 = 0.693$
Versus	room)	STDP Right	$F(2,18)=11.517, p<0.05, \eta 2 = 0.561$
Task- Difficulty	versus	L1NS Left	$F(2,19)=20.509, p<0.05, \eta 2 = 0.683$
	Auditory N-back (Dynamic light room)	L1NS Right	$F(2,18)=12.273, p<0.05, \eta 2 = 0.577$
		LPF Left	$F(2,18)=15.759, p<0.05, \eta 2 = 0.636$
		LPF Right	$F(2,18)=10.168, p<0.05, \eta 2 = 0.53$
Presentation versus	Auditory N-back (Dark room)	STDP Left	$F(2,19)=10.415, p<0.05, \eta 2 = 0.523$
Task- Difficulty	Versus	L1NS Left	$F(2,19)=11.691, p<0.05, \eta 2 = 0.552$
	Visual N-back	LPF Left	$F(2,19)=10.732, p<0.05, \eta 2 = 0.53$
	Auditory N-back (Dynamic light room)	STDP Left	$F(2,19)=3.717, p<0.05, \eta 2 = 0.281$
		STDP Right	$F(2,19)=4.442, p<0.05, \eta 2 = 0.319$
	versus	L1NS Left	$F(2,19)=3.954, p<0.05, \eta 2 = 0.294$
	Visual N-back	L1NS Right	$F(2,19)=3.781, p<0.05, \eta 2 = 0.285$
		LPF Left	$F(2,19)=3.872, p<0.05, \eta 2 = 0.29$
		LPF Right	$F(2,19)=3.295, p<0.05, \eta 2 = 0.258$

Fig. 9. Participant performing secondary task while driving.

5.3. Design

The participants had to undergo three trials of driving tasks by wearing Tobii glasses and EEG tracker on his/her head. The driving simulator did not accommodate any on-road traffic events. The experimental setup is illustrated in Fig. 9. We recorded baseline data by letting participants to freely drive without engaging in any secondary task. We recorded this trial as reference case (C1). In the second case, we instructed participants to follow the lane changing instructions while driving. We recorded this trial as case 2 (C2). In the third case, we instructed participants to drive by following lane changing instructions and perform a secondary task of selecting a button on the IVIS display on hearing an auditory cue. We recorded this trial as case 3 (C3). The dashboard display was designed like one of the dashboard systems of Jaguar Land Rover. The IVIS display was placed to the left of driving simulator. To summarise the three conditions:

- Driving without any secondary tasks (C1)
- Driving and following Lane changing instructions (C2)
- Driving with Lane changing instruction and perform a selection task on IVIS (C3)

5.4. Procedure

We instructed participants to wear Tobii Pro-glasses and EEG tracker. We instructed them to drive safely without veering off from the road. We collected the eye gaze position, pupil diameter and EEG data from par-

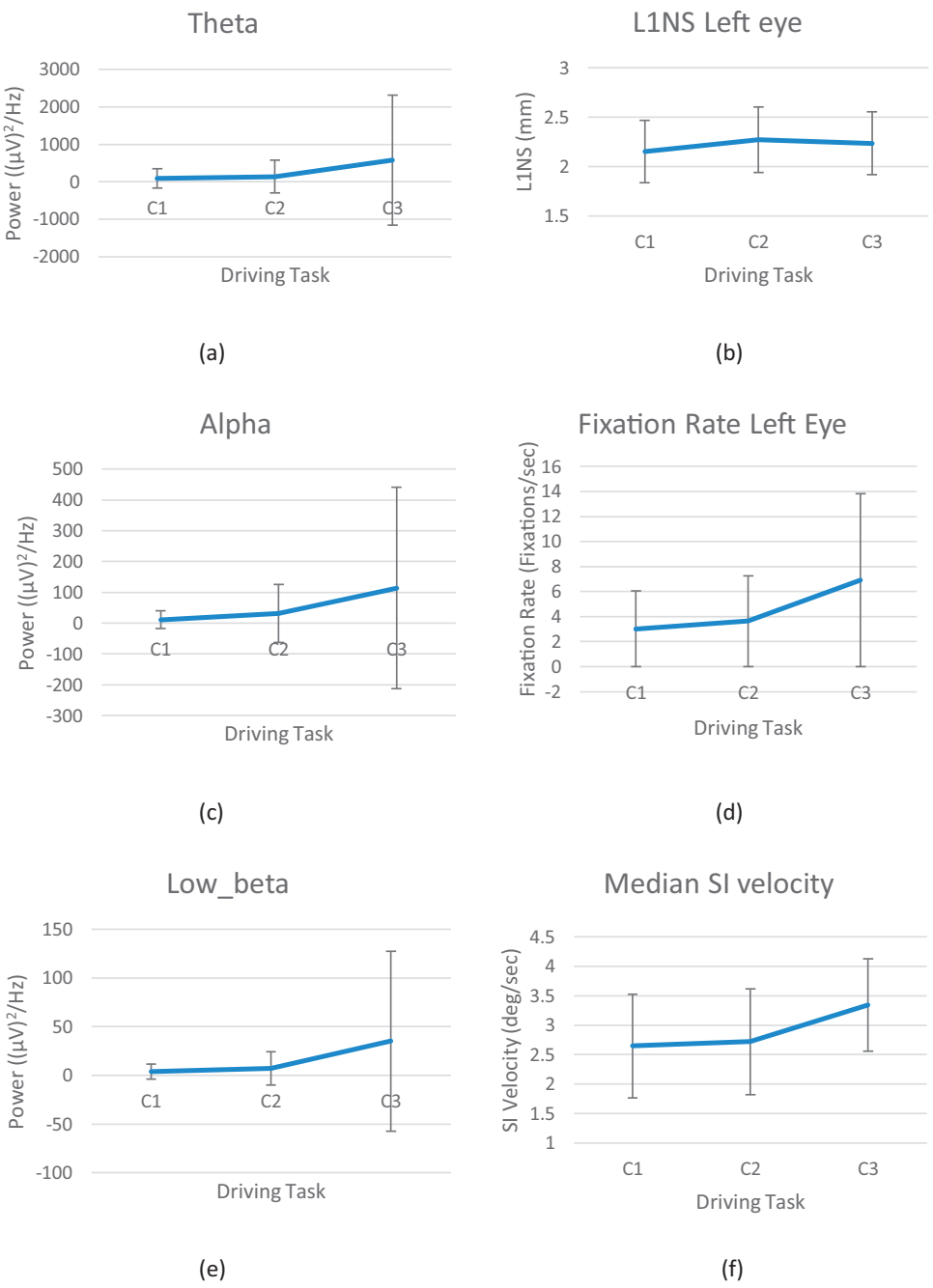

Fig. 10. (From Top left) (a)Average Theta, (b)L1NS left eye, (c)Alpha, (d)Fixation rate left eye, (e)low Beta and (f)median SI velocity in different driving conditions.

ticipants for all three cases. We analysed data to find if our parameters classify C1, C2 and C3.

5.5. Results

A Friedman test was performed on EEG bands as described in Table 7b.

We performed a Kolmogorov-Smirnov test on alpha (Fig. 10c), low beta (Fig. 10e) and theta (Fig. 10a) bands of EEG. We found that all three of them were not normally distributed. We then performed Signed-Rank test between each pair and found C3 was significantly ($p<0.05$) greater than C1 and C3 was significantly ($p<0.05$) greater than C2. We did not find significant difference in high beta and gamma bands.

We found the driving performances as described in Table 8b. A Kolmogorov-Smirnov test showed the groups were normally distributed. We found mean deviation from lane for C3 was significantly (t-test: $p<0.05$) greater than C1 and C2 was significantly (t-test: $p<0.05$) greater than C1. Average speed for C3 was significantly (t-test: $p<0.05$) less than C1 and C2 was significantly (t-test: $p<0.05$) less than C1. A repeated measure one-way ANOVA was performed on our metrics as described in Table 9.

We performed a Kolmogorov-Smirnov test on L1NS, STDP and LPF of pupil data for both eyes, saccade rate and fixation rate, and median SI velocity of gaze points of both eyes. We found that all of them were normally distributed. We then performed a t-test between each pair and found C2 was significantly ($p<0.05$) greater than C1, C3 was significantly ($p<0.05$) greater than C1 for L1NS (Fig. 10b), STDP and LPF of both eyes (Fig. 10). Though we found C3 significantly less than C2 for

Table 8b
Driving performance for each driving conditions.

	C1	C2	C3
Mean deviation from lane	0.784 (0.387)	2.349 (0.590)	2.386 (0.722)
Mean speed	53.281 (10.518)	48.444 (11.195)	46.412 (10.166)

Table 9
Repeated measure one-way ANOVA for each metric with effect size.

L1NS Right eye	$F(2,9)=21.273, p<0.05, \eta^2 = 0.825$
L1NS Left eye	$F(2,9)=18.574, p<0.05, \eta^2 = 0.805$
STDP Right eye	$F(2,9)=22.001, p<0.05, \eta^2 = 0.830$
STDP Left eye	$F(2,9)=19.265, p<0.05, \eta^2 = 0.811$
LPF Left eye	$F(2,9)=19.120, p<0.05, \eta^2 = 0.809$
LPF Right eye	$F(2,9)=22.825, p<0.05, \eta^2 = 0.835$
Saccade rate Left eye	$F(2,9)=30.812, p<0.05, \eta^2 = 0.873$
Saccade rate Right eye	$F(2,9)=20.913, p<0.05, \eta^2 = 0.823$
Fixation rate Left eye	$F(2,9)=31.041, p<0.05, \eta^2 = 0.873$
Fixation rate Right eye	$F(2,9)=21.066, p<0.05, \eta^2 = 0.824$
Median SI velocity	$F(2,9)=5.341, p<0.05, \eta^2 = 0.543$

L1NS, STDP and LPF of left eye, we did not find significant difference for those in right eye. We performed paired *t*-test between each pair and found C3 was significantly ($p<0.05$) greater than C1 and C3 was significantly ($p<0.05$) greater than C2 for fixation rate (Fig. 10d), saccade rate of both eyes and median SI velocity (Fig. 10f).

5.6. Discussion

The EEG band power significantly increased and driving performance significantly decreased from driving conditions C1 to C3. This confirms the increase in task difficulty of driving conditions from C1 to C3. Since we also found increasing trend in pupil and gaze-based metric values from C1 to C3, we infer that our proposed metrics agree with EEG band powers. This establishes the relation between EEG metrics and proposed ocular metrics which validates the usage of our proposed metrics as probes to estimate cognitive load. We found relatively large effect sizes in Low Beta (EEG), L1NS, LPF, STDP and largest in Fixation/saccade rate left eye for driving conditions. This infers that both pupil-based and gaze-based metrics were significantly able to distinguish between the cognitive states. The performance of left eye metrics was different from that of right eye metrics. This study gives us the evidence to consider pupil and gaze-based metrics for estimating cognitive load for automotive interfaces. It also shows that our metrics can be used in real-time as they are calculated using 1 s running window. This outcome is further confirmed by conducting an in-car study with professional drivers.

6. User study inside moving vehicles

Our third study investigated differences in values of ocular metrics for driving in real cars while drivers undertook secondary tasks as they do in natural driving situation. This user study was conducted to validate the ability of our proposed parameters to distinguish between different cognitive states caused by performing secondary tasks in cars. We hypothesised that L1NS, STDP, LPF, saccade rate, fixation rate, median SI velocity can distinguish between 'No Task' and secondary task conditions.

6.1. Participants

A set of 13 professional male drivers with an average age of 36 years (std: 8 years) participated in the study. All drivers were hired from a travel agency. All drivers had an average driving experience of 7150 km (std: 2700 km). The average number of years the drivers drove holding a valid license was 11 years (std: 7 years). All participants had a valid two-wheeler and four-wheeler Indian driving license.

We took all necessary permissions and consent from all drivers before undertaking trials and the study never interfered with the driving task.

6.2. Material

We used Tobii Pro-glasses 2 to record the first-person video as well as ocular parameters. We used Tobii Pro-software to export the data into TSV (Tab Separated Values) file. We used video tagging tool to find instants where the driver performed secondary tasks and used MATLAB to analyse data.

6.3. Design

We designed the study such that each driver had to start driving his vehicle from a fixed start point to a fixed location inside our university and return to the same start point. We recorded data for each driver during the trip. We asked them to perform secondary tasks while driving and calculated various metric values for estimating their cognitive load. We designed secondary tasks such that they cause visual, manual, cognitive or combined distractions. Drivers performed secondary tasks all by themselves without our intervention. We recorded the following secondary tasks and the type of distraction is mentioned within brackets

1. Driver turned on/off Air Conditioner (Visual and Manual)
2. Driver turned on/off music player or changed radio station (Visual and Manual)
3. Driver talking with passengers (Visual and Cognitive)
4. Driver opening/closing windows (Visual and Manual)
5. Driver received a call from an unknown number while driving (Visual, Manual and Cognitive)

We did not control the time of day for the study and collected data in both sunny, shadowy, dusk and cloudy conditions.

6.4. Procedure

We calibrated the eye gaze tracker for each driver before igniting the engine and after that did not interfere with the driving task. We instructed them to drive inside our institute premises from a pre-fixed location and come back to the same place. We tagged the video for timestamps of each secondary task event as illustrated in Fig. 11 where the consecutive timestamps correspond to start and end of an event. We also tagged the timestamps when driver neither performed any secondary tasks nor observed a road hazard. We calculated L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity corresponding to events. We checked if metric values were high for performing secondary tasks (Task) compared to performing only driving task (No Task).

6.5. Results

We calculated average metric values corresponding to 'Task' and 'No Task' events for each driver. We undertook a paired *t*-test on 'Task' and

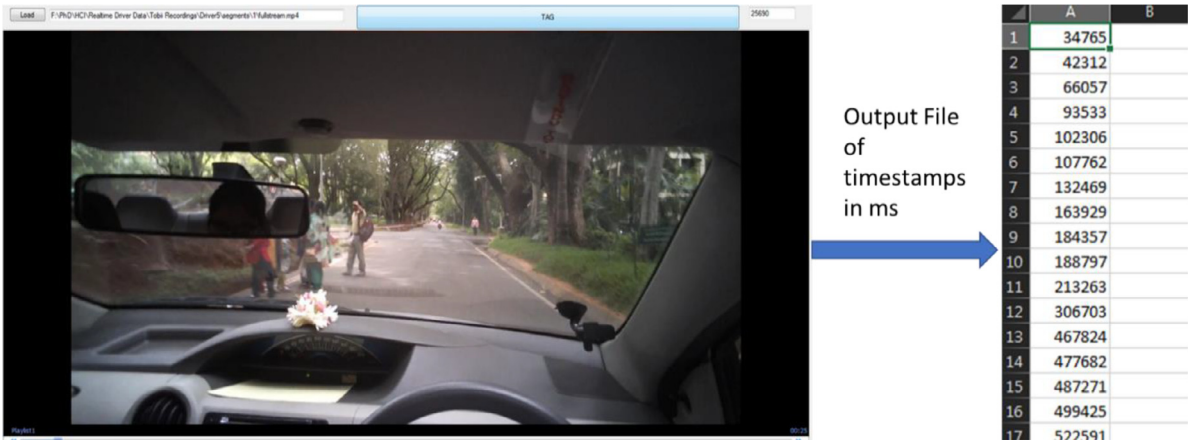

Fig. 11. Process of tagging video and generating timestamp file.

Table 10
Cohens D value for each metric.

1NS Left eye	0.388
STDP Left eye	0.350
LPF Left eye	0.378
Saccade rate Left eye	0.599
Saccade rate Right eye	0.539
Fixation rate Left eye	0.560
Fixation rate Right eye	0.640
Median SI velocity	0.244

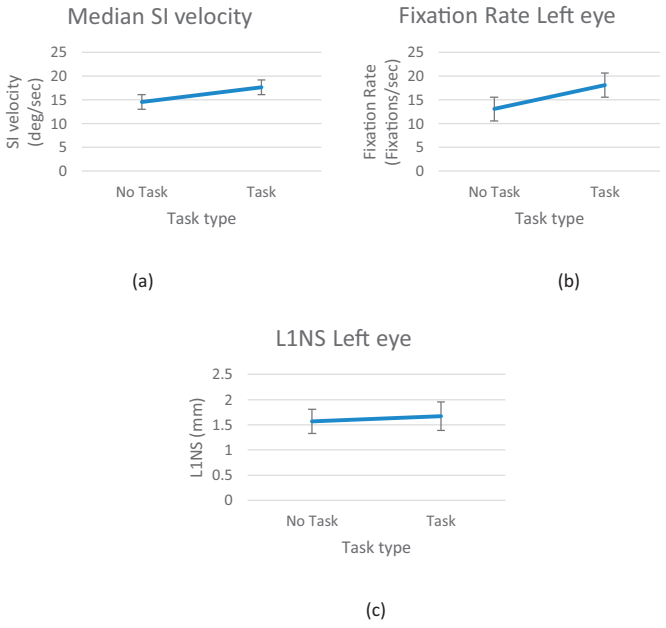

Fig. 12. (From Top left) (a)Median SI velocity, (b)Fixation rate and (c)L1NS corresponding to 'Task' and No Task.

'No Task' groups where each data point corresponded to each of 13 drivers. We reported Cohens D values as effect sizes of tests in Table 10.

We found that L1NS (Fig. 12c), STDP, LPF of left eye was significantly (t-test: $p<0.05$) higher for operating secondary task than driving without any secondary task. We found fixation rate and saccade rate (Fig. 12b) for both eyes were significantly (t-test: $p<0.05$) higher for operating secondary task than driving without any secondary task. We found that median SI velocity was significantly (t-test: $p<0.05$) higher for operating secondary task than driving without any secondary task (Fig. 12a).

6.6. Discussion

Our results showed that L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity were significantly higher for drivers while performing secondary tasks than when they were not doing any task. This clearly indicates that our proposed metrics were able to distinguish between 'Task' and 'No Task' performed by drivers. The metrics were significantly different only for left eye which might be an effect of right-handed driving by drivers. We found relatively largest effect size in Fixation rate right eye for 'Task' and 'No Task'. This infers that the gaze-based metrics are equally effective as pupil-based metrics in classifying cognitive states of drivers in real cars. In the previous sections, we validated our proposed metrics using secondary tasks.

We computed global thresholds for individual metrics using which we can classify 'Task' and 'No Task' for all the users. We used all the individual metrics with this approach and obtained highest accuracy of 68.8% with saccade rate of right eye (Fig. 15). The rest of the metrics gave lower classification accuracy. Hence, we found that obtaining a universal threshold for all drivers using a single metric is challenging. These global thresholds based on individual metric values may change based on the drivers in the dataset and uncertainty prevails over generalizability, we built a ML based model to classify the cognitive states using all ocular metrics.

7. Machine learning-based classification

In this section, we explained our ML models for binary classification. Our proposed ML-based system used L1NS, LPF, STDP, saccade rate, fixation rate and median SI velocity metrics as input features. We used Support Vector Classifier (SVC) with different kernels and compared their accuracy. We proposed NN model to increase accuracy of the system. The ML models classify driver's cognitive states into two states viz., 'No Task' and 'Task'. The working of our proposed ML based cognitive load monitoring system is illustrated in Fig. 13. We evaluated the performance of our ML models in terms of their classification accuracies.

7.1. Procedure

Initially we used our proposed metrics as input to ML models to predict 'No Task' and 'Task' classes. We started our prediction model by using SVC with Polynomial and RBF kernels. We then compared results

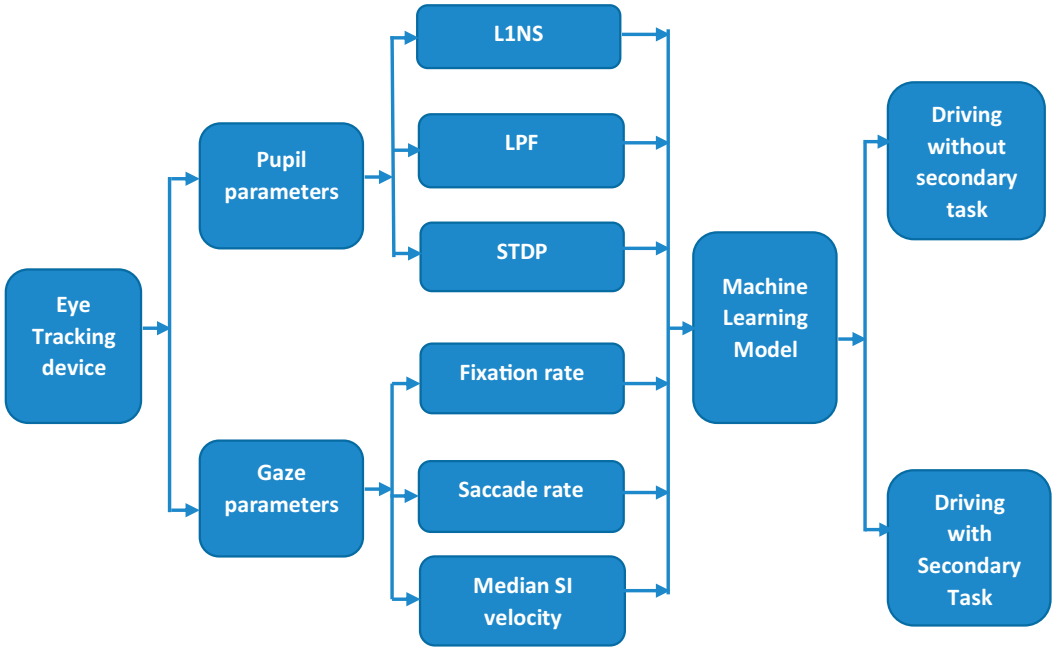

Fig. 13. Block diagram of cognitive load monitoring system.

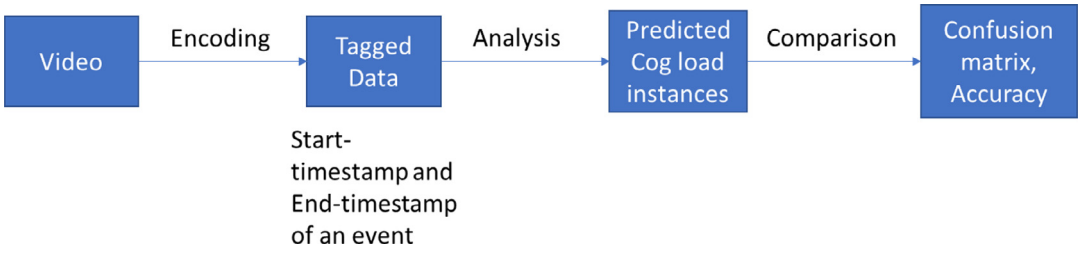

Fig. 14. Process of tagging video and calculating accuracy of classification of distraction.

of SVC model and NN model. The procedure to calculate accuracy of classifiers (ML models and individual metrics) is described in Fig. 14.

We calculated the accuracy of classification of 'Task' and 'No Task' in the existing dataset for individual metrics like L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity. We classified an event as 'Task' if the metric was above threshold and classified as 'No Task' if the metric was below threshold. We used two methods to calculate classification accuracy corresponding to two thresholds (global and individual). In former case the threshold (global) was calculated by taking average of all "No Task" metric values and used the single value as threshold for all drivers. In latter case we used the respective "No Task" metric value as threshold (individual) for each driver.

We took 'Task' region as positive and 'No Task' region as negative. We counted True positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) as follows:

TP: If (metric > threshold) and lies in 'Task' region
FP: If (metric >threshold) and lies in 'No Task' region
FN: If (metric <threshold) and lies in 'Task' region
TN: If (metric <threshold) and lies in 'No Task' region

Accuracy is calculated using the following equation.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

7.2. Results

We did empirical experiments with different combination of values for γ and C and obtained highest accuracy of 66.9% with the value of γ and C as 0.001 and 1 respectively. Polynomial kernel SVC showed promising results without scaling the dataset. We obtained 58.6% accuracy with value of γ, C, and d as 100, 1, and 6 respectively. To increase the classification accuracy, we used feed forward NN. We trained and tested our ML models on drivers' data described in previous section. We obtained highest accuracy of 75% using feed forward NN model. We compared the accuracies of individual metrics against accuracies obtained from ML models as illustrated in Fig. 15.

7.3. Discussion

We took average metric values corresponding to 'No Task' class of all drivers as global threshold. We evaluated the accuracy of our pupil and gaze-based metrics to classify the cognitive states and compared it with that of ML models for both global and individual threshold methods. We found increase in accuracy of classification by using NN model. Though our NN model takes both pupil-based and gaze-based metrics as input features and achieves a binary-classification accuracy of 75%, accuracy of saccade rate of right eye with global threshold is only next to NN model with a difference of 7%. Hence, we infer that binary cognitive state classification can also be performed using a low-cost eye tracker without pupil diameter-based measurement.

7.4. Machine learning model for multi-class classification

It may be argued that change in drivers' attention due to undertaking a few secondary tasks involving the head down display in central task can be more accurately detected by tracking head movement in

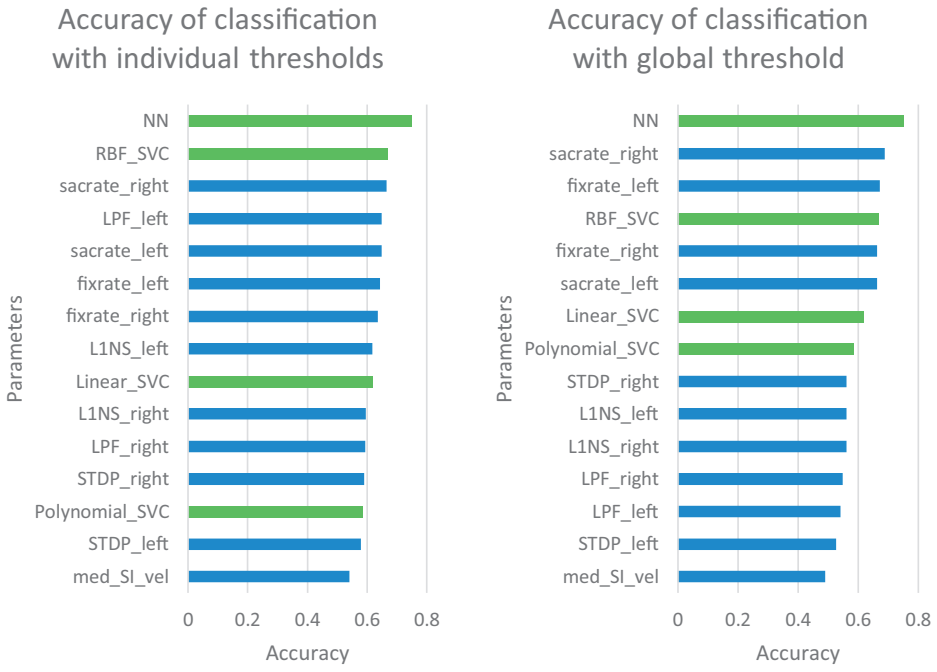

Fig. 15. Comparison of accuracies of individual metrics(blue) and ML models(green) with individual (left) thresholds and global (right) thresholds. In this figure L1NS_left: L1 Norm of spectrum of left Pupil, L1NS_right: L1 Norm of spectrum of right Pupil, STDP_left: Standard Deviation of left Pupil, STDP_right: Standard Deviation of right Pupil, LPF_left: Low Pass Filter of left pupil, LPF_right: Low Pass Filter of right pupil, med_SI_vel: Median velocity of Saccadic Intrusion, sacrate_left: Saccade rate of left eye, sacrate_right: Saccade rate of right eye, fixrate_left: Fixation rate of left eye, fixrate_right: Fixation rate of right eye.

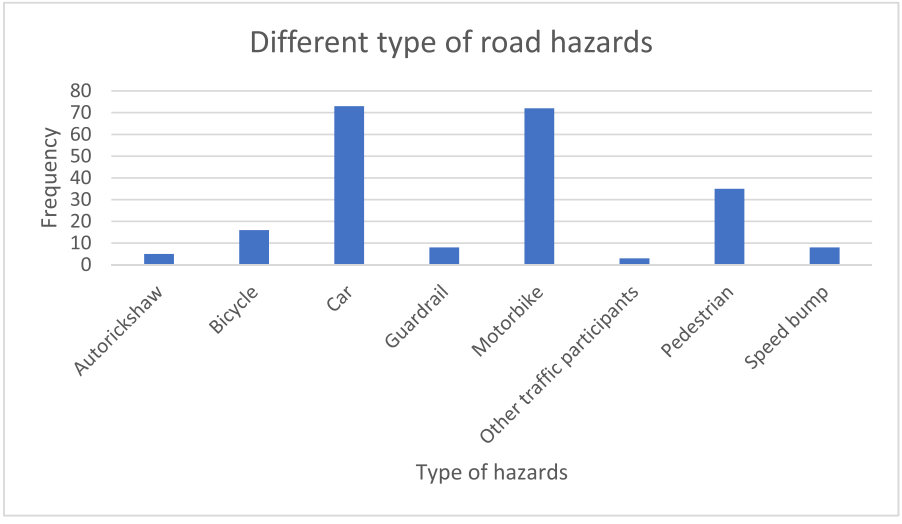

Fig. 16a. Details of different type of developed road hazards tagged in 13 videos.

addition to eye gaze movement. In fact, our cognitive load monitoring system can also detect distraction tracking only eye gaze [38]. However, to emphasize the utility of the machine learning models in terms of distinguishing amongst fine grain differences in drivers' cognitive load, we attempted to detect change in cognitive load when drivers perceive a road hazard. Perception of a road hazard does not involve head or eye gaze movement like operating head down display.

We followed the guidelines of Driver and Vehicle Standards Agency (DVSA), UK [16] and earlier research work [12] to identify developing road hazard in the first-person video of drivers. We tagged timestamps of all oncoming road hazard following the DVSA guidelines. Fig. 16a shows a comparative chart between different type of road hazard (vehicles, pedestrians, speed bumps, animals) for the set of driving samples used in the proposed system. We calculated L1NS, STDP, LPF, saccade rate, fixation rate and median SI velocity in time window duration of ± 2 s, ± 3 s, ± 4 s, and ± 5 s around the instances of each developing hazard [7]. The working of our proposed multi-class NN based cognitive load monitoring system is illustrated in Fig. 13.

Table 11
Accuracy of neural network in classifying distraction of driver for different label of hazard durations.

	± 2 secs	± 3 secs	± 4 secs	± 5 secs
Training	91.95%	94.62	92.47	84.52%
Test	71.15%	72.44%	70.50%	70.51%

We modified the NN model into multi-class classifier. We have trained and tested with 4 different datasets where hazard durations were marked in above-mentioned time windows. Table 11 shows accuracy of multi-class classifier in classifying different cognitive state of driver in different driving situations. Accuracy is calculated as sum of correct classifications divided by total number of classifications $((T_A + T_B + T_C)$ /

number of testing example), where T_A: TP of class A, T_B: TP of class B, and T_C: TP of class C). We found our model was able to classify 28 events out of 39 test events correctly (accuracy of 72.44 %) with ± 3 secs of time window corresponding to road hazard.

8. General discussion

In this paper, we proposed and evaluated the performance of pupil and gaze-based metrics from frequency and temporal domains for estimating cognitive load. Our pupil-based metrics from temporal domain has potential for deployment in real-time as it is robust to ambient light. L1NS algorithm uses frequency domain calculation and can be deployed in hardware supporting FFT computation. STDP algorithm does not require a frequency transform as we calculate the metric in time domain which allows deployment using a microcontroller. LPF algorithm works on frequency specific operations in time domain and can also be deployed using a microcontroller. We also demonstrated cognitive load estimation using our gaze-based metrics, which can be computed from any low-cost eye gaze tracker which does not support pupil diameter measurement. Our metrics are validated for their performance in simulated driving environment as well as real cars. Our gaze-based metrics did not show significant performance in psychometric tests. This might be due to the design of experiment which constrained participants to look into a region on wall such that the eye tracker did not miss detecting eyes. Though we requested participants to focus on the auditory task and not to focus on the region, it might have restricted the eye gaze movements which could have influenced the performance of gaze-based metrics in distinguishing cognitive states. We demonstrated real-time detection of cognitive state in real cars [sup 4]. Our metrics L1NS and LPF showed relatively high performance in N-back/arithmetic tests. Similarly, L1NS, STDP, LPF, fixation rate, saccade rate showed relatively high performance in simulator study and fixation rate showed relatively high performance in real car study. We observed that finding a single threshold value of metrics for all drivers is challenging. In this direction, we designed and tested with different type of ML based classification models which take multiple metrics as inputs to classify cognitive state and achieved significant increase in the classification accuracy. We were able to classify driver's cognitive state between driving task and secondary task with 75% accuracy using feed forward NN model.

8.1. Limitations and future work

We carried out our studies with a thin range of age groups and sample size. Though our studies showed significant increase in performance of our system, there is still room for improvement in terms of accuracy of classification. We are planning to investigate new ML models with and without memory for accommodating the associativity within data. Due to the influence of culture, gender and geographical location of a driver on driving behaviour and cognitive load, the sample size can be further expanded to increase the boundaries of inclusion. Following the investigation of binary classification of two states ('Task' and 'No Task'), we are planning to investigate multi-class classification of different tasks (music, talk, phone) and road hazards. We are analysing associativity between oncoming road hazards and pupil-based metric for a driver. In our future work, we will measure performance of neural network-based model for different road hazards with the help of Hazard Perception Test following the guidelines of Driver and Vehicle Standards Agency (DVSA) [DVLA] and earlier research work [12]. We are planning to use the classified states to give multimodal alert to drivers. We are also planning to evaluate our techniques with an automotive compliant camera-based eye tracking system.

8.2. Multimodal alert system

We are working on a multimodal alert system which can detect eyes-off-road distraction and cognitive states of driver to alert with haptic, auditory and visual feedback. We demonstrated our multimodal alert system in supplementary video [sup3]. When an eyes-off-road event is detected, it will alert the driver by an auditory sound followed by a voice note telling "please concentrate on driving". An LED (Light Emitting Diode) strip lights up with a blinking pattern to alert the driver visually. If it detects that the driver is undergoing distraction, it will lock secondary tasks from being operated by driver. In case a call or message is received, the system does not pop up any notification to the driver. The system will display only important IVIS functions on the screen. The icon size and colours will adapt according to the cognitive state of driver. When driver's cognitive load decreases, it will retract notifying driver about his missed calls and messages. The driver will then have access to operate locked secondary tasks. The working of alert system is illustrated in Fig. 16b.

8.3. Application to autonomous vehicle

Our cognitive load estimation system can also be integrated to autonomous vehicle and can enhance safety of the vehicle. Even if the vehicle is autonomous, there may be situation requiring the human driver to intervene either from inside the vehicle or from a remote location. In both cases, it would be necessary for the driver to stay alert and aware of the situation. The on-board obstacle detection system can be augmented with visual perception and cognitive load of the front seat passengers and can avoid accident if the human driver responses to a traffic participant even if the machine vision system fails to do so. Measuring passengers' cognitive load can be used as a metric of ride quality and driving parameters and routes can be adjusted based on that.

8.4. Application in comparing HMI

We developed an HMI (Human Machine Interface) evaluation tool (Fig. 17) to record and analyse the cognitive load of a user performing any task which can be used for comparative studies. This tool calculates the ocular metrics and streams in the form of live graphs along with videos of eyes and scene camera. We can add participants as well as events and record data corresponding to selected participant and event. We can compare the performance of participants in corresponding events by analysing corresponding CSV (Comma Separated Values) files. We can compare different tasks in a single HMI or same task in different HMI by comparing cognitive load estimated by our system.

8.5. Value addition

Though researchers [6] investigated several ML models for classifying cognitive states in controlled environments (laboratory or simulation), they found it challenging to implement their systems in uncontrolled environments. We have investigated the performance of our proposed system in an uncontrolled driving environment (real driving). Binias [6] used EEG data of participants to classify cognitive state of pilots in flight simulator. The implementation of such system is challenging as EEG is sensitive to body movements. Babu [3] and Davis [14] estimated cognitive load in actual flights from physiological parameters but there are not many similar studies in automotive domain beyond measuring glance duration [32]. Our proposed cognitive load monitoring system takes the advantage of using ocular parameters to estimate cognitive load and can be deployed in uncontrolled environments. Our proposed system is independent of head movement, ambient lighting and vibration in the environment. These advantages show the potential of our system in applications outside automotive like aviation, biomedical, pedagogy and many other scientific fields.

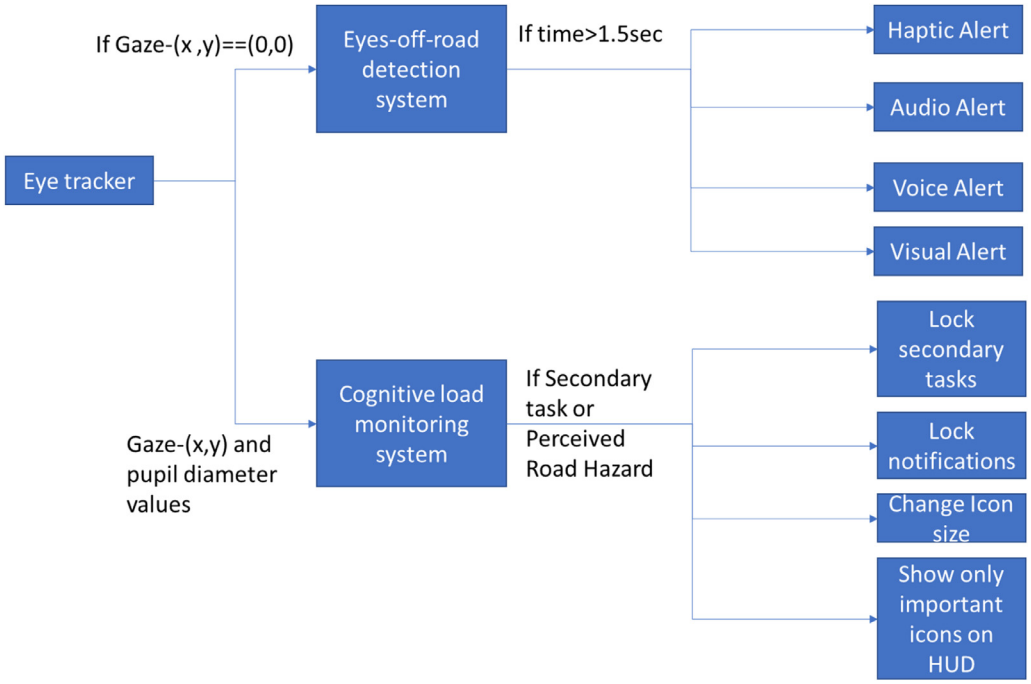

Fig. 16b. Working of alert system.

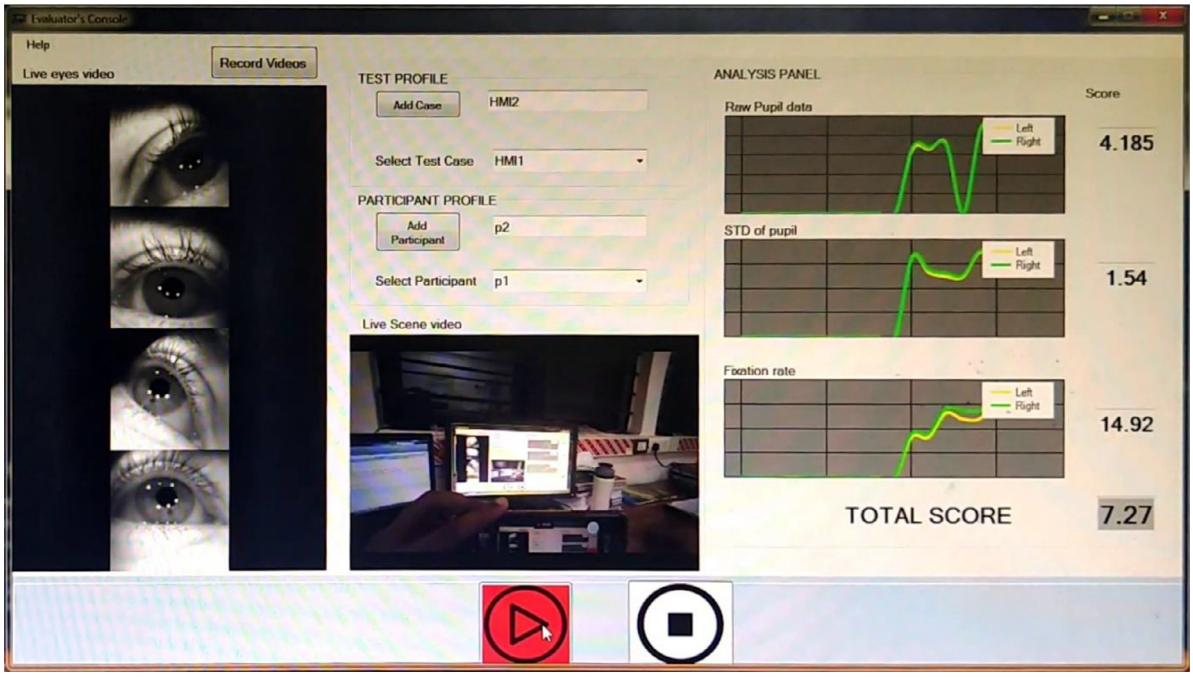

Fig. 17. Snapshot of HMI evaluation dashboard.

9. Conclusion

This paper presents a driver behaviour monitoring system based on physiological parameters in real driving situation under various ambient lighting conditions and traffic participants. We found that our proposed ocular metrics were able to distinguish between different cognitive states corresponding to task difficulties irrespective of changes in ambient lighting conditions in standard psychometric study. To classify differences in cognitive states due to operating secondary task, we compared various gaze and pupil dilation-based metrics and different Machine Learning models. We found that our ocular metrics were able to distinguish between differences in cognitive states corresponding to driving with and without undertaking secondary task. We also found that a feed forward Neural Network model outperformed individual metrics and other Machine Learning models with 75% accuracy (classifying between 'No Task' and 'Task'). Our results show that ocular parameters and Machine Learning models like Neural Network can be deployed for monitoring driver's behaviour in uncontrolled environments. Though we achieved a significant improvement in performance of binary classification by Neural Network-based models compared to other Machine

Learning models, we found rooms for improvement in the expansion of dataset and inclusion of wide range of age groups. We are planning to investigate further levels of 'Task' like operating music, maps, calls, Air Conditioner and perceiving a road hazard. This will be an extension of our work from binary to multi-level classification using Machine Learning models. We are also planning to improve the accuracy of classification by investigating Neural Networks with memory [25] which can accommodate patterns due to associativity within data. We are now investigating on constraints to integrate cognitive load monitoring system to an alert system which will decide when to enable/disable a secondary task and alert the driver based on amount and type of distraction.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.treng.2020.100008.

References

[1] R.V. Abadi, E. Gowen, Characteristics of saccadic intrusions, Vis. Res. 44 (23) (2004) 2675–2690.

[2] S. Afzal, P. Robinson, Natural affect data – collection & annotation in a learning context, in: Proceedings of Affective Computing & Intelligent Interaction (ACII), Amsterdam, 2009.

[3] M.D. Babu, J.D. Shree, G. Prabhakar, K.P.S. Saluja, A. Pashilkar, P. Biswas, Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies, J. Eye Mov. Res. 12 (3) (2019).

[4] O. Basir, J.P. Bhavnani, F. Karray, K. Desrochers, (2004). Drowsiness detection system, US 6822573 B2.

[5] J. Beatty, B. Lucero-Wagoner, The pupillary system, Handbook of Psychophysiology, 2, 2000.

[6] B. Binias, D. Myszor, K.A. Cyran, A machine learning approach to the detection of pilot's reaction to unexpected events based on EEG signals, Comput. Intell. Neurosci. (2018) 2018.

[7] P. Biswas, G. Prabhakar, Detecting drivers' cognitive load from saccadic intrusion, Transp. Res. Part F: Traffic Psychol. Behav. 54 (2018) 63–78.

[8] E. Bjäreholt, github page. 2014. https://github.com/ErikBjare/N-Back/.

[9] H. Boril, S.O. Sadjadi, J.H.L. Hansen, UTDrive: emotion and cognitive load classification in-vehicle scenarios, Proceeding of the 5th biennial workshop on DSP for in-vehicle systems, 2011.

[10] M. Braun, F. Weber, F. Alt, (2020). Affective automotive user interfaces–reviewing the state of emotion regulation in the car. arXiv preprint arXiv:2003.13731.

[11] C. Coulacoglou, D.H. Saklofske, in: Psychometrics and Psychological assessment: Principles and Applications, Academic Press, 2017, pp. 91–130.

[12] D. Crundall, P. Chapman, S. Trawley, L. Collins, E. Van Loon, B. Andrews, G. Underwood, Some hazards are more attractive than others: drivers of varying experience respond differently to different types of hazard, Accident Anal. Prevent. 45 (2012) 600–609.

[13] M. Dalmaso, L. Castelli, P. Scatturin, G. Galfano, Working memory load modulates microsaccadic rate, J. Vis. 17 (3) (2017) 6 -6.

[14] I. DAVIS, Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions, Aviat. Space Environ. Med. (1994).

[15] A.T. Duchowski, C. Biele, A. Niedzielska, K. Krejtz, I. Krejtz, P. Kiefer, M. Raubal, I. Giannopoulos, The Index of pupillary activity measuring cognitive load vis-à-vis task difficulty with pupil oscillation, CHI 2018.

[16] [DVSA] Free Hazard Perception Test. Available: https://www.theory-test-online.co.uk/free-hazard-perception-test-demo, Accessed on 25/04/2020.

[17] L. Fridman, B. Reimer, B. Mehler, W.T. Freeman, Cognitive load estimation in the wild, Proceeding of the Conference on Human Factors in Computing Systems, 2018.

[18] X. Gao, H. Yan, H.J. Sun, Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons, J. Vis. 15 (3) (2015) 3 -3.

[19] R. Gavas, D. Chatterjee, A. Sinha, Estimation of cognitive load based on the pupil size dilation, in: Proceeding of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), AB, Banff, 2017, pp. 1499–1504. 2017.

[20] E. Granholm, R.F. Asarnow, A.J. Sarkin, K.L. Dykes, Pupillary responses index cognitive resource limitations, Psychophysiology 33 (4) (1996) 457–461.

[21] J.A. Healey, R.W. Picard, Detecting stress during real-world driving tasks using physiological sensors, IEEE Trans. Intell. Transp. Syst. 6 (2) (2011) 156–166.

[22] E.H. Hess, The Tell-Tale Eye, Van Nostrand Reinhold Company, 1975.

[23] St. John, Kobus M., D. A., J.G. Morrison, D. Schmorrow, Overview of the DARPA augmented cognition technical integration experiment, Int. J. Hum. Comput. Interact. 17 (2) (2004) 131–149.

[24] T. De Jong, Cognitive load theory, educational research, and instructional design: some food for thought, Instruct. Sci. 38 (2) (2010) 105–134.

[25] S.M. Kouchak, A. Gaffar, Detecting Driver Behavior Using Stacked Long Short Term Memory Network With Attention Layer, IEEE Transactions on Intelligent Transportation Systems, 2020.

[26] K. Krejtz, A.T. Duchowski, A. Niedzielska, C. Biele, I. Krejtz, Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze, PLoS ONE 13 (9) (2018) e0203629.

[27] T. Kundinger, P.K. Yalavarthi, A. Riener, P. Wintersberger, C. Schartmüller, Feasibility of smart wearables for driver drowsiness detection and its potential among different age groups, Int. J. Pervasive Comput. Commun. (2020).

[28] Y. Lee, L.N. Boyle, Visual attention in driving: the effects of cognitive load and visual disruption, Hum. Factors. (2007).

[29] Y. Liang, J.D. Lee, A hybrid bayesian network approach to detect driver cognitive ditraction, Transp. Res. Part C 38 (2014) 146–155.

[30] S. Marshall, The index of cognitive activity: measuring cognitive workload, in: Proceedings of the 7th Conference on Human Factors and Power Plants, 2002 7–5.

[31] S. Marshall, Identifying cognitive state from eye metrics, Aviat. Space Environ. Med. 78 (Suppl. 1) (2007) B165–B175.

[32] Visual-Manual NHTSA, Driver distraction guidelines for in-vehicle electronic devices: notice of proposed federal guidelines, Fed. Regist. 77 (37) (2012) 11199–11250.

[33] D.C. Niehorster, R.S. Hessels, J.S. Benjamins, (in prep). GlassesViewer:boril Open-source software for viewing and analyzing data from the Tobii Pro Glasses 2 eye tracker 2019.

[34] F. Onorati, R. Barbieri, M. Mauri, V. Russo, L. Mainardi, Reconstruction and analysis of the pupil dilation signal: application to a psychophysiological affective protocol, in: Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, July, pp. 5–8.

[35] O. Palinko, A.L. Kun, A. Shyrokov, P. Heeman, Estimating cognitive load using remote eye tracking in a driving simulator, in: Proceedings of the Symposium on Eye-Tracking Research & Applications, 2010, pp. 141–144.

[36] G. Prabhakar, P. Biswas, Eye gaze controlled projected display in automotive and military aviation environments, Multimodal Technol. Interact. 2 (1) (2018) a.

[37] G. Prabhakar, N. Madhu, P. Biswas, Comparing pupil dilation, head movement, and eeg for distraction detection of drivers, in: Proceedings of the 32nd International BCS Human Computer Interaction Conference, 32, 2018, pp. 1–5.

[38] G. Prabhakar, A. Ramakrishnan, L.R.D. Murthy, V.K. Sharma, M. Madan, S. Deshmukh, P. Biswas, Interactive gaze and finger controlled HUD for cars, J. Multimodal User Interface 14 (2019) 101–121.

[39] E. Redlich, Ueber ein eigenartiges Pupillenphänomen; zugleich ein Beitrag zur Frage der hysterischen Pupillenstarre, Deutsche medizinischeWochenschrift 34 (1908) 313–315.

[40] D. Schnelle-Walka, S. Radomski, Automotive multimodal human-machine interface, in: The Handbook of Multimodal-Multisensor Interfaces, Association for Computing Machinery and Morgan & Claypool, 2019, July, pp. 477–522.

[41] T.M. Sezgin, P. Robinson, Affective video data collection using an automobile simulator, in: Proceedings of the International Conference of Affective Computing, 2007.

[42] E. Siegenthaler, F.M. Costela, M.B. McCamy, Di Stasi, J. Otero-Millan, A. Sonderegger, ..., S. Martinez-Conde, Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes, Eur. J. Neurosci. 39 (2) (2014) 287–294.

[43] J. Sweller, Cognitive load during problem solving: effects on learning, Cogn. Sci. 12 (2) (1988) 257–285.

[44] J. Sweller, J.J. Van Merrienboer, F.G. Paas, Cognitive architecture and instructional design, Educ. Psychol. Rev. 10 (3) (1998) 251–296.

[45] Tobii Pro Glasses 2 Product Description. 2018. Retrieved October 8, 2018 from https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-glasses-2-product-description.pdf/?v=1.95.

[46] S. Tokuda, G. Obinata, E. Palmer, A. Chaparo, Estimation of mental workload using saccadic eye movements in a free-viewing task, in: Proceedings of the 23rd international conference of the IEEE EMBS, 2011, pp. 4523–4529.

[47] A. Westphal, Ueber ein im katatonischen stupor beobachtetes Pupillenphänomen sowie Bemerkungen über die Pupillenstarre bei Hysterie, Deutsche medizinische Wochenschrift 33 (1907) 1080–1084.

[48] Y. Yoshida, H. Ohwada, F. Mizoguchi, Classifying cognitive load and driving situation with machine learning, Int. J. Mach. Learn. Comput. 4 (3) (2014).

[49] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A Survey of affect recognition methods: audio visual & spontaneous expressions, IEEE Trans. PAMI 31 (1) (2009) 39–58.