2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
June 28-30, 2023. Seattle, Washington, USA

Model-Based Estimation of Mental Workload in Drivers Using Pupil Size Measurements

Prarthana Pillai*, Balakumar Balasingam*, *Senior Member, IEEE*, and Francesco N. Biondi†

***Abstract*—Passenger vehicles are increasingly adopting the use of automated driving systems (ADS) to help ease the workload of drivers and to improve road safety. These systems require human drivers to constantly maintain supervisory control of the ADS. For safe adoption and ADS, the attention or alertness of the driver needs to be continuously monitored. Past studies have demonstrated pupil dilation as an effective measure of cognitive load. However, the raw pupil data recorded using eye trackers are noisy which may result in poor classification of the cognitive load levels of the driver. In this paper, an approach to reduce the noise raw pupil size data obtained from eye trackers used by ADS is proposed. The proposed approach uses a Kalman filter to filter out high-frequency noise that arises due to sudden changes in ambient light, head/body movement, and measurement noise. Data collected from 16 participants were used to demonstrate the performance of the model-based pupil-size filtering approach presented in this paper. Results show an objective improvement in the potential to distinguish changes in pupil size due to various levels of cognitive workload experienced by participants.**

***Index Terms*—Cognitive load detection, Expectation-Maximization algorithm, Eye-tracking, Human-computer interface, Kalman filter, Physiological signals, Pupil size, and State-space model.**

I. Introduction

The present-day Automated Driving Systems (ADS) are mostly in the ranges of level 2 or 3 (L2/L3). In L2/L3 autonomous systems, [1], the ADS is capable of performing certain driving tasks, such as, keeping the vehicle on a lane and maintaining its peed to preset levels by the driver; the human driver still bears the ultimate responsibility of the vehicle and needs to always maintain a supervisory control of the vehicle [2]. However, ADS failure is shown to predominantly occur due to driver distraction [3]. These observations point to the need to have an effective Driver Monitoring Systems (DMS) that can monitor the alertness level of the driver and warn them about impending dangers.

Cognitive load detection has become an actively researched arena in the automobile industry in the past few years [5]. Unlike manual and visual load, cognitive load is difficult to be measured through direct and non-invasive measurements. Researchers suggest empirical methods to estimate cognitive load [6]. These methods involve estimating the cognitive load by collecting subjective data using rating scales [7], performance data using primary and secondary task techniques [8]–[10], behavioural data using response times to tactile or auditory stimulus [11], and physiological data using physiological measurements [12]. Heart rate, heart rate variability, reaction time and pupil size are some examples of physiological measures that can be used to estimate the cognitive load of the driver.

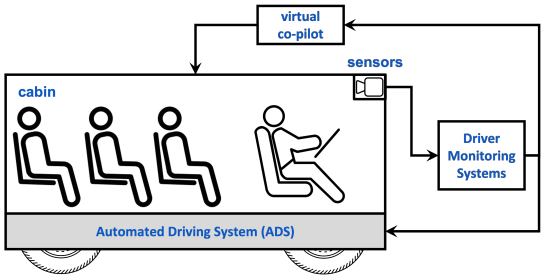


Fig. 1: **Automated Driving Systems.** Flowchart depicting cognitive load detection and response flow in ADS [4].

Eye-tracking measures are considered an alternative for cognitive load detection in ADS as they can be recorded non-invasively [13]. Several eye-tracking metrics, such as the pupil dilation [14], eye-gaze patterns [4], and eye-blink patterns [15], can be utilized to quantify the cognitive load experienced by the driver. It has been shown that pupillary reflex dilation occurs as a result of sensory and motor movements (tactile, auditory or gustatory), and mental and emotional efforts [13]. Past studies have shown that the mean and variance of the pupil size increases with cognitive difficulty [16] and that eye-tracking can be used to detect the changes in pupil size for different conditions of cognitive difficulty [17]. Therefore, with accurate measurement and classification of cognitive load conditions using eye-tracking measures, more reliable ADS can be developed.

Studies involving ADS employ low-cost portable cameras such as webcams and infrared cameras that can be easily installed in vehicles to continuously record pupil size at high sampling rates [18]. However, when it comes to eye-tracking in a driving environment, numerous factors contribute to noise and uncertainty in the obtained pupil size measurements [19]. Pupil size is obtained by applying signal and image processing algorithms on the image of the person captured by a video/infrared camera; the performance of these algorithms is affected by the quality and resolution of the images. In-vehicle eye trackers are likely mounted on a fixed platform such as

Submitted to *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, May 2023.

*The authors are with the Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Windsor, Ontario, N9G 3P4, Canada. E-mail: {pillaip,singam}@uwindsor.ca, Contact TP: +1(519) 253-3000 ext. 5431, Fax: +1(519) 971-3695

†The author is with the Faculty of Human Kinetics, University of Windsor, 401 Sunset Avenue, Windsor, Ontario, N9G 3P4, Canada. E-mail: {fbiondi}@uwindsor.ca, Contact TP: +1(519) 253-3000 ext. 2444.

the dashboard; the exposure to the eyes may vary as the driver moves their head and may result in measurement noise. This paper presents signal processing approaches for filtering pupil size data recorded using an infrared low-cost eye-tracker for drivers on a driving simulator. The proposed filtering method can therefore be extended to estimate the drivers' cognitive load in practical ADS.

This paper is organized as follows: An overview of the experimental setup, measurement devices used, participants, and the data collection procedure is briefed in Section II. Section III and Section IV contain the descriptive and inferential analysis for the classification into three conditions of the experiment for raw and normalized pupil data respectively. Further, this section introduces signal-to-noise ratio (SNR) as a measure of distinguishing varying levels of cognitive load of drivers from pupil size measurements. Section V describes a model-based filtering approach to improve the SNR for the classification of cognitive load. Section VI introduces some approaches for automated vetting data for model training and classification. Section VII provides the results of the new filtering approach in terms of SNR improvement and filtered pupil size plots for all data collected during the experiment. Finally, Section VIII concludes the paper.

II. Experimental Details

In order to demonstrate driver distraction detection, experimental data was collected from 16 participants while they drove on a medium-fidelity driving simulator. Sixteen participants in age ranges from 19 to 32 years (M = 24, SD = 3) were recruited for this study. All the participants were from the student and staff population at the University of Windsor. All participants were required to have a valid Ontario G2 license [20] or equivalent for at least two years with no fault in driving record for one year. In order to emulate driver distraction, participants were asked to participate in a number-back (n-back) task [21] of different difficulty levels (0-back and 2-back.) In the 0-back task, participants had to repeat out loud the number they just heard. In 2-back task, participants had to repeat out loud two numbers previous to the number they just heard. For the first two numbers, no response was required. The duration of each N-back task was five minutes during which 114 numbers were announced and the participants response to the audio stimulus was recorded. The participants also performed a 'Control' condition where they performed only driving and DRT tasks without N-back. A medium-fidelity driving simulator software called OpenDS [22] was used in the experiment. A Logitech G29 driving wheel [23] with pedals was used in this experiment. The driving wheel was mounted on the table in front of the desktop and the pedals were below the table. In order to validate the cognitive difficulty experienced by the participants, a Detection Response Task (DRT) [24], [25] and the subjective measure, NASA-TLX [26] were used. The Gaze-Point (GP3) eye-tracking system [27] was used to collect pupil size, eye-gaze, fixations and eye-blinks data. The eye-tracker was mounted on the desktop along the eyeline of the participant to record the eye-tracking data while driving. In the remainder of this paper, we consider only the pupil size data collected through the above experiment for cognitive load classification. More details are available in [4].

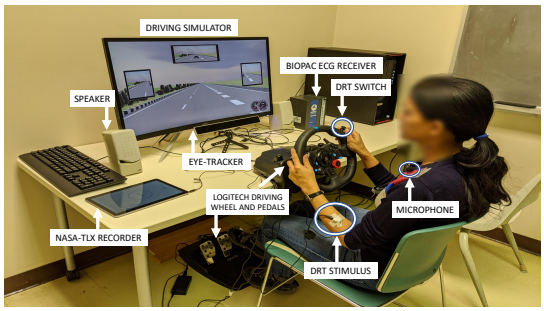


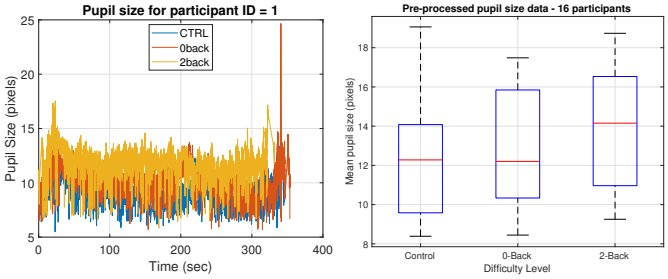
Fig. 2: **Experimental setup.** The apparatus and the setup used for all 16 participants are shown. ECG belt (not shown in the figure) was worn by the participant in contact with their skin forming a triangle around the heart [4].

III. Classification Using Raw Pupil Size Data

In Figure 3(a), the pupil size data from participant ID01 is shown for three difficulty conditions. The (collective) mean value of the pupil size and the (collective) standard deviation for each difficulty level are given by

$$\mu_c = 12.4704,\ \sigma_c = 2.7104 \tag{1}$$
$$\mu_0 = 12.9748,\ \sigma_0 = 3.1139 \tag{2}$$
$$\mu_2 = 13.9822,\ \sigma_2 = 3.1395 \tag{3}$$

Figure 3(b) shows the mean pupil of each participant as a box plot.



(a) The raw data corresponding to three difficulty conditions.

(b) Box plot of mean pupil size.

Fig. 3: **Pupil size.**

Statistical analysis [14] using the raw data concluded that the pupil diameter was significantly different for the different conditions during the experiment with $F(2, 30) = 12.105, p < 0.05, \eta_p^2 = 0.045$. Post-hoc analyses with a Bonferroni adjustment revealed that two of the 3 pairwise differences: {'control', '2-back' } and {'0-back', '2-back'} were significantly different with $p < 0.05$ whereas there was

no statistically significant difference in the {'control', '0-back' } pair.

The ability to distinguish data from two different groups, i and j, can be quantified using the *signal to noise ratio*, defined as

$$\mathrm{SNR}_{i,j} = 20\log\left(\frac{|\mu_i - \mu_j|}{\max\{\sigma_i, \sigma_j\}}\right) \quad (4)$$

where $i, j \in \{\text{'control'}, \text{'0-back'}, \text{'2-back'}\}$, $i \neq j$, and μ_i and σ_i are the mean pupil size and its standard deviation, respectively. The standard deviations σ_i and σ_j are assumed to be comparable in magnitude.

Using the above definition, the SNR between three different pairs is given as

$$\mathrm{SNR}_{\mathrm{c},0} = 20\log\left(\frac{|12.4704 - 12.9748|}{3.1139}\right) = -15.8\,\mathrm{dB} \quad (5)$$

Similarly, $\mathrm{SNR}_{\mathrm{c},2} = -6.3\,\mathrm{dB}$ and $\mathrm{SNR}_{0,2} = -9.9\,\mathrm{dB}$, Here, the subscripts $\mathrm{c}, 0$, and 2 are used to refer to 'control', '0-back', and '2-back' difficulty conditions, respectively. That is, $i, j \in \{\mathrm{c}, 0, 2\}$.

One can notice that the higher the SNR the more likely that different difficulty level pairs can be separated (classified) based on the observed mean pupil size. This is confirmed by the statistical analysis reported earlier in this section using the same data: that there is statistically significant difference between the {'control', '2-back'} ($\mathrm{SNR}_{\mathrm{c},2} = -6.3\,\mathrm{dB}$) and {'0-back', '2-back'} $\mathrm{SNR}_{0,2} = -9.9\,\mathrm{dB}$ pairs and there is no significant difference between {'control', '0-back'} pair which had the lowest SNR of $\mathrm{SNR}_{\mathrm{c},0} = -15.8\,\mathrm{dB}$.

Our goal in the remainder of this paper is to improve the SNR using various signal-processing approaches in order to enhance classification performance.

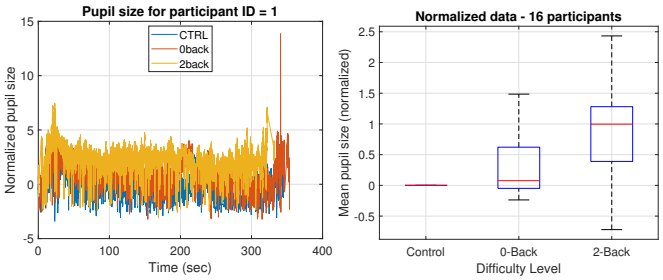
IV. Classification Using Normalized Pupil Size Data

The baseline pupil size might be different for each participant due to their physical features [28]. Hence, it is important to normalize these differences so that the data corresponding to each difficulty level can be compared during classification. We use the following formula to normalize data from each participant using data collected during each of their 'control' level experiment.

$$\tilde{p}_c = \frac{p_c - \mu_c}{\sigma_c}; \quad \tilde{p}_0 = \frac{p_0 - \mu_c}{\sigma_c}; \quad \tilde{p}_2 = \frac{p_2 - \mu_c}{\sigma_c} \quad (6)$$

where p_c, p_0, and p_2 denote the pupil size measurements corresponding to the 'control', '0-back', and '2-back' conditions, respectively and $\tilde{p}_c, \tilde{p}_0$, and $\tilde{p}_2$ refer the corresponding normalized data.

Figure 4(a) shows normalized pupil size data corresponding to the data shown in Figure 3(a) . Figure 4(b) shows the normalized mean pupil data as a box plot. Statistical analysis concluded that all three pairwise differences: {'control', '2-back'}, {'0-back', '2-back'}, and {'control', '0-back'} were statistically significantly different with $p < 0.05$.



(a) The normalized data correspond-ing to three difficulty conditions.

(b) Box plot of the normalized mean pupil size.

Fig. 4: **Normalized pupil size data.**

Table I shows the computed SNR values before and after normalization using the approach described in this section. It can be noticed that the SNR improved for all pairs, which conforms to the statistical analysis of the data, i.e., normalization improved the ability to classify different cognitive load conditions as indicated by the improved SNR. In the next section, the proposed approach to filter pupil size data is detailed.

TABLE I: **Comparison of SNR before and after normalization**

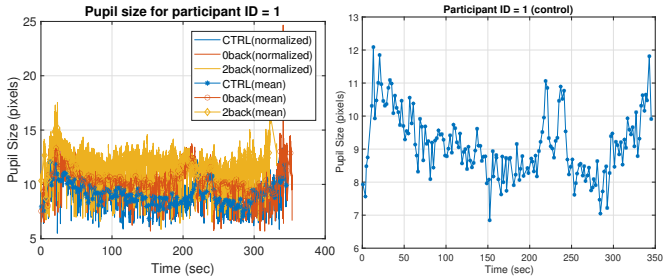
Difficulty pairs	Raw data	Normalized data
'control' vs. '0-back'	$-15.8\,\mathrm{dB}$	$-4.5\,\mathrm{dB}$
'control' vs. '2-back'	$-6.3\,\mathrm{dB}$	$0.9\,\mathrm{dB}$
'0-back' vs. '2-back'	$-9.9\,\mathrm{dB}$	$-2.6\,\mathrm{dB}$

V. SNR Improvement Through Model-Based Filtering

As seen by the definition of SNR (4), and through the statistical analysis and Table I, it is evident that the variance in the pupil size measurements is an important factor affecting the ability detect changes in pupil size, which is an indicator of cognitive load. If the variance in the pupil size measurements were to be low, it will lead to an increase in SNR and hence so does the ability to distinguish pupil size. This section presents a signal-processing approach to reduce the noise variance in pupil size measurement; the signal-processing approach is selected in a way that noise in pupil size measurement due to various high-frequency disturbances will be suppressed while the information pertaining to cognitive load will be retained.

Figure 3(a) shows the pupil size data in pixels collected at a sampling rate of 60 Hz over a 30-second period using a Gazepoint GP3 eye tracker. The accuracy of the recorded pupil data depends on many factors, including, the quality of the camera, the angle of camera orientation to the eyes, proximity, external light, and the robustness of estimation and detection algorithms against the head movement, to name a few. Due to this, the measured pupil size will be noisy. Further, it is assumed in this paper that the pupil size measurement noise is distributed zero-mean white Gaussian [29].

We saw in Section IV that the pupil size data needs to be normalized in order to remove discrepancies pertaining to individual physiological features. In order to further improve classification performance, we will focus on reducing the measurement noise through model-based filtering. First, considering that the measurement noise is white, an averaging method is employed to reduce measurement noise. Figure 5 shows the raw pupil size data (of Figure 3(a)) along with its averaged measurements over a 2-second window.



(a) (b)

Fig. 5: **Measurement noise reduction through averaging.** The raw pupil size data along with the averaged values over a 2-second block. Averaging serves as a pre-processing before applying model-based filtering on the data. The time length of each data block is also important — averaging over a longer window of data might suppress valuable pupil reflex information due to cognitive loading. One of the averaged data from Figure (a) is shown in (b).

It must be noted that the averaging should be done carefully to avoid losing sensitive information related to cognitive load. Figure 5(b) shows just one of the averaged data, corresponding to the 'control' level from Figure 5(a), in order to further illustrate this. It must be noted that the jump in pupil size during the 150-200 second duration is in response to realistic cognitive demand imposed by the driving simulator. Considering that cognitive load change is observed to change over the span of several seconds, it is decided to limit the data averaging window to a maximum of 2 seconds. Since the frequency of the eye-tracker used in the experiment is 60 Hz, the 2-second data window has, on average, 200 pupil size measurements. A brief inspection of Figure 5(b) shows that averaged measurements result in reduced variance. Indeed, it can be shown that [30], averaging results in the maximum likelihood estimate of the parameter (pupil size in this case) with reduced variance under the assumption that the measurement noise is distributed zero-mean Gaussian.

A closer look at Figure 5(b), one of the averaged measurements, gives the following observations related to the experiment: The increase in pupil size during 0-25 seconds is due to the participant starting the vehicle and familiarizing themself with the driving environment; the drop in pupil size during 25-150 seconds is due to incident-free driving on a straight highway; the increase during 150-175 seconds and then 350-375 seconds are due to encountered challenges on the driving path. Even though the 2-second averaged data in Figure 5(b) reasonably reflects the realistic cognitive load changes along the driving path, the data still has spiky features representing noise. From now on, we will refer to this as the measurement noise.

Remark 1. *The two second window for averaging is selected to make sure that high frequency noise that is typical due to body movement, cardiac activity, sudden changes in ambient light, e.g., change in light intensity caused by shades in traffic, are reduced due to averaging.*

Next, a state-space model is introduced to filter the pupil size data that already underwent a two-second averaging process described above. This model allows to be tuned in a way that pupil size of certain characteristics are retained and the effect of others reduced. First, the following state-vector is introduced

$$\mathbf{x}(k) = [x(k) \quad \dot{x}(k)]^T \tag{7}$$

where k denotes time index, $x(k)$ the pupil size in pixels at time k, and $\dot{x}(k)$ denotes the rate of change of pupil in pixels/seconds at time k.

The change of state vector from time instance k to time instance $k+1$ is modelled through the following process or plant equation [30]

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{\Gamma} v(k) \tag{8}$$

where

$$\mathbf{F}_k = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Delta^2/2 \\ \Delta \end{bmatrix} \tag{9}$$

and Δ denotes the sampling time. The *process noise* term $v(k)$ is modelled as a zero-mean Gaussian white noise with standard deviation σ_v. From this assumption, the *process noise covariance matrix* can be shown to be [30]

$$\mathbf{Q}_k = E\left[\mathbf{\Gamma} v(k) v(k)^T \mathbf{\Gamma}^T\right] = \begin{bmatrix} \frac{1}{4}\Delta^4 & \frac{1}{2}\Delta^3 \\ \frac{1}{2}\Delta^3 & \Delta^2 \end{bmatrix} \sigma_v^2 = \bar{\mathbf{Q}}_k \sigma_v^2 \tag{10}$$

The process noise variance σ_v^2 is one of the important design parameters for the proposed application in this paper. The process noise is used to model the unknown factor in how pupil size changes over time. Higher process noise indicates rapid changes in pupil size and vice versa. For example, let us inspect the pupil size changes in Figure 5(b). At first, the pupil size changed from 8 to 11 pixels in approximately 10 seconds, resulting in 0.3 pixels/second rate. During the next increase (from 175 sec. to 225 sec.) it took 50 seconds for an approximately 2-pixel increase, resulting in 0.04 pixels/seconds rate. At the last stretch, there is an increase of 3 pixels over 75 seconds, resulting in 0.04 pixels/second rate. From 10 s to 75 s pupil size dropped from 11 to 9 pixels, resulting in 0.04 pixels/second. Finally, there were large sectors of data where there are no changes at all (0 pixels/second). The quantity σ_v^2 represents the variance of all such change rates in pupil size.

In the context of the proposed application, an approximate value of the process noise variance can be obtained. The best way to come up with a reasonable value for the process noise variance is to learn from real-world eye-tracking data. Several experimentations with the data showed $\sigma_v = 0.01\,\text{pixels/s}^2$ to be a suitable value to represent typical variance in pupil size changes in averaged pupil size observations in a two-second window. In [31], this was confirmed through a machine-learning approach based on the Expectation-Maximization (EM) algorithm.

Let us denote the pupil size out of the two-second average process as $z(k)$; from this point, $z(k)$ will be treated as the 'measurement' that relates to the state vector defined in (7) as follows

$$\begin{aligned} z(k) &= \mathbf{H}\mathbf{x}(k) + w(k) \\ &= [1\ \ 0]\mathbf{x}(k) + w(k) \end{aligned} \tag{11}$$

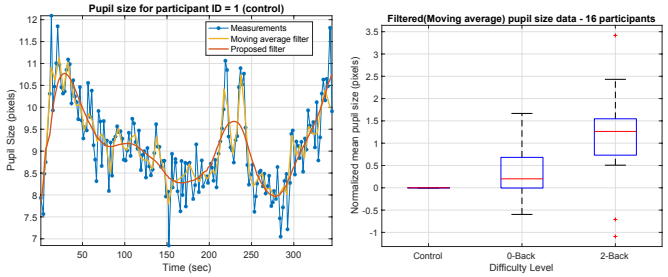
here, the *measurement noise* is modelled as a zero-mean white Gaussian noise with the following variance

$$\sigma_r^2 = E\left\{w(k)^2\right\} \tag{12}$$

Figure 5(b) shows a real-world example of the measurement noise $w(k)$ based on the data recorded from participant-1. This measurement noise variance σ_r^2 is dependent on various factors, including the quality of the recording device. A data-driven approach, based on the expectation maximization (EM) algorithm [31] is used to estimate the measurement noise parameter σ_r. Based on the data from all 16 participants, the standard deviation of the measurement error is found to be approximately $\sigma_r = 1$ pixels (see details in [31]).

The proposed state-space model to filter the pupil size consists of the process equation (8) and the measurement equation (11). The parameter σ_v of the process model is selected based on the characteristics of the cognitive load dynamics and the parameter σ_r of the measurement noise is estimated based on data-driven approaches as detailed above. Based on these discussions so far, the model parameters are found to be $\sigma_v = 0.01\,\text{pixels}/s^2$ and $\sigma_r = 1\,\text{pixels}$. In the remainder of the data analysis, the above two model parameters will be used. Once the model and its parameters are identified, the Kalman filter [30] will be used to recursively estimate (or filter) the state $\mathbf{x}(k)$ which has the desired pupil size as its first element.

Figure 6 demonstrates the performance of the proposed pupil size filtering approach using data from one of the participants. In Figure 6(a), the two-second averaged pupil size measurements (i.e., 'measurements') are shown using the marker '*'. The output of the proposed Kalman filter is also shown on the same plot in a solid line. As a comparison, Figure 6(a) also shows a regular averaging approach for the purpose of reducing the variance in the noisy pre-processed data. This smoothing-based approach to filtering the raw data uses a 5-point moving average. Later, the performance of the smoothing approach and the proposed model-based Kalman filtering approach are compared side by side and it was concluded that the proposed approach outperforms the moving average filter. Figure 6(b) shows the normalized mean pupil data filtered using the smoothing average filter as a box plot. This box plot will later be compared to the one generated through the Kalman filtering approach, shown in Figure 8.



(a) (b)

Fig. 6: **Filtered pupil size data.** (The averaged data in Figure 5(b) is sent through the model-based filter presented in Section V to obtain the filtered output shown in yellow. A 5-point average filter is also applied to the data from Figure 5(a).

VI. Automated Approach to Data Selection

Physiological data acquisition systems encounter an unusual amount of errors due to various types of disturbances. This section presents two automated approaches to detect such abnormal measurements so that they can be removed from further analysis.

A. Outlier Removal

A preliminary inspection of the pupil size data showed occurrences of negative values and outliers, i.e., unrealistic pupil size values. Based on visual inspection, it was decided to remove pupil size data that is below 5 pixels and above 25 pixels from further consideration. Each eye-gaze point data output from the Gaze point GP3 eye-tracker is accompanied by a flag of either '0' or '1'. It was described by the manufacturer that the flag values of '0', correspond to unreliable (less confident) estimates. Hence, only the gaze point data corresponding to flag '1' was extracted for further consideration. This type of outlier removal in pupil size data is in line with our previous studies conducted by the authors [14] and studies conducted by other researchers [32].

B. Missing Data Index

The outlier removal described in subsection VI-A resulted in loss of data. If the loss of data is too great for a given scenario (participant or condition) the entire dataset belonging to that condition was excluded form further analysis. The following data quality measure is used to decide wether or not to retain the data for further analysis

$$Q_{\text{data}} = \left(\frac{\text{number of retained data points}}{\text{total number data points}}\right) \times 100 \tag{13}$$

where $Q_{\text{data}} \in [0, 100\%]$. Here, higher the Q_{data} the better is the quality of the data.

C. Normalized Innovation Squares

The Kalman filtering process gives a way of quantifying the measurements by computing the normalized innovation square (NIS) [30] that is defined as

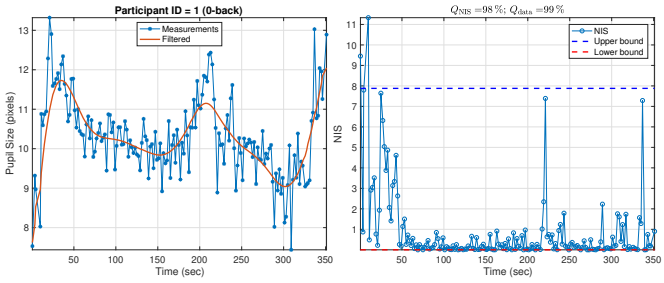
$$\text{NIS}(k) = \nu(k)^T S(k)^{-1} \nu(k) \tag{14}$$

where $\nu(k)$ is the innovation (the difference between the predicted measurement and the observed one) of the Kalman filter at time index k. When a Kalman filter processes data that conforms to the assumed model, the NIS values will stay within a specific limit [31]; when the Kalman filter is fed with measurements that do not conformed to the underlying model assumption, the NIS values jump out of their predetermined bound. Based on this observation, the following quality measure is defined

$$Q_{\text{NIS}} = \left(\frac{\text{number of NIS}(k) \text{ within the bound}}{\text{length of data}} \right) \times 100 \tag{15}$$

where $Q_{\text{NIS}} \in [0, 100\%]$. Once again, higher the Q_{NIS} the better is the quality of the data.

Figure 7 shows demonstrations of data quality indices using pupil size data from participants 1 and 7, respectively. In Figures 7, 98% of the NIS values were within the bounds — indicating that the data fit well to the hypothesized model in Section V. Further, 99% expected data were present, indicating that the eye-tracker functioned well and that the participant followed all the guidelines of the experiment.



(a) Measurements vs. filtering (b) Normalized innovation squares

Fig. 7: **Quality monitoring.** Pupil size data belonging to participant#1 is shown above. This data represents one of the best scenarios in the dataset: 99% of the expected data was present and the NIS was within the limit for 98% of the time.

VII. Results

This section provides an approach to remove data that suffered quality issues:

- *Data score.* For each dataset ('control', '0-back', and '2-back') of each participant (1-16), a score of '1' is given if Q_{data} is less than 75%.
- *Model score.* For each dataset ('control', '0-back', and '2-back') of each participant (1-16), a score of '1' is given if Q_{NIS} is less than 75%.

Using the above scoring system, each participant could receive a score of zero to six; zero indicating no quality issues and six indicating quality issues with all three datasets of a participant. Thus, datasets from participants 7, 10 and 16 were removed, resulting in data from 13 participants for further analysis pertaining to cognitive load classification.

Figure 8 shows the mean pupil sizes of the filtered data from the remaining 13 participants.

Fig. 8: **Filtered pupil size data.** The raw data corresponding to three difficulty conditions is shown. Each level lasted approximately 5 minutes and the red line depicts the median across the particular condition.

Table II shows the newly computed SNR using the filtered data (13 participants) in comparison to previous values shown in Table I. The SNR values computed for the pupil size data using the moving average filter are also added to this table for comparison with the proposed approach. It can be observed that the moving average filter did not result in an improved signal-to-noise ratio. The proposed model-based filtering using the Kalman filter approach resulted in significant improvement in all three difficulty pairs.

TABLE II: **Comparison of SNR after filtering and screening**

Pairs	C-0	C-2	0-2
Raw data	−15.8 dB	−6.3 dB	−9.9 dB
Normalized data	−4.5 dB	0.9 dB	−2.6 dB
Moving average filter	−4.3 dB	0.6 dB	−2.9 dB
Kalman filter	−2.4 dB	7.2 dB	4.3 dB

VIII. Conclusions and Discussion

In this paper, the problem of cognitive load classification in drivers using pupil size data is considered. Particularly, the pupil size data collected from a low-cost eye tracking device is evaluated for potential applications in Automated Driving Systems (ADS). This paper presented a signal processing approach that is designed to remove some form of high-frequency noise from the measured pupil size data while retaining possible changes in pupil size as a result of changes in cognitive load. The proposed approach consists of a linear state-space model; some of the parameters of this model are

estimated based on data-driven approaches whereas some other parameters were selected based on prior information about the nature of pupil size dynamics as a result of cognitive load. Table III contains the SNR values, computed according to (4), for raw, normalized and filtered pupil size data collected from 16 participants. It also contains the SNR values computed for the Detection Response Task (DRT) which is the ISO standard for cognitive load estimation. In addition, the Table III contains additional measures of cognitive load such as heart rate, $n-$back accuracy and eye-gaze based measures [4] calculated from data collected during this experiment. From this table, it can be observed that the SNR for normalized and filtered pupil size data is comparable to the ISO standard of DRT response times. Among the two physiological measures, pupil size and heart rate, SNR values were observed to be greater for pupil size. Although eye-gaze data can be used for better estimation of driver's cognitive load, pupil size has its advantages in that eye-gaze can be controlled by the participant whereas pupil size cannot be. Also, pupil size can be used in conjunction with other metrics to improve the accuracy of cognitive load estimation. Thus, the signal processing techniques introduced in this paper are applicable to improve individual datasets and are valid regardless of the sample size.

TABLE III: **Comparison of SNR based on different metrics of cognitive load.** The last three measures are eye-gaze metrics. [4]

Measure of cognitive load	$\mathrm{SNR}_{c,0}$	$\mathrm{SNR}_{c,2}$	$\mathrm{SNR}_{0,2}$
Response time	-4.99	0.06	-6.01
Pupil size (Raw)	-15.81	-6.35	-9.87
Pupil size (Normalized)	-4.5	0.9	-2.6
Pupil size (Filtered)	-2.4	7.2	4.3
Pre-processed heart rate	-24.06	-12.03	-6.92
$n-$back accuracy			8.83
Entropy in eye movements	10.41	11.43	1.25
Nearest neighbor index	2.26	0.13	-7.35
Entropy of gaze transitions	-12.31	4.84	-0.54

REFERENCES

[1] Y. S. Razin and K. M. Feigh, "Hitting the road: Exploring human-robot trust for self-driving vehicles," in *IEEE International Conference on Human-Machine Systems*, pp. 1–6, 2020.

[2] D. L. Fisher, M. Lohrenz, D. Moore, E. D. Nadler, and J. K. Pollard, "Humans and intelligent vehicles: The hope, the help, and the harm," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 56–67, 2016.

[3] C. Olaverri-Monreal and T. Jizba, "Human factors in the design of human–machine interaction: An overview emphasizing v2x communication," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 4, pp. 302–313, 2016.

[4] P. Pillai, B. Balasingam, Y. H. Kim, C. Lee, and F. Biondi, "Eye-gaze metrics for cognitive load detection on a driving simulator," *IEEE/ASME Transactions on Mechatronics*, pp. 1–8, 2022.

[5] V. Borisov, E. Kasneci, and G. Kasneci, "Robust cognitive load detection from wrist-band sensors," *Computers in Human Behavior Reports*, vol. 4, p. 100116, 2021.

[6] B. Xie and G. Salvendy, "Prediction of mental workload in single and multiple tasks environments," *International journal of cognitive ergonomics*, vol. 4, no. 3, pp. 213–242, 2000.

[7] E. Galy, J. Paxion, and C. Berthelon, "Measuring mental workload with the nasa-tlx needs to examine each dimension rather than relying on the global score: an example with driving," *Ergonomics*, vol. 61, no. 4, pp. 517–527, 2018.

[8] N. von Janczewski, J. Wittmann, A. Engeln, M. Baumann, and L. Krauß, "A meta-analysis of the n-back task while driving and its effects on cognitive workload," *Transportation research part F: traffic psychology and behaviour*, vol. 76, pp. 269–285, 2021.

[9] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," in *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–9, 2018.

[10] B. Mehler, B. Reimer, and J. A. Dusek, "Mit agelab delayed digit recall task (n-back)," *Cambridge, MA: Massachusetts Institute of Technology*, vol. 17, 2011.

[11] K. Stojmenova and J. Sodnik, "Detection-response task—uses and limitations," *Sensors*, vol. 18, no. 2, p. 594, 2018.

[12] M. Lohani, B. R. Payne, and D. L. Strayer, "A review of psychophysiological measures to assess cognitive states in real-world driving," *Frontiers in human neuroscience*, vol. 13, p. 57, 2019.

[13] R. Gavas, D. Chatterjee, and A. Sinha, "Estimation of cognitive load based on the pupil size dilation," in *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1499–1504, 2017.

[14] F. N. Biondi, B. Balasingam, and P. Ayare, "On the cost of detection response task performance on cognitive load," *Human factors*, p. 0018720820931628, 2020.

[15] S. Chen and J. Epps, "Using task-induced pupil diameter and blink rate to infer cognitive load," *Human–Computer Interaction*, vol. 29, no. 4, pp. 390–413, 2014.

[16] P. Ayare, *Cognitive Load Detection for Advanced Driver Assistance Systems*. PhD thesis, University of Windsor (Canada), 2019.

[17] P. Mannaru, B. Balasingam, K. Pattipati, C. Sibley, and J. Coyne, "Cognitive context detection for adaptive automation," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, pp. 223–227, 2016.

[18] D. Yi, J. Su, L. Hu, C. Liu, M. Quddus, M. Dianati, and W.-H. Chen, "Implicit personalization in driving assistance: State-of-the-art and open issues," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 3, pp. 397–413, 2019.

[19] X. He, L. Wang, and Y. Chen, "The study of conversion between the pupil diameter measured by eye tracker and the real value," in *IEEE 10th International Conference on Industrial Informatics*, pp. 508–511, 2012.

[20] "Ontario Ministry of Transportation Driver's Licence Classification Chart." http://www.mto.gov.on.ca/english/trucks/handbook/section1-1-2.shtml.

[21] "Delayed digit recall (N-back) task website." http://agelab.mit.edu/delayed-digit-recall-n-back-task.

[22] "OpenDS website." https://opends.dfki.de/.

[23] "Logitech website." https://www.logitechg.com/en-ca/products/driving/driving-force-racing-wheel.html.

[24] I. O. for Standardization, "Road vehicles transport information and control systems Detection Response-Task (DRT) for assessing attentional effects of cognitive load in driving," 2016.

[25] "DRT (Detection Response Task) website." https://www.redscientific.com/detection-response-task.html.

[26] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, pp. 139–183, 1988.

[27] "Gazepoint eye tracker website." https://www.gazept.com/.

[28] S. E. Kuchinsky, J. B. Ahlstrom, K. I. Vaden Jr, S. L. Cute, L. E. Humes, J. R. Dubno, and M. A. Eckert, "Pupil size varies with word listening and response selection difficulty in older adults with hearing loss," *Psychophysiology*, vol. 50, no. 1, pp. 23–34, 2013.

[29] S. Bottos and B. Balasingam, "Tracking the progression of reading using eye-gaze point measurements and hidden markov models," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7857–7868, 2020.

[30] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[31] P. Pillai, B. Balasingam, A. Jaekel, and F. Biondi, "Kalman filtering to track changes in pupil size for automated driving systems," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pp. 1–6, IEEE, 2021.

[32] M. E. Kret and E. E. Sjak-Shie, "Preprocessing pupil size data: Guidelines and code," *Behavior research methods*, vol. 51, no. 3, pp. 1336–1342, 2019.