Estimating Cognitive Load under Varying Light Intensity

A Novel Method for Quantifying Perceived Light Intensity for Cognitive Load Esimation

Chris Smit

MSc Mechanical Engineering
Vehicle Engineering
Perception and Modelling

Estimating Cognitive Load under Varying Light Intensity

A Novel Method for Quantifying Perceived Light Intensity for Cognitive Load Esimation

by

Christiaan Olivier Smit

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday September 25, 2020 at 14:00.

Student number:	4237684	
Project duration:	November, 2019 – September, 2020	
Thesis committee:	Dr. ir. J.C.F. de Winter,	Delft University of Technology, supervisor
	Dr. ir. R. Happee,	Delft University of Technology
	Ir. J.C.J. Stapel,	Delft University of Technology, supervisor
	M.Sc. F. Doubek,	Dr. Ing. h.c. F. Porsche AG, supervisor
	M.Sc. N. von Janczewski,	Dr. Ing. h.c. F. Porsche AG, supervisor

This thesis is confidential and cannot be made public until September 1, 2021.

Preface

Dear reader,

In front of you, you have what has been my main occupation for the past year. If I successfully defend this thesis on September the 25th of 2020, I will get to call myself a Master of Science. While it certainly feels like an overstatement to say that I then have mastered science, I have learned so much from this thesis and all of my time at the Delft University of Technology. I would like to take the opportunity to thank everyone who has made this possible.

First of all there is Joost, an inexhaustible source of knowledge who might recommend an interesting paper at 4 AM, or share a plot that shows a new insight he has gained on the data. Your commitment not only to the research, but also to your students means a lot to me. Jork, your encouraging and thorough feedback always made me excited to go work on the next problem. The both of you have always made it easy for me to reach out with problems, practical questions or new ideas. Your input has been of immense value to me. Thank you for pushing me to explore different perspectives and lifting this work to an academic level.

Fabian, thank you for sending me on this exciting adventure in Stuttgart. I have really enjoyed our hipster meetings in Holzapfel. You and Niko have done a great job at ensuring that the research could go on smoothly, despite the circumstances. Niko, your data of course was the foundation of my thesis, but of even larger value were the psychological insights you have to offer. Thank you both for always taking the time to answer my questions. Thanks for your supervision, but also the funny and personal moments we have shared.

2020 has turned out to be everything but ordinary. Due to the COVID-19 pandemic, my workplace quickly changed from the comfortable Weissach office to my parents' attic, until I found a place of my own. I'd like to thank my parents Nanne-Geert and Marian for enduring me moving back home, but mostly for supporting me throughout my entire studies. I hope you agree that your patience with me is now paying off.

Last, but not least, I thank my wonderful girlfriend Trine. I can always count on your support, whenever and wherever it may be, pandemic or not. Thank you for enduring me spending entire weekends and nights behind a laptop, trying to improve my classifiers by a minuscule amount. Your encouragements mean more to me than you know.

For much of this research I was a six hour drive removed from the Porsche facilities. While doing everything remotely sometimes brought its challenges, I am proud of the result. This thesis consists of a paper that has been submitted for the CHI 2021 conference and some appendices. As the CHI strongly encourages papers to be no longer than strictly necessary, the thesis is on the short and compact side. The appendices, submitted to the CHI as complementary material, should give it the substance of a master's thesis.

All the best,
Chris

Amsterdam, September 2020

Abstract

Measuring cognitive load is essential for understanding driver performance. Under- and overload can result in dangerous situations on the road. Cognitive load can be estimated by monitoring the diameter and movements of the pupils, but during measurements external influences such as changes in light intensity affect pupil diameters. In this paper, we present a novel method for quantifying light intensity with a head-mounted eye-tracker by weighting pixel values around the gaze direction. We demonstrate its effectiveness in cognitive load classification systems that use pupil metrics only. 54 participants in two separate studies have carried out n-back tasks during a simple driving task in a driving simulator. The data is classified by cognitive task (baseline, 1-back, 2-back) with the Random Forest algorithm. The resulting systems are 92.5% accurate with and 85.9% accurate without gaze features available, but are unable to generalise to participants unseen in the training phase of the algorithm.

Contents

1
Introduction

Cognitive load is an important measure in the quest for understanding driver performance. A cognitive over- or underload is associated with reduced cognitive performance, potentially leading to dangerous situations in traffic [1]. Tracking the cognitive load a driver experiences is a step in providing real-time feedback and improving driver assistance systems, ultimately contributing to a safer vehicle environment. In a vehicle environment, it is important to have a non-intrusive way of measuring cognitive load. Back in 1964, Hess and Polt already observed changes in pupil sizes related to cognitive load [2]. Later, in multiple studies, various other behaviours of the eye, including blinks, fixations and saccades, have been linked to an increased cognitive load [3]. Modern eye-trackers aid in observing such changes automatically, but several challenges remain today. One of these challenges is providing accurate measurements in situations where light intensity can vary, for instance in a vehicle. Like cognitive load, light intensity provokes pupillary responses. Where load-induced dilations are usually up to 0.5 mm, pupil diameters can vary from 2 to 8 mm under different light intensities. This effect of light can overshadow the effect of cognitive load. Therefore, the effects of light intensity need to be taken into account for accurate cognitive load estimation. Research has aimed to subtract light-evoked pupillary responses from the total dilations, so only a task-evoked pupillary response remains [4, 5]. These approaches require the light intensity to be constant during the interval of measuring cognitive load. To measure cognitive load automatically in an environment with varying light conditions, a different approach is needed. Wavelet transforms on the pupil diameter have been used to estimate cognitive load, regardless of the light conditions [6–8]. The underlying idea is that the pupil dilations induced by cognitive load occur at a different frequency than dilations induced by other external influences as light.

In this paper, we present a novel way to quantify light intensity and derive two novel features for use in cognitive load classification. We will do so by combining two studies into one dataset and examining them with machine learning. Section 2 summarises the research this work was built upon. In Section 3 we explain how the light intensity measure can be derived from images of a forward-facing camera and gaze marker coordinates. Next, in Section 4, we will describe two features of the perceived light intensity: the median light intensity and the Relative Change in Light Intensity (RCLI). Then we will demonstrate these features alongside other features in a 3-class problem analysed with a Random Forest classifier [9]. The classes are three tasks of varying cognitive load: baseline (just driving, low load), 1-back (on top of driving, medium load) and 2-back (on top of driving, high load). As a benchmark, we will present two systems of features that use known and well-researched measures of cognitive load; the median pupil diameter and the percentage of eye closure (PERCLOS) [10]. A novel measure of the variability of the pupil diameter is the Relative Change in Pupil Diameter (RCPD). The RCLI and the RCDP are used because the relative changes over time help paint a more complete picture, on top of the information median values provide. On top of these features, one of these systems will take the median gaze marker coordinates into account and the other will not. We will demonstrate the benefit of taking the light intensity features into account on top of the previously mentioned features. We show that with this approach a higher classification accuracy was reached than when using the wavelet transform techniques IPA and LHIPA [7, 8].

Scripts used in this research can be found on the online repository at: `https://github.com/C-O-Smit/estimating-cognitive-load`

2
Related Work

Cognitive load can be estimated automatically with machine learning algorithms and pupillometry. When the classifier of a machine learning algorithm is trained on recordings of situations where the cognitive load condition was known, it can compare unseen data to these instances. In a comparison between different algorithms, decision tree algorithms showed the advantage of a high performance and transparency of the algorithms' inner workings, allowing for intuitive identification of important predictive variables [11]. Fridman et al. presented two novel ways to estimate cognitive load in a driving environment with a 3-class classification approach, achieving 77.7% and 86.1% accuracy [12]. Both methods relied on gaze direction for the cognitive load detection, the former explicitly and the latter implicitly. The higher the cognitive load, the more concentrated the position of the eye and thus the gaze direction was. The more accurate approach was indirectly influenced by the light intensity, as it was a deep learning approach that used images of the eye directly as its input. The other method used a bivariate Hidden Markov Model with the pupil position and the blink state as input, not taking the light intensity into account.

The Index of Cognitive Activity (ICA) popularised wavelet transform approaches to cognitive load estimation [6]. The cognitive task-evoked dilations are separated from the light-evoked dilations with a wavelet transform, after which the results are quantified in a score called the ICA. The ICA was found to have a strong correlation with lane deviation in a driving simulator [13]. Rerhaye et al. found that a higher ICA was an indication for a higher workload during spatial processing, but also found inconclusive results when evaluating an inhibition task, the Stroop task [14]. They conclude that the validity of the ICA is highly task-specific and that measuring workload with the ICA remains questionable. Two open-source alternatives are the IPA (Index of Pupillary Activity) and the LHIPA (Low/High Index of Pupillary Activity) [7, 8]. While based on the same principle as the ICA, they use a different wavelet and different tresholding approaches. The IPA and more so the LHIPA were reported to be sensitive to changes in cognitive load, but not sufficiently so to distinguish between task difficulty levels.

The Percentage Change in Pupil Diameter (PCPD) is often cited as a reliable measure for cognitive load, generally increasing with an increase in load [15, 16]. The PCPD is calculated by dividing the current pupil diameter by the average pupil diameter of the participant during baseline conditions. A drawback is that the light intensity has to be constant during a recording. Gaze direction is also cited as an indication of cognitive load [17]. In a driving environment, the gaze direction is increasingly concentrated to the centre of the road as cognitive load increases. This is attributed to tunnel vision; the shedding of less essential tasks such as checking the mirrors or the dashboard instruments.

3

A Proposed Measure for Light Intensity

The Pupillary Light Response (PLR) is a function of the Corneal Flux Density (CFD, i.e. the product of luminance and subtended area) and the eccentricity [18, 19]. A measure for light intensity that aims to take the PLR into account should reflect this. The light intensity was derived out of images from the forward-facing camera in the vicinity of the gaze marker. The forward-facing camera recorded at 30 Hz, the gaze markers were derived from eye-facing cameras that recorded at 60 Hz. The light intensity feature was computed at 60 Hz where every video frame of the scene was matched with the two sets of gaze marker coordinates with the closest matching timestamps. The light intensity is measured as the weighted 8-bit RGB values of concentric, touching but not overlapping rings centred at the gaze marker, where the light intensity (LI) values per ring are weighted following a light intensity as used in the OpenCV software package: $LI_{ring} = 0.2989R + 0.5871G + 0.1140B$ [20]. These weights per colour channel are based on how sensitive the human eye is to the different channels. The rings each have a weight determined by their average eccentricity. Figure 3.1 shows a simplified example of the rings; in reality 11 rings with a diameter corresponding to an eccentricity of up to 15 degrees were used, as a circle of that eccentricity captures 99.8% of the potential light intensity humans can experience according to the relation described in Figure 3.2.

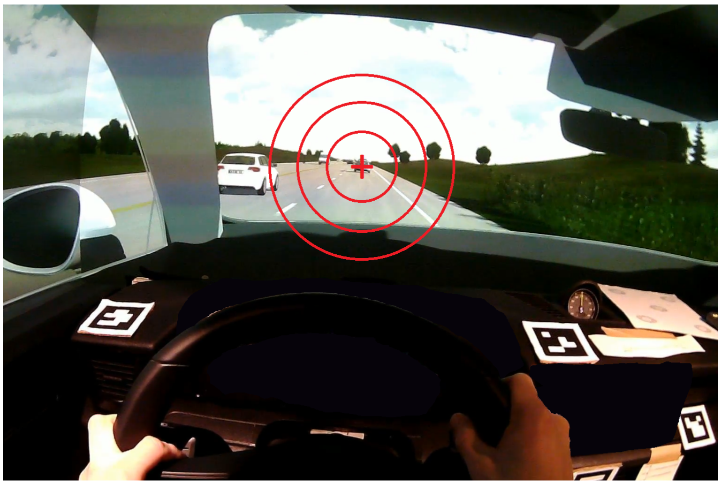


Figure 3.1: Illustration of the derivation of the light intensity, with the gaze marker depicted as a red cross, surrounded by rings in which each pixel has an equal contribution to the light intensity of that ring

The eccentricity was defined as the angle between the direction of the light that reaches the eye and the line of sight. The weight for the eccentricity was determined by fitting a Gaussian distribution to the empirical relation between perceived light intensity and the eccentricity as found by Wright and Nelson [21]. This research stems from 1936, but the work on how light refracts inside of the human eye is still deemed relevant in the 21st century [22]. This relation was determined by increasing the eccentricity of the light source and visually comparing the light intensity to a light source with an eccentricity of zero. The light source with an eccentricity of zero would be dimmed until its intensity was equal to that of the light source with non-zero

Table 3.1: Ring weights for the light intensity calculation

Outer diameter (°)	0.5	1	1.5	2	2.5	3.5	5	7.5	10	12.5	15
Ring weight	0.0050	0.015	0.024	0.033	0.040	0.10	0.18	0.28	0.20	0.093	0.034

eccentricity. After performing this procedure for a range of eccentricity levels, the standard deviation (SD) of the distribution was determined to be 5 degrees; the resulting relation is shown in Figure 3.2. From the vertical view angle of the camera and the height of the image in pixels, the eccentricity was converted to a radius in pixels using basic trigonometry. For every ring, the weight corresponding to the average eccentricity of the ring is taken; e.g. for the ring with an inner diameter of 2° and an outer diameter of 2.5° the average eccentricity is 2.25° with a corresponding weight of 0.144. To get the weight of the ring this number is multiplied with the area of the ring. The ring weights are then normalised so their sum is 1. For the 11 rings used in this research the weights are given in Table 3.1. The rings with the lowest outer diameters have a small area and therefore a low weight. The ring from 5° to 7.5° has the highest weight, a result of its area multiplied by the perceived light intensity for its average eccentricity. The outer rings have lower weights, as the perceived light intensity for their eccentricity is lower. The rings were chosen over a continuous function to decrease the computational complexity. The total light intensity is calculated by taking the sum of the light intensities per ring multiplied by their ring weights, as in Equation 3.1. If a ring partially falls of the screen due to the gaze marker being too close to the edge, the ring is not taken into account. The remaining rings' weights are adjusted so their sum remains 1. The result is a single light intensity value between 0 and 255 that is updated 60 times per second.

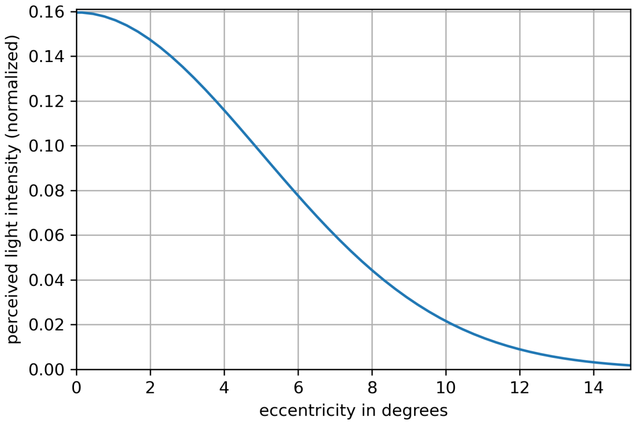


Figure 3.2: The perceived light intensity for varying eccentricity, a Gaussian distribution (SD = 5°) derived from Wright and Nelson [21]

From 0 to 8 degrees the relation between the light intensity and the eccentricity of Figure 3.2 is fairly consistent with the relation between foveal acuity and the eccentricity as described by Duchowski [23]. The foveal acuity was reported to be fairly constant within the central 2°, drop approximately linearly from there to the 5° foveal border and drop more sharply from there. A difference between the two distributions is that the Gaussian distribution describing the light intensity has tails where the intensity does not drop as sharply with the increasing eccentricity, where the foveal acuity diminishes more rapidly. Using a tailed distribution for the light intensity is physiologically plausible; where the foveal acuity is associated with photoreceptor cells called cones, the perception of light intensity is associated with both cones and other photoreceptor cells called rods. The cones are mostly found in the centre of the retina, the rods at its edge [24]. The rods have a maximum density around an eccentricity of 20 degrees. A tailed distribution also corresponds with the effects of light diffraction [25].

$$LI = \sum w_{ring} \cdot LI_{ring} \tag{3.1}$$

4

Estimating Cognitive Load

4.1. From time-series to features

The features of Table 4.1 were derived from the time-series of the light intensity, the pupil diameters and the gaze marker coordinates. All features were calculated over a 6 second time period, consistent with other research [12]. An equal amount of footage was cut from the start and the end of the recording, to make the length of the recording a multiple of 6 seconds. The remaining footage was cut into segments adjacent in time. For all features but the percentage of eye-closure (PERCLOS) the blinks were excluded and linear interpolation was applied to the gaps. The feature Relative Change in Pupil Diameter (RCPD) was used as the index of the variability of the pupil diameter. It is defined as the absolute sum of changes in pupil diameter in 6 seconds, divided by the mean diameter over the same 6 seconds, as in Equation 4.1. Similar to the RCPD the Relative Change in Light Intensity (RCLI) was defined as the absolute sum of changes in light intensity, divided by the mean over the same time interval, as in Equation 4.2. Changes in light intensity and changes in cognitive load both provoke changes in pupil diameter. By measuring light intensity and modelling its pupillary response, the confounding effects of light can be compensated, leading to a more accurate cognitive load estimation.

$$RCPD = \frac{\sum_{t=t_0}^{T} |\frac{\partial PD}{\partial t}|}{\mu(PD)} \tag{4.1}$$

$$RCLI = \frac{\sum_{t=t_0}^{T} |\frac{\partial LI}{\partial t}|}{\mu(LI)} \tag{4.2}$$

Out of the processed time series of pupil diameters the following four features were calculated per data point: the Median Pupil Diameter (MPD) of the left eye, the MPD of the right eye, the RCPD of the left eye and the RCPD of the right eye. While the features of the left and the right eye will be strongly correlated, sensitivity to light intensity is stronger in the non-dominant eye [26]. These metrics are therefore included for each eye individually. A fifth feature, PERCLOS averaged over both eyes, was calculated from the unprocessed diameter time series, before excluding the blinks. These five features are part of all four presented systems. From the time-series of the gaze marker x and y coordinates in pixels the medians were used as features. As the gaze marker coordinates are strongly task-dependent, two systems with these features and two systems without these features will be presented.

From the time series of light intensity, two features were calculated: the median and the RCLI. To investigate the added benefit of these measures two systems with and two systems without light intensity features will be presented. The systems without light intensity features will serve as a performance baseline. In total four systems will be presented; Table 4.1 provides an overview.

4.2. Classification with Random Forest

A random forest classifier was trained on the data [9]. The outcome of a random grid search decided the range in which the hyperparameters would be optimal [27]. The final hyperparameter tuning was done with an exhaustive grid search in that range. The hyperparameters were scored with a 10-fold cross-validation score. This by cutting the shuffled dataset into 10 equally sized portions. Ten classifiers were each trained on nine portions of the dataset and then tested on the remaining portion, so that every data point was trained on nine times and tested on once. The cross-validation score is the average testing accuracy of the 10 classifiers.

Table 4.1: The features used in the four presented cognitive load classification systems

Benchmark 1	System 1	Benchmark 2	System 2
left MPD	left MPD	left MPD	left MPD
right MPD	right MPD	right MPD	right MPD
left RCPD	left RCPD	left RCPD	left RCPD
right RCPD	right RCPD	right RCPD	right RCPD
PERCLOS	PERCLOS	PERCLOS	PERCLOS
median gaze marker X	median gaze marker X		median light intensity
median gaze marker Y	median gaze marker Y		RCLI
	median light intensity		
	RCLI		

As optimal settings a forest of 1000 trees with no bootstrapping was found, all other settings were standard as in the sci-kit learn library. These parameter settings were used for all systems.

We will also present another way of dividing the data between train and test sets in the 10-fold analysis, known as group K-fold. In this split, all the data of an entire participant is put in one of the 10 portions. The portions are of approximately equal size, but because the data of a participant must belong to one group in its entirety, the groups are not perfectly equal in size. The result is that the classifier is only tested on participants that are completely unseen in the training phase of the algorithm. The goal is to see how well the classifier generalises to unseen participants. The systems using group K-fold were trained on the three classes "baseline", "1-back" and "2-back".

5

The Dataset

In the training phase two independently recorded studies were used, which will be referred to as Study 1 and Study 2. The data of these studies was joined in a combined dataset that the classifiers were trained on. Neither of the studies has been recorded with the direct purpose of cognitive load classification. Statements made in this subsection apply to both studies, statements made in the subsection of the individual study apply to that study only. Participants were recruited through a newsletter for all employees of the Porsche Entwicklungszentrum in Weissach. Both studies were recorded to investigate in-vehicle interaction systems (IVIS), but in all recordings used in this research no IVIS tasks were being performed.

The pupil diameters, depicted in Figure 5.1, were found by taking the means and standard deviations of the features 'median right pupil diameter' and 'median left pupil diameter' per cognitive task. Error bars indicate the standard deviation. All data was recorded with Ergoneers head-mounted eye-trackers (Dikablis Professional or Dikablis Glasses 3) in a high-fidelity driving simulator with motion dynamics projecting a driving environment around a mock-up cockpit of a road-legal vehicle. The cockpit is placed on the eMove eM6-640-1800 by E2M Technologies, a 6-DOF moving base platform with an actuator stroke of 640 mm. The projection was provided by projectors with a resolution of 3840 by 2160 pixels and a refresh rate of 60 Hz projecting on the front, the side walls and the ceiling, creating a field of view of 180°. At the end of its service life the projector lamps brightness has reduced by 50%, lowering the light intensity in the simulator. Whilst driving secondary 1-back and 2-back tasks were both presented audibly for a duration of 2.5 minutes per task, as prescribed by ISO TS 14198:2019 standards. Four series of 10 digits were presented through an audio file, with 2.25 seconds spacing between the stimuli. Responses were given verbally, as prescribed by the MIT AgeLab [28]. Baseline data was collected during driving with no secondary task. The driving task for Study 1 and Study 2 consisted of following a lead-vehicle on the right lane of a multi-lane highway. NASA-TLX mental data was collected with both studies through a questionnaire after the cognitive task. Data where pupils were wrongly detected in eye-brows, glasses or other faulty locations was excluded. If over half of the segments contained faulty pupil detections, the whole recording was excluded. This was checked through manual inspection. Participants who reported any level of simulator sickness were excluded. They either mentioned simulator sickness themselves, or were asked about it if they showed symptoms.

For both eyes, the average pupil diameters were larger during a task associated with a higher cognitive load, likely caused by task-evoked pupil dilations. The median pupil diameters were also found to be significantly different per cognitive task by testing with an unpaired t-test, apart from the difference between baseline and 1-back for the left eye (t(1727) = -1.70, p = 0.09). Means and standard deviations are given in Table 5.1, sample sizes in Table 5.2.

The light intensity varied within the recordings, mainly because the participants would switch between looking at the instruments in the black dashboard of the cockpit and looking at the projected road. Checking instruments such as the speedometer is not directly related to the cognitive task, but to the driving task. We did not find a significant difference in the average median light intensity between 1-back and 2-back recordings (unpaired t-test, t(2146) = 0.45, p = 0.65), but for baseline recordings the light intensity was 9% lower on average (unpaired t-test between baseline and 1-back, t(1727) = 3.24, p = 0.001). This is caused by two things. The first reason is that the average light intensity for Study 1 is lower than for Study 2 and 49% of baseline recordings came from Study 1, where for 1-back and 2-back this number was 43% and 44% respectively. The higher ratio of darker conditions makes the average light intensity for baseline recordings lower than for the other classes. The second reason is that during the baseline recordings, participants look left at the overtaking traffic more often. The darker cockpit frame appears in the field of view, resulting in a lower median light intensity for the segment. With the higher cognitive load of 1-back and 2-back tasks, the gaze direction is more centred, yielding a higher median light intensity.

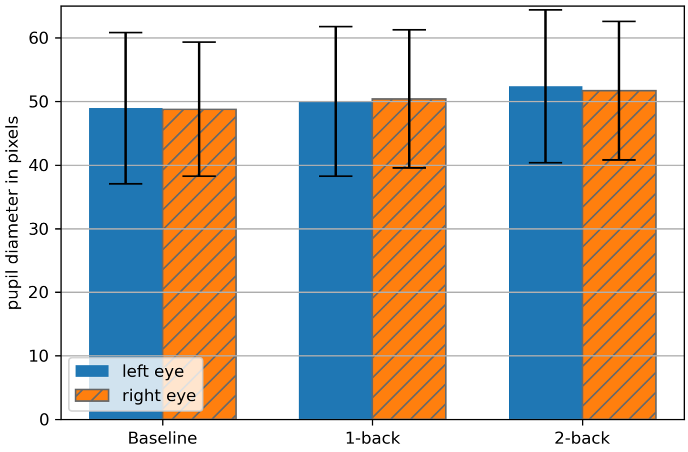


Figure 5.1: Average pupil sizes during cognitive tasks of the combined dataset of Study 1 and Study 2 with standard deviations as error bars

The means and standard deviations per class of all features is given in Table 5.1. The gaze marker X coordinate is horizontal and ranges from 1 (left) to 1920 (right) on screen. The Y coordinate is vertical and ranges from 1 (up) to 1080 (down) on screen. Participants can also look off-screen leading to values out of these ranges.

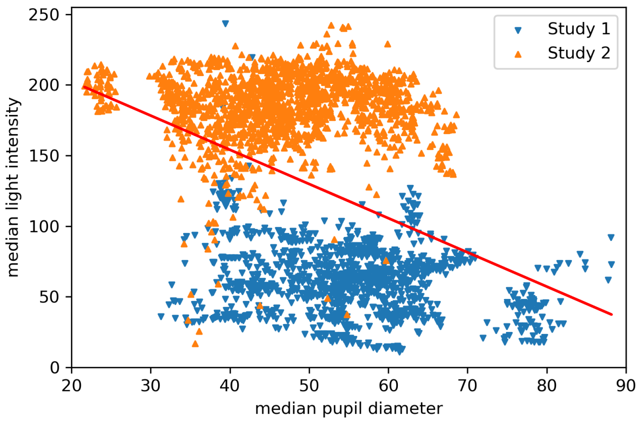


Figure 5.2: Median values per 6 second segment with pupil diameters averaged over both eyes for Study 1 and Study 2 and the median light intensity as described in Section 3 with a least-squares fitted line

Figure 5.2 illustrates that while there is a negative correlation between the median light intensity and the median pupil diameter, there is high variance. This variance is caused by a variety of factors, including natural differences between participants, emotional state and cognitive load. The product-moment correlation coefficient (r) is -0.40 and the rank-order correlation coefficient (ρ) is -0.35. A higher light intensity is correlated with a lower pupil diameter. Figure 5.2 also illustrates the average light intensity differences between Study 1 and Study 2, caused by differences in lighting. The correlations also hold within the studies (Study 1: $\rho = -0.41$ and $r = -0.45$ with a sample size of 1259, Study 2: $\rho = -0.33$ and $r = -0.38$ with a sample size of 1539, p < 0.001 for both individual studies as the combined dataset).

Table 5.1: The means and standard deviations (SD) of all features per class for the combined dataset of Study 1 and Study 2

feature	baseline mean	baseline SD	1-back mean	1-back SD	2-back mean	2-back SD
left MPD (pixel)	48.93	11.87	49.98	11.77	52.36	11.98
right MPD (pixel)	48.75	10.55	50.38	10.83	51.69	10.88
left RCPD (-)	2.678	1.868	2.868	2.428	2.853	2.868
right RCPD (-)	2.492	1.965	2.257	1.451	2.466	1.580
PERCLOS (-)	0.053	0.048	0.059	0.048	0.072	0.058
median gaze X (pixel)	941.9	127.7	965.6	148.7	977.1	142.8
median gaze Y (pixel)	291.7	168.0	286.4	179.2	293.9	155.6
median light intensity (-)	120.2	63.19	130.3	66.61	131.6	66.22
RCLI (-)	13.10	9.846	11.39	6.661	11.82	7.677

5.1. Study 1

The recordings were made with the Dikablis Professional in 2017, in a study to driver distraction when operating IVIS through various interaction methods [29]. After exclusion, data of 24 participants remained. The simulator simulated driving behind a lead-vehicle traveling at 110 km/h on a three-lane highway, no lane changes occurred. 1-back, 2-back and 3-back tasks were recorded whilst driving as a benchmark to compare with several IVIS tasks.

5.2. Study 2

These recordings were made with the Dikablis Glasses 3 in 2019 in a study on driver distraction when operating various in-vehicle interaction systems. After exclusion the data of 30 participants was included in the dataset. The simulated driving task consisted of driving behind a lead-vehicle with a speed of 80 km/h in one of two equiprobable situations presented at random, one with less and one with more traffic. In both cases there would be no lane changes by the participant or other traffic. The simulated road was a two-lane highway. In both cases interaction with the traffic was limited to following the lead-vehicle. Concerning perceived cognitive workload measured by the NASA-TLX mental dimension, the differences between the two different road conditions “less traffic” (N=16, mean=12.9, SD=4.6) and “more traffic” (N=16, mean=13.5, SD=3.9) were not statistically significant (paired t-test, t(30) = -0.71, p = 0.48). Therefore, we did not separate road type conditions within our analysis. 1-back and 2-back tasks were recorded whilst driving as a benchmark to compare with several IVIS tasks.

Table 5.2: numbers of participants and datapoints per cognitive condition and recording circumstances

cognitive condition	nr. of participants	nr. of datapoints
baseline Study 1	29	321
baseline Study 2 less traffic	22	172
baseline Study 2 more traffic	22	157
1-back Study 1	28	469
1-back Study 2 less traffic	17	292
1-back Study 2 more traffic	22	318
2-back Study 1	27	469
2-back Study 2 less traffic	22	315
2-back Study 2 more traffic	17	285

5.3. Pre-processing time series

From the 60 Hz recordings of both eyes time series of the pupil diameter and coordinates (x,y) of a gaze marker were extracted. In the original unprocessed signal blinks and other moments where no pupil was detected were depicted with 0 values, these were used to calculate the PERCLOS. The blinks, including one frame before the blink and 4 frames after the blink, were excluded for the time series of the pupil diameters, gaze marker coordinates and the light intensity. All time series were linearly interpolated to a common time vector before splitting them in fragments of 6 s. Every 60 Hz fragment consisting of 360 samples of different features will be referred to as a single data point, representing a 6 second period. Data points with pupil diameters smaller than 20 pixels or greater than 90 pixels, a PERCLOS of over 0.25, or a light intensity lower than 0 or higher than 255 were assumed to be detection or interpolation errors and excluded from analysis. Combined Study 1 and Study 2 provide 2798 data points after exclusion, corresponding to over 4.5 hours of recordings.

Participants were recorded performing one of three tasks; 1-back, 2-back and baseline, that each are recorded in three circumstances; Study 1, Study 2 "less traffic", Study 2 "more traffic". The classifier separates these 9 groups, but we only distinguish by cognitive task (baseline, 1-back and 2-back) when calculating the results as in Table 6.1. Table 5.2 gives an overview of the sizes of the classes.

6

Results

The 10-fold cross-validation scores of the Random Forest classifiers with and without gaze features and with and without the light intensity features are presented in Table 6.1. Both systems show significant improvement with the use of the light intensity features.

Table 6.1: Classification accuracies of Benchmark 1, System 1, Benchmark 2 and System 2

	without LI (Benchmark)	with LI (System)
without gaze (2)	74.8%	85.9%
with gaze (1)	88.9%	92.5%

Feature importances are calculated with the Gini importance metric. To highlight the importance for classifying cognitive load instead of what study the data belongs to, feature importances have been calculated on a system trained directly on the 3 cognitive conditions (baseline, 1-back and 2-back), not on the 9 classes from Figure 6.2. The feature importances calculated on System 1, with gaze and LI features, shown in Figure 6.1, identify the median X coordinate (horizontal) of the gaze marker to be the most important feature for cognitive load classification, followed by the median light intensity. The median light intensity has a higher importance than well-established measures as PERCLOS and the median pupil diameters. The gaze marker X coordinate has a higher importance than its Y coordinate (vertical). Features calculated on the right pupil diameter have a higher importance than features calculated on the left pupil diameter.

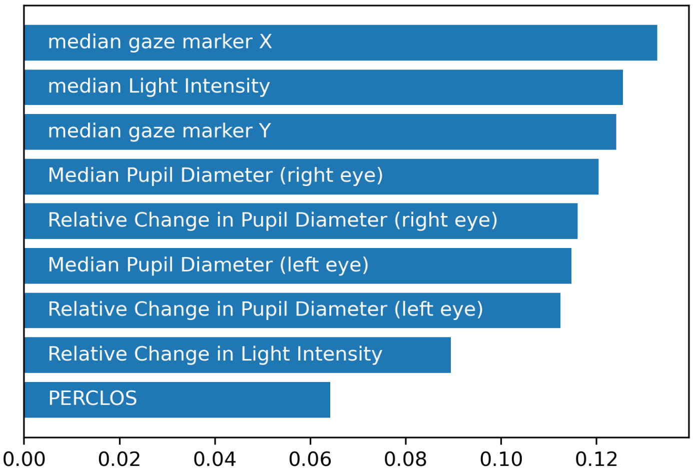
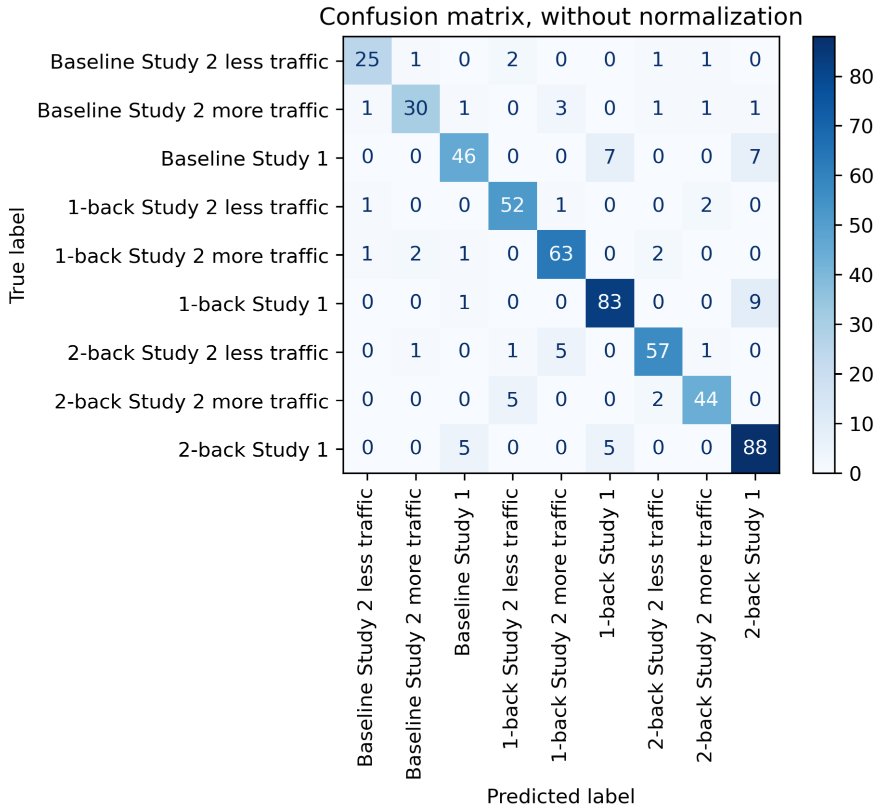


Figure 6.1: Normalized feature importances for System 1 trained on the combined dataset of Study 1 and Study 2

Figure 6.2 was calculated on the system without gaze features, since the gaze features are strongly task dependent. The image was created by training a single classifier on 80% of the data and classifying the remaining 20%. 278 test samples were correctly classified as belonging to the correct cognitive task of Study 2. In 6 out of those 278 samples were the "less traffic" and "more traffic" conditions mixed up. It is important to note that for every cognitive task in Study 2 a participant was given either the "less traffic" or the "more traffic" condition. As a result, the classes contain data from different participants. When only distinguishing by cognitive task, the classes (baseline, 1-back, 2-back) contain data from the same participants.



True label \ Predicted label	Baseline Study 2 less traffic	Baseline Study 2 more traffic	Baseline Study 1	1-back Study 2 less traffic	1-back Study 2 more traffic	1-back Study 1	2-back Study 2 less traffic	2-back Study 2 more traffic	2-back Study 1
Baseline Study 2 less traffic	25	1	0	2	0	0	1	1	0
Baseline Study 2 more traffic	1	30	1	0	3	0	1	1	1
Baseline Study 1	0	0	46	0	0	7	0	0	7
1-back Study 2 less traffic	1	0	0	52	1	0	0	2	0
1-back Study 2 more traffic	1	2	1	0	63	0	2	0	0
1-back Study 1	0	0	1	0	0	83	0	0	9
2-back Study 2 less traffic	0	1	0	1	5	0	57	1	0
2-back Study 2 more traffic	0	0	0	5	0	0	2	44	0
2-back Study 1	0	0	5	0	0	5	0	0	88

Figure 6.2: Confusion matrix of the 9-class System 2 Random Forest classifier

6.1. Testing on unseen participants

When using group K-fold to test on participants unseen in the training phase, the systems' accuracies are much lower. When examining the results of Table 6.2, it should be noted that guessing at random has an expected accuracy of 33.3%. While the introduction of the light intensity features improved the accuracy for both System 1 and System 2, the introduction of gaze features actually decreased accuracy for the Benchmark classifiers. When training on 9-classes as in Figure 6.2, the System 2 classifier with group K-fold yielded an accuracy of 35.5%.

Table 6.2: Classification accuracies of Benchmark 1, System 1, Benchmark 2 and System 2 when using group K-fold

	without LI (Benchmark)	with LI (System)
without gaze (2)	35.2%	35.4%
with gaze (1)	34.7%	36.7%

As shown in Figure 6.3, the inaccuracy of System 2 is not evenly distributed over the classes, the "baseline" class was predicted correctly 58% of the time. System 1 had an accuracy of 55% for the baseline class, Benchmark 1 and Benchmark 2 yielded accuracies of 46% and 38% respectively. For all classifiers the baseline class had the highest accuracy.

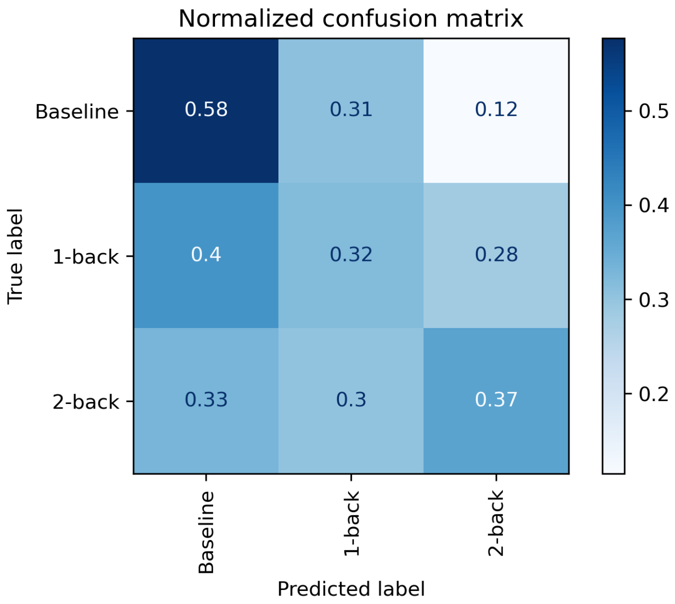


Figure 6.3: Confusion matrix of the 3-class System 2 Random Forest classifier using group K-fold

6.2. Combining the two studies

The classifier with gaze and light intensity features (System 1) has been trained on the individual studies. Where the classifier had an accuracy of 92.5% on the combined set, Study 1 and Study 2 yielded accuracies of 89.9% and 90.5% respectively in k-fold cross-validation (k=10) with the same hyperparameters as used for the combined dataset. Figure 6.4 shows that the classifiers trained on individual studies outperform the classifier trained on the combined until 1100 and 1300 samples. Only after the classifiers trained on individual studies run out of training samples does the classifier trained on the combined studies reach a higher accuracy.

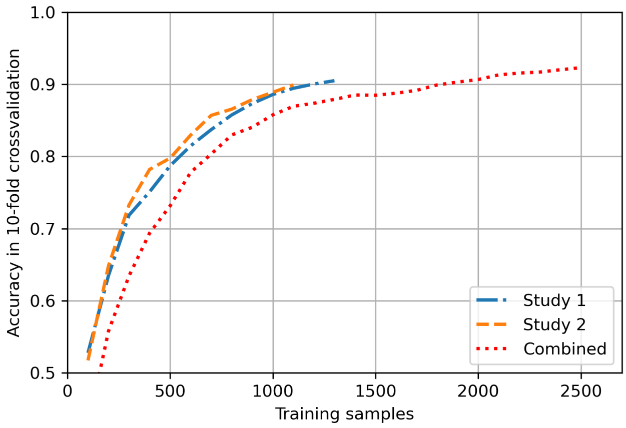


Figure 6.4: Learning curves of a Random Forest classifier trained on Study 1, Study 2 and the combined studies

6.3. Less accurate systems

Variant model architectures, pre-processing methods and feature sets were evaluated but did not yield a better performance than the methods presented above. Systems with a trained XGBoost classifier [30] yielded 1 and 3% lower accuracy scores than the Random Forest classifier, with the feature sets of System 1 and System 2 respectively. Systems where the IPA or the LHIPA was added to the feature set did not have an increased accuracy with respect to the here presented systems [7, 8]. Attempts of smoothing the time series with a median filter followed by a Savitzky-Golay filter reduced the importance of the RCPD features and ultimately classification accuracy. The standard deviations of gaze marker coordinates, pupil diameters and the light intensity over the 6 second segments have been added as features. All reduced the accuracy of the systems, also after re-tuning the hyperparameters. All these less accurate systems benefited from the light intensity features, of which the median consistently ranked as the second most or most important feature.

7 Discussion and Future Work

The results show a significant boost in accuracy when the light intensity features are incorporated, both for systems with and without gaze features. Individually none of the features are a good predictor for cognitive load, but when they are combined a Random Forest classifier can provide accurate estimations. With an accuracy of 92.5% System 1, with gaze features, can compete with state of the art cognitive load estimation techniques [12], but gaze features are very dependent on the task at hand [17]. With an accuracy of 85.9% System 2, without gaze features, is a helpful estimation tool as well. System 2 may generalise better to situations with different tasks that involve different eye and head movements, as it is solely based on the pupil diameter and the light intensity.

However, it is fair to say that the accuracies of the classifiers collapse when testing on unseen participants. The systems are not able to generalise to data of participants who have not been included in the dataset during training and perform marginally better than guessing at random. It appears that the Random Forest classifier strongly relied on individual differences in the chosen features. This is partially because of how the features are designed; e.g. the absolute scale of pupil sizes does not correct for the natural differences in pupil sizes. Only using features that reflect relative changes rather than absolute properties might yield better results. Not being able to generalise to unseen participants is not uncommon for driver assessment applications. Often individual calibration is required to adjust monitoring systems to a driver's driving style and physical properties [31].

The learning curves of Figure 6.4 suggest accuracy could still improve by training the classifier on more studies on n-back tasks [32]. As adding new studies will increase the variance in the dataset it will take more training samples for the classifier to perform well. At the end of the training phase adding one or more studies could result in not only a higher accuracy, but also a more robust classification model.

The light intensity as described in this paper is calculated with the use of the gaze marker coordinates of the participants. As a result the light intensity features also partially capture the gaze behaviour of the participants. This means their contribution to the classification is not just limited to the light intensity's influence on the pupil diameters.

While effective, this light intensity feature is not the perfect representation of the perceived light intensity. Lens distortions, and camera properties and settings are not accounted for [33]. With standardisation it would be easier to translate the light intensity feature between different cameras.

As the datasets were already recorded, we have not tested a controlled range of light intensities. Especially very dark circumstances where the pupils are fully dilated and very light circumstances where the pupils are fully constricted should be explored, as well as scenarios with large fluctuations. In different circumstances it may be beneficial to extract more features from the light intensity than the median and the RCLI to provide a more detailed image.

This research has been conducted with head-mounted eye-trackers. To work with less invasive remote eye-trackers one or more additional cameras are required to capture the field of view of the participant. The computation of the gaze marker and getting the correct line of sight would not be as straightforward as with head-mounted eye-trackers, but this is achievable with extra calculations.

8
Conclusion

Cognitive load estimation based on pupil measurements remains a challenge and many steps still have to be taken before it is a solved problem. Being able to estimate cognitive load accurately in varying lighting conditions is a step in the right direction. The light intensity measure and its derived features as presented in this paper are promising tools to estimate cognitive load more accurately, both when gaze direction can and cannot be used for load estimation. The features are relatively simple to implement and no extra recordings are needed to compensate for the effects of light intensity. We have shown that when implemented in a 3-class classification problem, the features reduced the error 32.4% with and 43.7% without using gaze features. The systems were found to be unfit for classification of participants not included in the training data, with classification accuracies only marginally better than guessing at random. We found that smoothing of the time-series, usage of the IPA and LHIPA features or an XGBoost classifier did not improve accuracy. Both systems presented in this paper only rely on pupil measurements, the available driving simulator data has not been used for classification.

Bibliography

[1] Nathalie Pattyn, Xavier Neyt, David Henderickx, and Eric Soetens. Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? *Physiology & behavior*, 93(1-2):369–378, 2008.

[2] Eckhard H Hess and James M Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

[3] Gerhard Marquart and Joost de Winter. Workload assessment for mental arithmetic tasks using the task-evoked pupillary response. *PeerJ Computer Science*, 1:e16, 2015.

[4] Oskar Palinko and Andrew Kun. Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. 2011.

[5] Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5776–5788, 2016.

[6] Sandra P Marshall. The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*, pages 7–7. IEEE, 2002.

[7] Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[8] Andrew T Duchowski, Krzysztof Krejtz, Nina A Gehrer, Tanya Bafna, and Per Bækgaard. The low/high index of pupillary activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[9] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[10] Jun Chen, Qilin Zhang, Long Cheng, Xudong Gao, and Lin Ding. A cognitive load assessment method considering individual differences in eye movement data. In *2019 IEEE 15th International Conference on Control and Automation (ICCA)*, pages 295–300. IEEE, 2019.

[11] Allan Fong, Ciara Sibley, Anna Cole, Carryl Baldwin, and Joseph Coyne. A comparison of artificial neural networks, logistic regressions, and classification trees for modeling mental workload in real-time. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 1709–1712. SAGE Publications Sage CA: Los Angeles, CA, 2010.

[12] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T Freeman. Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–9, 2018.

[13] Maximilian Schwalm, Andreas Keinath, and Hubert D Zimmer. Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for assistance and automation*, (1986): 1–13, 2008.

[14] Lisa Rerhaye, Talke Blaser, and Thomas Alexander. Evaluation of the index of cognitive activity (ica) as an instrument to measure cognitive workload under differing light conditions. In *Congress of the International Ergonomics Association*, pages 350–359. Springer, 2018.

[15] Jan-Louis Kruger, Esté Hefer, and Gordon Matthew. Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, pages 62–66, 2013.

[16] Rahul Gavas, Debatri Chatterjee, and Aniruddha Sinha. Estimation of cognitive load based on the pupil size dilation. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1499–1504. IEEE, 2017.

[17] Johan Engström, Emma Johansson, and Joakim Östlund. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation research part F: traffic psychology and behaviour*, 8(2):97–120, 2005.

[18] Xiaofei Hu, Rumi Hisakata, and Hirohiko Kaneko. Effects of stimulus size, eccentricity, luminance, and attention on pupillary light response examined by concentric stimulus. *Vision Research*, 170:35–45, 2020.

[19] Philip A Stanley and A Kelvin Davies. The effect of field of view size on steady-state pupil diameter. *Ophthalmic and Physiological Optics*, 15(6):601–603, 1995.

[20] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.

[21] William David Wright and JH Nelson. The relation between the apparent intensity of a beam of light and the angle at which the beam strikes the retina. *Proceedings of the Physical Society*, 48(3):401, 1936.

[22] Johannes J Vos. Depth in colour, a history of a chapter in physiologie optique amusante. *Clinical and Experimental Optometry*, 91(2):139–147, 2008.

[23] Andrew T Duchowski. Eye tracking methodology. *Theory and practice*, 328(614):2–3, 2007.

[24] Roy H Steinberg, Miriam Reid, and Paula L Lacy. The distribution of rods and cones in the retina of the cat (felis domesticus). *Journal of Comparative Neurology*, 148(2):229–248, 1973.

[25] Joseph B Keller. Geometrical theory of diffraction. *Josa*, 52(2):116–130, 1962.

[26] Haruhisa Sekiya, Sachiko Hasegawa, Kazuo Mukuno, and Satoshi Ishikawa. Sensitivity of nasal and temporal hemiretinas in latent nystagmus and strabismus evaluated using the light reflex. *British journal of ophthalmology*, 78(5):327–331, 1994.

[27] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.

[28] Bruce Mehler, Bryan Reimer, and Jeffery A Dusek. Mit agelab delayed digit recall task (n-back). *Cambridge, MA: Massachusetts Institute of Technology*, page 17, 2011.

[29] Nikolas Pärsch, Martin Baumann, Arnd Engeln, and Lutz W.H. Krauß. Change the way to manage an in-vehicle menu selection and thereby lower cognitive workload? In *Proceedings of the 6th Driver Distraction and Inattention conference, Gothenburg, Sweden, October 15-17*, DDI '18, 2018. URL `http://ddi2018.org/wp-content/uploads/2018/10/S2.4-P%C3%A4rsch.pdf`.

[30] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[31] Gulbadan Sikander and Shahzad Anwar. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2339–2352, 2018.

[32] Claudia Perlich. Learning curves in machine learning., 2010.

[33] Junhee Park, Seong-Chan Byun, and Byung-Uk Lee. Lens distortion correction using ideal image coordinates. *IEEE Transactions on Consumer Electronics*, 55(3):987–991, 2009.

[34] Fabian Doubek, Erik Loosveld, Riender Happee, Joost de Winter, Fabian Doubek, and Erik Loosveld. Take-over quality: Assessing the effects of time budget and traffic density with the help of a trajectory-planning method.

Appendix A: The feature PERCLOS

In Figure 8.1 the means and standard deviations of the feature with the lowest importances that still improved classification are depicted: PERCLOS. The means would suggest that PERCLOS could be an important variable for classification, but the high stand deviation shows that PERCLOS had high variations between the individual measurements, this was even the case with measurements from the same recording. By itself PERCLOS is not a reliable indicator for cognitive load, but a combination of many non-perfect indicators can still result in accurate classification.

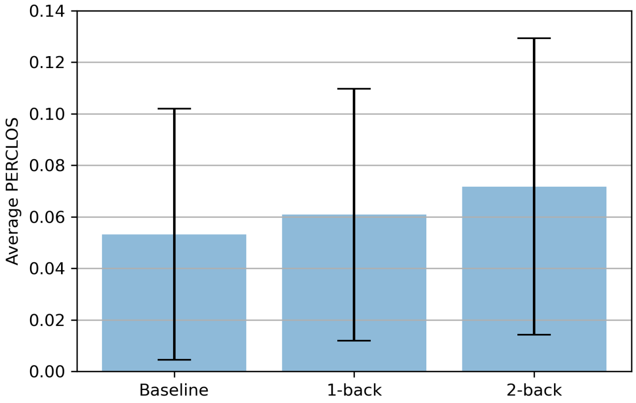


Figure 8.1: Mean PERCLOS per cognitive task with the standard deviation as error bars

Appendix B: Tables with means and standard deviations

The tables below depict the means and standard deviations of Study 1 and Study 2 that have been merged into one dataset. Study 2 has also been split into the conditions "less traffic" and "more traffic". Please note that Table 8.3 and Table 8.4 do not have the same participants in every class (cognitive task), as the data is separated per traffic condition. This makes them less suitable for comparison between cognitive tasks.

Table 8.1: The means and standard deviations of all features per class for Study 1

feature	baseline mean	baseline SD	1-back mean	1-back SD	2-back mean	2-back SD
left MPD (pixel)	52.23	13.47	54.49	12.50	58.42	11.70
right MPD (pixel)	51.82	10.98	55.53	9.782	56.81	9.577
left RCPD (-)	2.099	1.410	2.612	2.670	2.854	3.668
right RCPD (-)	1.500	1.355	1.274	0.622	1.639	1.419
PERCLOS (-)	0.061	0.050	0.049	0.043	0.066	0.059
median gaze marker X (pixel)	897.4	137.0	921.8	178.2	936.2	167.3
median gaze marker Y (pixel)	186.2	160.2	166.9	178.5	192.6	162.3
median light intensity (-)	60.77	22.50	62.97	25.01	61.26	25.05
RCLI (-)	14.38	9.172	10.87	8.348	13.53	10.27

Table 8.2: The means and standard deviations of all features per class for the combined dataset of both traffic conditions of Study 2

feature	baseline mean	baseline SD	1-back mean	1-back SD	2-back mean	2-back SD
left MPD (pixel)	45.74	8.998	46.17	9.893	47.55	9.989
right MPD (pixel)	45.76	9.201	46.15	9.838	47.64	10.20
left RCPD (-)	3.244	2.075	3.081	2.184	2.853	2.035
right RCPD (-)	3.460	1.984	3.075	1.434	3.112	1.386
PERCLOS (-)	0.045	0.045	0.068	0.050	0.077	0.057
median gaze marker X (pixel)	985.3	100.5	1002	105.7	1009	110.2
median gaze marker Y (pixel)	394.7	96.19	385.8	102.8	373.1	90.86
median light intensity (-)	178.1	24.30	186.4	26.24	186.6	20.56
RCLI (-)	11.85	5.032	11.82	4.786	10.48	4.289

Table 8.3: The means and standard deviations of all features per class for the less traffic condition of Study 2

feature	baseline mean	baseline SD	1-back mean	1-back SD	2-back mean	2-back SD
left MPD (pixel)	46.70	9.679	49.45	7.807	44.45	10.31
right MPD (pixel)	45.99	9.317	49.45	6.764	44.59	11.12
left RCPD (-)	3.012	1.684	3.746	2.679	2.283	1.183
right RCPD (-)	3.654	1.715	3.210	1.573	3.174	1.407
PERCLOS (-)	0.054	0.049	0.071	0.052	0.069	0.047
median gaze marker X (pixel)	924.2	81.44	979.2	93.82	1032	105.7
median gaze marker Y (pixel)	414.2	100.9	386.4	110.4	374.3	79.72
median light intensity (-)	179.2	23.93	189.3	21.55	187.2	22.89
RCLI (-)	11.07	3.771	12.20	4.206	10.05	4.014

Table 8.4: The means and standard deviations of all features per class for the more traffic condition of Study 2

feature	baseline mean	baseline SD	1-back mean	1-back SD	2-back mean	2-back SD
left MPD (pixel)	44.68	8.056	43.38	10.60	50.98	8.371
right MPD (pixel)	45.51	9.067	43.34	11.09	51.01	7.789
left RCPD (-)	3.499	2.407	2.515	1.423	3.483	2.533
right RCPD (-)	3.247	2.223	2.960	1.293	3.043	1.360
PERCLOS (-)	0.035	0.038	0.065	0.048	0.085	0.065
median gaze marker X (pixel)	1052	73.35	1021	111.2	983.8	109.4
median gaze marker Y (pixel)	373.2	85.77	385.3	95.80	371.6	101.7
median light intensity (-)	176.9	24.65	183.9	29.43	185.9	17.60
RCLI (-)	12.71	6.007	11.49	5.207	10.96	4.527

Appendix C: Signal smoothing

Figure 8.2 shows the pupil diameter signal as I initially wanted to process it. In the original signal blinks were depicted with 0 values. These values were removed. Because the pupil detection was inaccurate 1 frame before the zeroes and up to 4 frames after, these samples were also removed. The resulting gaps were interpolated linearly. A median filter with a width of 3 samples was applied to remove outliers, followed by a second order Savitzky-Golay filter with a width of 31 samples. The systems yielded accuracies that were 6% lower than when not applying the filters. A system was trained with just the median filter applied, its accuracy was 4% lower than the systems presented in this paper. I did continue applying linear interpolation to gaps created by excluding the blinks and surrounding samples.

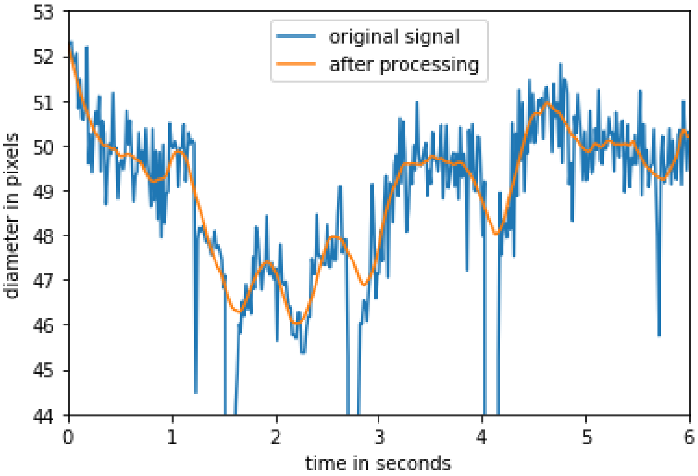


Figure 8.2: Signal smoothing not applied in final systems

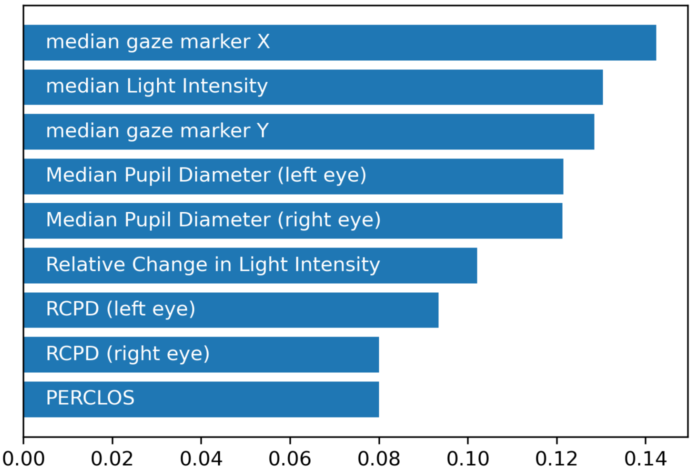


Figure 8.3: Normalized feature importances for System 1 trained on the combined dataset of Study 1 and Study 2 with smoothing filters applied

The drop in performance when applying smoothing suggests that a lot of the predictive information of the time series for pupil diameters is found in the fast dilations and contractions, at least when examined with the feature sets of System 1 and System 2. Even just removing single outlier samples with the median filter decreased accuracy with 4%. The smoothing removes some detail, especially affecting the Relative Change in Pupil Diameter. When comparing the feature importances of System 1 trained on the smoothed time series in Figure 8.3 with the feature importances of Figure 6.1, it becomes apparent that it is mainly the importance of RCPD for both eyes that has changed. It could be that the RCPD captures the same rapid dilations and constrictions that the ICA, IPA and LHIPA aim to find [6–8]. However, smoothing filters or not, the features IPA and the LHIPA did not improve classification accuracy for the systems of this paper.

Appendix D: The correlation between light intensity and pupil diameter

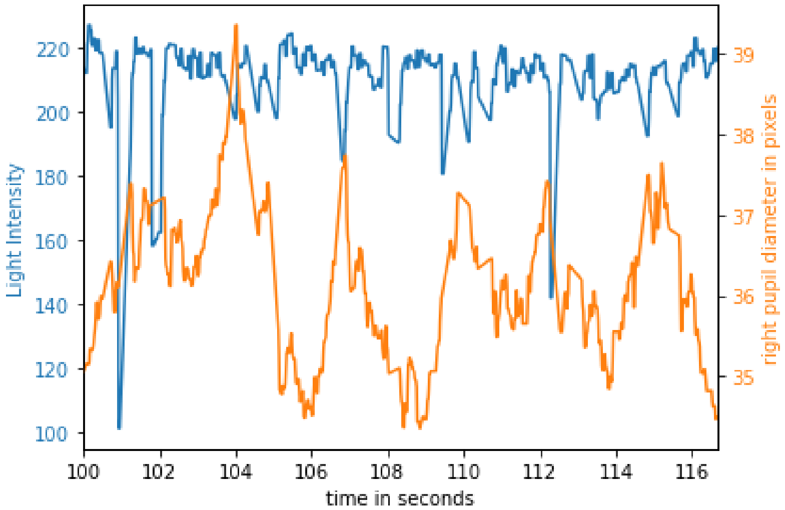


Figure 8.4: Light Intensity vs pupil diameter

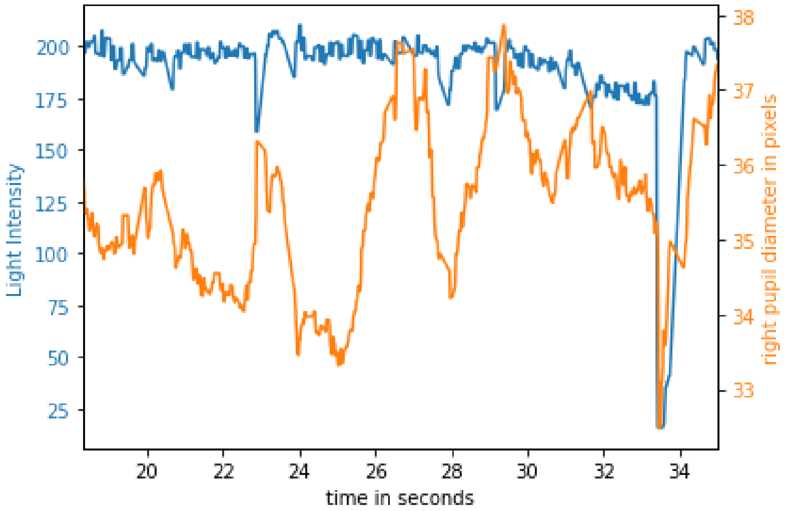


Figure 8.5: Light Intensity vs pupil diameter

All correlation coefficients mentioned in this appendix are the Pearson's product-moment correlation coefficient (r). When the light intensity increases, pupils generally constrict. In Figure 8.4 we see peaks in the light intensity coinciding with drops in pupil diameter. The correlation coefficient for this segment is -0.23. While for a large part we see the same behaviour in Figure 8.5, we also see a sharp drop can be seen in both the Light Intensity and the pupil diameter at t = 33.5 seconds. This drop cannot be explained by the pupillary light response. The correlation coefficient for this segment is positive, 0.28. In the screenshot of the eye-camera of

Figure 8.6, it shows that these coinciding drops were an erroneously processed half-blink. The eye is closed partially for a short time, causing the system to detect a smaller pupil. The eye never fully closes, the original signal shows no zeroes that signal no pupil is detected and as a result the half-blink and surrounding frames are not dealt with in processing. The gaze coordinates are calculated from the pupil position, which is also shifted by the partial detection of the pupil. The gaze direction ends up on the black dashboard, the result is a simultaneous drop in light intensity and pupil diameter.

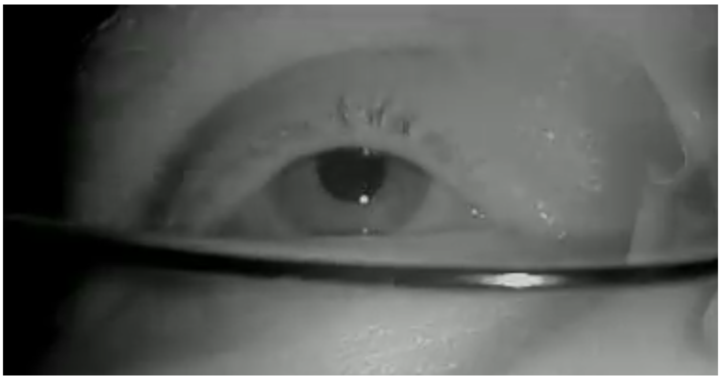


Figure 8.6: Partially closed eye at t = 33.5

For the complete two minute recording the correlation coefficient is 0.03, suggestion there is little to no correlation between the two variables. For the data of Study 1 and Study 2 combined it is -0.38, but there is a lot of variation between participants, recordings and even segments of the same recording. Note, r = -0.38 for the raw data, for the processed medians r = -0.40. Table 8.5 lists the means and standard deviations of the correlation coefficients of all individual segments, grouped per cognitive task and recording condition. The average correlations range from -0.086 to 0.044, with standard deviations ranging from 0.270 to 0.358. This indicates very low correlations, with large differences between the correlations. Adding this correlation as a feature in classification decreased accuracy with a half percent.

Table 8.5: means and standard deviations per cognitive tasks of correlation coefficients of individual segments between the light intensity measure and the average pupil diameter

cognitive condition	mean	standard deviation	sample size
baseline Study 1	0.044	0.332	321
baseline Study 2 less traffic	-0.051	0.318	172
baseline Study 2 more traffic	0.001	0.279	157
1-back Study 1	-0.086	0.358	469
1-back Study 2 less traffic	-0.014	0.299	292
1-back Study 2 more traffic	-0.002	0.358	318
2-back Study 1	-0.061	0.354	469
2-back Study 2 less traffic	0.020	0.270	315
2-back Study 2 more traffic	-0.026	0.317	285

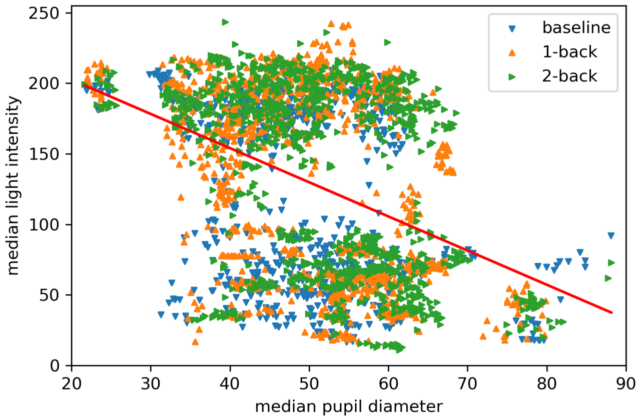


Figure 8.7: Median values per 6 second segment with pupil diameters averaged over both eyes for Study 1 and Study 2 per cognitive task

Figure 8.7 contains the same data points as Figure 5.2, but separated by cognitive task. The distinction per cognitive class in this graph seems to be less clear than the distinction per study.

Appendix E: Hyperparameter tuning

The following hyperparameters have been tuned; the number of trees, maximum depth of the tree (max depth), the minimum number of samples required to split an internal node (min samples split), the minimum number of samples required to be at a leaf node (min samples leaf) and whether bootstrapping would be used in the trees or not. The number of trees means the number of individual decision trees that vote in the random forest algorithm. Generally, the more trees, the higher the performance, but the more resources required for training the algorithm. At some point the performance shows asymptotic behaviour, performance no longer increases when the number of trees increases, but the amount of required resources will still rise. For these reasons the optimal value for the number of trees is the lowest number that still results in maximum performance. The maximum depth of the tree is the number of nodes a tree is allowed to have. As the standard setting in the scikit-learn library no maximum depth is imposed. The "min samples leaf" is set to 1 if not specified. It can work as a smoothing factor, especially when using regression instead of classification. The "min samples split" is set to 2 if not specified, meaning a node can split as soon as there is more than one sample. If bootstrapping is applied, samples are drawn with replacement. As a result not every tree is trained on the entire dataset. This increases the variance in individual trees, but can reduce the variance of the overall forest without increasing its bias. Since this does not always work out, both settings have been tested. By default bootstrapping is used by the scikit-learn library.

Since none of these variables are continuous, a grid was used instead of a continuous range. Together the parameter options of Table 8.6 form 2160 possible combinations. Trying them all would be inefficient. To find the best hyperparameters for the Random Forest classifiers in an efficient way, first a randomised grid search was applied. Out of the 2160 combinations, 100 are picked at random. The best tried combination had 800 trees, no maximum depth, a "min samples split" of 2, a "min samples leaf" of 1 and no bootstrapping.

Table 8.6: The grid of hyperparameters investigated with a randomised grid search

number of trees	max depth	min samples split	min samples leaf	bootstrapping
200	10	2	1	yes
400	20	5	2	no
600	30	10	4	
800	40			
1000	50			
1200	60			
1400	70			
1600	80			
1800	90			
2000	100			
	110			
	none			

Hyperparameters that the random grid search determined to be optimal at their standard setting have been set to their standard setting. The two hyperparameters that were not standard, number of trees and bootstrapping, were varied. The outcome of the random grid search was used as input for an exhaustive grid search, all eight combinations of Table 8.7 have been investigated. Two systems returned the maximum performance of 92.5% accuracy, 1000 trees without bootstrapping and 1200 trees without bootstrapping. The final hyperparameters were 1000 trees, no maximum depth, a "min samples split" of 2, a "min samples leaf" of 1 and no bootstrapping. Only the hyperparameters that are not standard are mentioned in the paper.

Table 8.7: The grid of hyperparameters investigated with an exhaustive grid search

number of trees	max depth	min samples split	min samples leaf	bootstrapping
600	none	2	1	yes
800				no
1000				
1200				

The parameters were tuned on all four systems, the two systems with light intensity features and their two baselines. The same parameters were found for all but Baseline 2, which reached its maximum accuracy with 600 trees. To make the comparison between systems as fair as possible the same parameters were used for all systems, so Baseline 2 was trained with a forest of 1000 trees.

Appendix F: Motion simulator specifications

Table 8.8: Specifications of the eMove eM6-640-1800 motion simulator [34]

	Excursions		Velocity	Acceleration
	Single DOF	Non-single DOF		
Surge	-0.48 to 0.60 [m]	-0.64 to 0.63 [m]	0.8 [m/s]	7 [m/s^2]
Sway	-0.50 to 0.50 [m]	-0.66 to 0.66 [m]	0.8 [m/s]	7 [m/s^2]
Heave	-0.41 to 0.41 [m]	-0.41 to 0.41 [m]	0.6 [m/s]	10 [m/s^2]
Roll	-23.8 to 23.8 [deg]	-29.2 to 29.2 [deg]	35 [deg/s]	250 [deg/s^2]
Pitch	-23.7 to 26.0 [deg]	-28.2 to 32.9 [deg]	35 [deg/s]	250 [deg/s^2]
Yaw	-25.4 to 25.4 [deg]	-28.7 to 28.7 [deg]	40 [deg/s]	500 [deg/s^2]

Appendix G: A study on take-over maneuvers

To test generalisability, a study on take-over maneuvers was analysed with the Random Forest classifier trained on n-back data. This separate study was not used in the training phase. Six different take-over maneuvers were recorded with the Dikablis Glasses 3 in the same driving simulator set-up as Study 1 and 2. In all scenarios the vehicle would be driving autonomously on the right lane of a two-lane highway. As the vehicle would approach roadworks in the right lane, the participant would get a take-over warning. The objective was to take over control of the vehicle and safely merge into the left lane to avoid collision with the obstacle. The difference between the scenarios was the time between the obstacle and the warning, and the traffic density in the left lane. Figure 8.8 shows the objective ratings for these six scenarios, where time budget (TB) is the time between warning and obstacle and traffic density (TD) is the traffic density on the left lane. Error bars indicate the standard error. Scenario 6 had a 20 second TB and the lowest TD, and was found the least critical and complex. Scenario 1 had the lowest TB, and the highest TD and was found to be the most critical and complex. The footage was cut to start from the obstacle warning and end 5 seconds after passing the obstacle. The remaining footage was cut into 6 second segments which were processed and classified by the Random Forest classifiers.

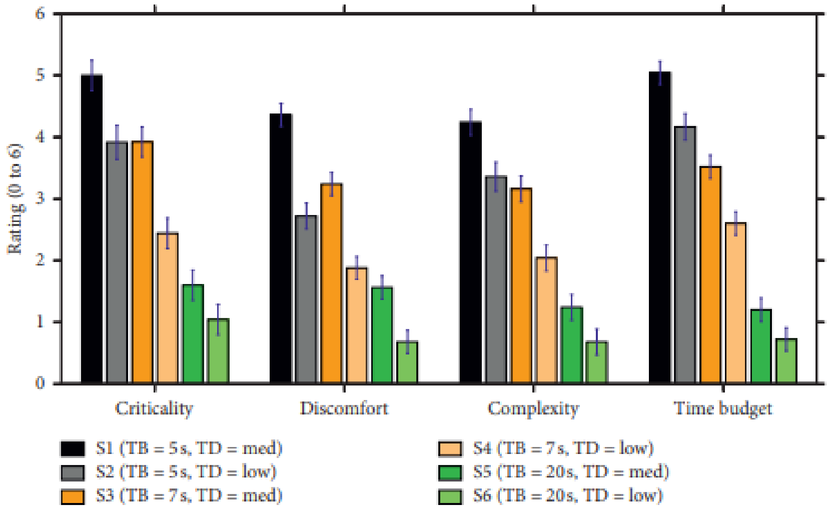


Figure 8.8: Subjective ratings of 6 take-over maneuvers [34]

Table 8.9: Segments of take-over maneuvers analyzed with n-back classifier System 1

Maneuver	Baseline	1-back	2-back
S1	83.5%	1.9%	14.5%
S2	82.0%	1.5%	16.5%
S3	87.5%	3.0%	9.5%
S4	76.4%	1.1%	22.4%
S5	75.6%	3.7%	20.7%
S6	74.1%	1.5%	24.4%

Table 8.10: Segments of take-over maneuvers analyzed with n-back classifier System 2

Maneuver	Baseline	1-back	2-back
S1	51.7%	32.4%	15.9%
S2	48.9%	28.9%	22.2%
S3	50.5%	32.5%	17.0%
S4	43.3%	33.1%	23.6%
S5	34.1%	36.9%	29.0%
S6	37.3%	32.8%	29.9%

Table 8.9 shows the results of classifying segments of the study with 6 different take-over maneuvers with the System 1 classifier. For all maneuvers the category "Baseline" is awarded the majority of segments. In all

cases "1-back" is awarded the least segments by the classifier. Table 8.9 was created with the classifier System 1, with gaze and LI features. The classifier System 2 in Table 8.10, without gaze features, shows more balance between the classes. This still defies the expectations.

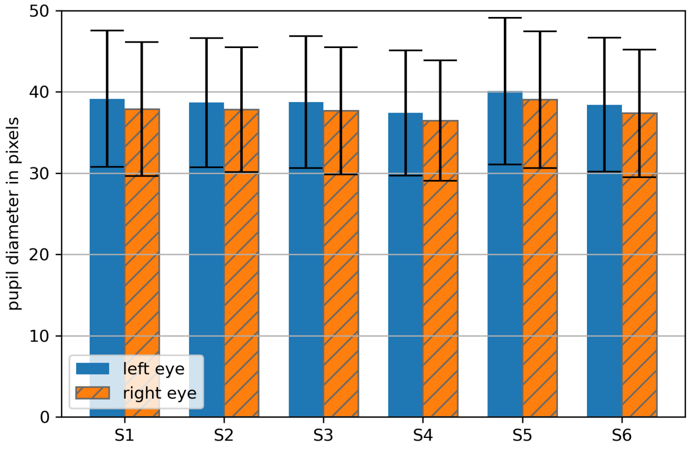


Figure 8.9: Average pupil sizes during take-over maneuvers

We found that the n-back classifier was not suitable for analysing the study of take-over maneuvers. The expectation was that S1 would show the highest cognitive load, as it was rated the most critical and least comfortable. This could have shown by e.g. a higher share of segments classified as 2-back and a lower share of segments classified as Baseline than for a maneuver associated with a lower criticality, like S6.

The take-over study was done with different participants than the n-back studies. As we have found, the classifiers are not able to generalise to unseen participants. With this in mind it is not surprising that the classifiers were not able to classify this results as the subject ratings would suggest. Another plausible explanation for the results is that the cognitive load of an overtake maneuver is not constant. To be expected is a spike in load when the warning is given, possibly followed by a 5-second roller-coaster of emotions with fear, panic, and in case of success, possibly relief or euphoria. System 1 and System 2 require the load to be constant during a 6-second segment. The n-back tasks provided the participants with a constant level of cognitive load, the take-over tasks did not. A third plausible explanation is that the load experienced during a take-over maneuver is not in the range the n-back classifiers were trained on.

An interesting note is that when gaze features are included, almost all segments are classified as baseline, while 1-back is the smallest class. When the gaze features are not taken into account, the classes are more balanced. The average pupil sizes, shown in Figure 8.9, did not increase in situations where we expected a higher load, in contrast to what we have seen with the n-back studies. As the average pupil diameters of the over-take study were lower than those of the n-back studies, pupil sizes were normalised for all studies when analysing the take-over scenarios. This had little to no effect on classification results.

Appendix H: Merging 9 recording circumstances to 3 classes

The systems that do not use group K-fold are trained on 9 classes of Figure 6.2, instead of directly on the 3 classes that we want to distinguish by: "baseline", "1-back" and "2-back". We do not get into this in the paper because the implications are relatively small and it distracts from the main point of the paper, but in this appendix we aim to clarify why we do this and why it improves the accuracy.

When guessing at random the expected value of the accuracy of a 3-class problem is 1/3. For a 9-class problem this is 1/9, which could suggest that a 9-class classification problem is inherently more difficult. Of course a classifier is not guessing at random, machine learning algorithms can benefit from more clearly defined classes. When grouping the data of Study 1 and Study 2, we are grouping data with very different properties. Not only is the median light intensity different between the studies, we found that there are large differences between subjects. We suspect that our classifiers benefited from having 9 more distinct classes, yielding higher classification accuracies than when training them on the 3 classes.

After the 9-class classification, the classes were merged to the 3 cognitive conditions, 'baseline', '1-back' and '2-back'. This means that if a sample was erroneously classified as '1-back Study 1', while it belonged to '1-back Study 2 Less Traffic', it would now be counted as being classified correctly. This extra step improved classification, but not by as much as choosing a 9-class system over a 3-class system did. For System 1, with gaze features, the indirect way of creating the classes described in Section 5.3 improves accuracy with 1.87%. In the testing of the 10 classifiers that make up the total system accuracy, 9 samples were corrected by the class merging, accounting for 0.32% of the 1.87% accuracy improvement. For System 2, without gaze features, this improvement was 0.58%. For System 2, 8 of the 2798 samples were initially classified incorrectly but corrected by the merging, accounting for just under half of the total accuracy improvement. Note that this is a slightly different result from Figure 6.2. This is because that image was created out of a single train and test split of 80% and 20% respectively of the data instead of a 10-fold cross-validation.

For the systems trained on the group K-fold splits the advantage of this indirect way of creating classes was not present. As the accuracy of the systems was low overall, the benefit of more distinct classes was overshadowed by the increased difficulty of a 9-class problem with respect to the difficulty of a 3-class problem. For this reason the systems were trained directly on the three cognitive tasks.

Appendix I: T-testing on participants averages

The following tables have been calculated on the combined dataset of Study 1 and Study 2. The averaged values per participant per cognitive condition were compared between cognitive conditions with paired and unpaired t-tests. It shows that when pairing the t-tests, the p-values are lower than when performing unpaired t-tests, indicating that a more significant difference can be found between the cognitive classes.

Table 8.11: Paired t-testing of the median light intensity averaged per participant per cognitive condition

	baseline	1-back	2-back
baseline	-	t(49) = -1.48 p = 0.144	t(49) = -1.32 p = 0.191
1-back	t(49) = -1.48 p = 0.144	-	t(46) = 0.410 p = 0.684
2-back	t(49) = -1.32 p = 0.191	t(46) = 0.410 p = 0.684	-

Table 8.12: Unpaired t-testing of the median light intensity averaged per participant per cognitive condition

	baseline	1-back	2-back
baseline	-	t(49) = -0.347 p = 0.729	t(49) = -0.399 p = 0.691
1-back	t(49) = -0.347 p = 0.729	-	t(46) = 0.0466 p = 0.963
2-back	t(49) = -0.399 p = 0.691	t(46) = 0.0466 p = 0.963	-

Table 8.13: Paired t-testing of the median right pupil diameter averaged per participant per cognitive condition

	baseline	1-back	2-back
baseline	-	t(49) = 3.97 p <0.001	t(49) = -5.80 p <0.001
1-back	t(49) = 3.97 p <0.001	-	t(46) = -5.37 p <0.001
2-back	t(49) = -5.80 p <0.001	t(46) = -5.37 p <0.001	-

Table 8.14: Unpaired t-testing of the median right pupil diameter averaged per participant per cognitive condition

	baseline	1-back	2-back
baseline	-	t(49) = 0.676 p = 0.500	t(49) = -1.70 p = 0.0922
1-back	t(49) = 0.676 p = 0.500	-	t(46) = -0.999 p = 0.320
2-back	t(49) = -1.70 p = 0.0922	t(46) = -0.999 p = 0.320	-

Table 8.15: Paired t-testing of the median left pupil diameter averaged per participant per cognitive condition

	baseline	1-back	2-back
baseline	-	t(49) = -1.23 p = 0.223	t(49) = -5.58 p <0.001
1-back	t(49) = -1.23 p = 0.223	-	t(46) = -5.32 p <0.001
2-back	t(49) = -5.58 p <0.001	t(46) = -5.32 p <0.001	-

Table 8.16: Unpaired t-testing of the median left pupil diameter averaged per participant per cognitive condition

	baseline	1-back	2-back
baseline	-	t(49) = -0.377 p = 0.707	t(49) = -1.99 p = 0.0492
1-back	t(49) = -0.377 p = 0.707	-	t(46) = -1.57 p = 0.119
2-back	t(49) = -1.99 p = 0.0492	t(46) = -1.57 p = 0.119	-

TU Delft

PORSCHE