Towards Real-Time, Nonintrusive Estimation of Driver Workload in a Simulated Environment

van Gent, Paul; Farah, Haneen; van Nes, Nicole; van Arem, Bart

Publication date
2017

Document Version
Accepted author manuscript

Published in
Proceedings of the 6th International Conference on Road Safety & Simulation (RSS)

Citation (APA)
van Gent, P., Farah, H., van Nes, N., & van Arem, B. (2017). Towards Real-Time, Nonintrusive Estimation of Driver Workload in a Simulated Environment. In *Proceedings of the 6th International Conference on Road Safety & Simulation (RSS): 17-19 October 2017, The Hague, Netherlands* Article 283

This paper will be published in: Proceedings of the Road Safety and Simulation Conference 2017
17-19 October 2017, The Hague, Netherlands

Towards Real-Time, Nonintrusive Estimation of Driver Workload: A Simulator Study

Paul van Gent[a], Haneen Farah[a], Nicole van Nes[b], Bart van Arem[a]

a. Department of Transport&Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology. Stevinweg 1, 2628CN Delft.
p.vangent@tudelft.nl; h.farah@tudelft.nl; B.vanArem@tudelft.nl
b. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV), Bezuidenhoutseweg 62, 2594AW Den Haag.
Nicole.van.nes@swov.nl

Abstract

The aim of this research is to work towards building an open-source, platform-independent algorithm capable of predicting driver workload in real-time and in a non-intrusive way. To work towards a system that can also be implemented in on-road settings, we aimed at using off-the-shelf, non-intrusive sensors that could be implemented into the steering wheel and dashboard of current and future generations of cars, making them non-intrusive.

In order to build the initial predictive model, a driving simulator experiment was performed. Nineteen participants were required to drive a virtual replication of the Dutch A67 C-ITS corridor between Eindhoven and Venlo. We attempted to induce driver workload by varying weather, traffic composition, traffic density and by asking participants to perform various manoeuvres such as lane changing, merging and exiting. We measured heart rate, skin response, blink and performance measures.

Results show that within individuals and within the experimental group, workload was predictable with a high correct rate in both individual models as well as group models. We also evaluated how well the models would generalise when used outside of the experimental setting. Preliminary results for this generalisation are poor. We discuss possible reasons for this and next steps we are planning to take to increase this performance.

Keywords

Driver workload; machine learning; workload prediction; driving simulator

1. Introduction

1.1 Background

Research into the relation between workload and physiological measures is not new [1], [2]. Yet, no consensus has been reached on which measures work consistently well when predicting driver workload. The literature shows that different measures might show different effects in different settings ([3], [4]).

Recently, some focus in traffic research has shifted more toward data-driven methods in an attempt to explain the phenomenon, with some studies achieving interesting results [5], [6]. However, these methods rely on sensing systems, measures or approaches that may not be practical in settings outside the lab, and the resulting models are not usually available for use in other studies. The aim of this research is to lay the groundwork for an open-source, platform-independent algorithm capable of predicting driver workload in real-time, using nonintrusive, off-the-shelf sensing methods. To work towards a system that can also be implemented in on-road settings, we aimed at using sensors that could also be implemented in the steering wheel and dashboard of current and future generations of cars, making them non-intrusive.

1.2 Objectives

The objectives of this study are to create a classification model capable of predicting driver workload in driving simulator settings. We will use only low-cost, off-the-shelf sensing equipment that can easily be integrated into the driving environment. We use a data-driven machine learning approach. This means we include a large selection of variables and let the algorithms perform variable selection based on how well they contribute to the to-be-predicted classes. The datasets, algorithms and models following from the research will be published open source on the project website once the research is complete and results published.

2. Defining Practical Measures for Workload Prediction

We surveyed the driver workload literature and identified measures linked to driver workload. These are heart rate, galvanic skin response, eye blinks, breathing rate, electroencephalogram (EEG) and performance measures. Following this, we evaluated the measures based on the ease of measurement, their capacity to interfere with the driving task, whether they are easy to measure with off-the-shelf sensors and whether these sensors can be integrated into the driving environment. We describe the physiological and performance measures below. How the criteria apply to the measures is displayed in table 1.

2.1. Physiological Measures

In order to define physiological measures for use in workload classification we surveyed the literature and found heart rate, galvanic skin response, blink rate, breathing rate and EEG as often occurring measures.

Heart rate (HR) has been used in most experiments related to driver workload, and is linked to driver workload consistently [7]–[9]. The heart signal is often split into heart rate and heart rate variability measures. The heart rate is a simple measure of the heart period, expressed in the beats per minute and the inter-beat interval. Heart rate variability measures describe how the heart rate signal varies over time, and can be divided into time-domain measures and frequency-domain measures [8], [10]. In the time-domain we include the beats per minute (BPM), the inter-beat interval (IBI), the median absolute deviation of intervals between heart beats (MAD), the standard deviation of intervals between heart beats (SDNN), the root mean square of successive differences between neighbouring heart beat intervals (RMSSD), the standard deviation of successive differences between neighbouring heart beat intervals (SDSD), the proportion of differences between successive heart beats greater than 50ms and 20ms (pNN50, pNN20, resp.). In the frequency domain we include two frequency bands: low frequency (LF, 0,04-0,15Hz), which is related to short-term blood pressure variation, and high frequency (HF, 0,16-0,5Hz), which is a reflection of breathing rate [10].

Galvanic Skin Response (GSR) is a measure of the electrical conductivity of the skin. It reflects the activation of the sympathetic branch of the autonomous nervous system, which functions as a system to prepare the body for action when situational demands increase, such as in the fight-or-flight-response [1]. GSR has been linked to (driver) workload [11] and stress [12]. Like HR, GSR is often expressed in time-domain measures and frequency-domain measures. In the time-domain it is often split into tonic and phasic components [13]. The tonic component represents the low-frequency variation in the GSR signal and is thought to represent overall psycho-physiological activation [14]. The phasic component consists of more short-term variation in the signal. It is thought to represent shorter responses to external stimuli, typically occurring 1-3 seconds after an external stimulus [14]. In the time-domain we include the mean value of the GSR signal, the difference between the largest and smallest value in the signal, the standard deviation of the signal and the median-absolute-deviance (MAD) of the signal. In the frequency domain, the 0.03 – 0.5Hz frequency band is of interest [15].

Blink data has been linked to heightened driver workload as well as drowsiness (underload) [11], [16]. In one experiment, Benedetto et al. [16] had participants perform a lane change task in a simulator in both single and dual-task settings. They report that short blinks become more frequent as visual load increases, and the occurrence of long blinks increases as ‘time on task’ increases. There is also evidence for a different effect for different types of load: high perceptual load led to lower blink rates, whereas high cognitive load led to higher blink rates [17]. We include the blink rate (number of blinks given a fixed time interval), blink duration and blink interval.

Breathing rate has been used in predicting driver workload [9], although the results are somewhat mixed [18]. Measures found in the literature are breathing rate [19] and the frequency spectra in 0-0.1Hz, 0.1-0.2Hz, 0.2-0.3Hz and 0.3-0.4Hz [20]. We had trouble locating a non-intrusive breathing sensor, and thus did not include this measure in the experiment, although nonintrusive sensing methods have been proposed [21].

The EEG (Electroencephalogram) measures localised brain activity with sensors attached to the scalp using induction. It has been used successfully in experiments predicting workload [9]. We did not include EEG measurements, as this measure does not satisfy the criteria. It requires complex set-ups that are not easy to integrate into the driving environment, takes considerable time to set up each time it is used and has the potential to interfere with the driving task since a large apparatus is attached to the head of the driver.

2.2 Performance Measures

Performance measures reflect how the control the driver exerts over the vehicle varies across conditions. If a driver becomes distracted or the driving task imposes a high load, control over the vehicle may suffer as a result. Performance measures often found are related to steering wheel movements (steering wheel angle, steering wheel reversals, steering wheel entropy, all related to the small variation in steering wheel position indicative of lane-keeping) [6], [22], to variation in the lateral and longitudinal axis [23], headway [24] or time-to-collision [25]. We include these measures in the data collection.

Table 1: Measures evaluated by criteria as set forth in the text
(++ : very good, + : good, +/- : fair, - : inadequate, -- :very inadequate)

Measure	Ease of measurement	Intereference with Driving Task	Off-the-shelf sensors	Integration into Driving environment
HR	+	++	++	+
GSR	++	++	++	+
Blink Rate	+	++	+/-	+/-
Breathing	+/-	+	+/-	+/-
EEG	--	--	-	--
Performance	++	++	+	++

3. Method

In this section briefly discuss the equipment used in the experiment, the scenarios we designed and the procedure that was followed when collecting data.

3.1 Equipment

The driving simulator experiment was performed in a fixed-base, medium-fidelity driving simulator manufactured by GreenDino. It consisted of a dashboard mock-up with three 4K displays (resolution of each display: 4096*2160 pixels) providing roughly 180-degree vision, Fanatec steering wheel and pedals, custom blinker control and a desktop computer running Windows 10. The simulation itself ran in Unity3D. Simulator data was logged at 50Hz.

The participants' faces were filmed with a GoPro HERO+ camera situated on the dashboard, running at 1080p/30fps with the purpose of detecting blinks. A second GoPro camera filmed from behind the participant and was used to visually inspect and link any artefacts in the data to what was happening either on the simulated road or in the simulator.

Physiological measures were recorded using off-the-shelf sensors, powered by an Atmel ATMega328p embedded processor board. Heart rate was recorded at the left index finger using a photoplethysmographic (PPG) method [21]. Skin response was measured at the ring and pinkie finger of the left hand using a set of standard cuffs with metal contacts on the inside, coupled with a signal amplifying and noise reduction circuit. See Figure 1 for an illustration of the sensors on the left hand. Data was logged at 100Hz to a laptop running Windows 7.

Figure 1: (a) The GreenDino fixed-base, medium-fidelity driving simulator; (b) Demonstration of physiological sensors attached to the left hand. The PPG sensor is placed at the index finger, the GSR sensor cuffs at the middle and ring finger.

3.2 Scenarios

Two scenarios were created in Unity3D: a high workload scenario and a low workload scenario. The road was based on a section of the A67 C-ITS corridor in the Netherlands, between Eindhoven and Venlo. The scenarios were driven in three weather conditions: clear weather, light fog (visibility approx. 150 meters) and heavy fog (visibility below 25 meters). This gave a total of six scenarios for each participant.

CAD drawings for the geometry of the road and surrounding terrain were secured from the open data program of the Dutch government. Road geometry was extracted from these datafiles to generate an accurate road model, and textured in Autodesk 3DS Max. Surrounding terrain was generated using data from the CAD drawings combined with heightmap data available from the Microsoft Bing Maps API. Patches of forest and canals were generated automatically from satellite imagery, and where necessary adjusted by hand. The location and content of traffic signs was inferred from Google Streetview. Signs were designed in 3DS Max, textured with Photoshop and placed in the scenario at the corresponding location.

Figure 2: (a) Screenshot of merging into dense truck column in the heavy fog condition; (b) screenshot of traffic jam alongside closed left lane with accident in the good weather condition.

The high workload scenario ran from Eindhoven to Someren (15.9 km). Participants started on an on-ramp and subsequently had to merge into a dense platoon of trucks traveling with a time headway of approx. 0.2 seconds, a manoeuvre that has been shown to increase mental effort required from the driver [26]. The trucks were programmed to only give way after a driver attempted to merge into the very narrow gap, this was designed as such to be stressful. After approximately 4 kilometres, participants encountered slow moving traffic on the right lane and an empty left lane, designed to prompt them to start driving left. While passing the slow-moving traffic, an ambulance came up behind with 160 km/h, putting the participant in the stressful position of finding and merging into a tight gap in the slower moving right lane. The ambulance travelled 40 km/h above the speed limit of 120 km/h, which is in accordance with Dutch regulations regarding maximum ambulance speed. The ambulance exhibited standard auditory and visual signals. Three kilometres after encountering the ambulance, a game of ’20 questions’ [27] was played to simulate an engaging (phone) conversation. Participants had to guess what person, animal or object the experimenter had in mind by asking at maximum 20 questions. The experimenter would only respond with either yes or no. Near the end of the simulation the right lane was closed due to an accident, with slow moving (<15km/h) traffic on the left lane. The 20 questions game was played until the traffic jam was encountered, if the participant finished before reaching the traffic jam, the game was restarted with a different subject. Two kilometres downstream from the traffic jam, participants were instructed to take the exit and stop the car. The data between each ‘event’ was not taken into account for the analysis.

The low workload scenario ran from Someren to Venlo (20.5 km). It served as the control drive, there were no events and only light traffic on the road. After merging onto the highway, participants drove until they reached the designated exit, where they were instructed to exit and stop the car.

3.3 Procedure

A pilot study was performed with three participants driving a single scenario, to test the recording equipment and the experimental procedure. The equipment functioned properly, however two participants expressed confusion when asked to rate their mental workload on a 7-point scale. Subsequently, we decided to split the question in two questions regarding mental effort and difficulty, rather than a single question based on workload.

Approval for the experiment was obtained from the TU Delft ethics committee. Upon registration, participants received a copy of the informed consent, instructed to sign it and bring it to the first session. Participants were seated in the driving simulator and were allowed to relax for a few minutes so that their physiological signals had time to return to their baseline after walking to the experiment room. Physiological sensors were attached to the index finger, middle finger and ring finger of the left hand, and the signal quality was verified. One minute of physiological baseline data was recorded after participants had been seated for approximately three minutes. The procedure of the drive was explained to the participants. If requested, the game of 20 questions was explained and, if still unsure, a test round was played prior to the drive to familiarise the participant with the procedure. Participants were instructed to drive at their own pace, but not exceed the speed limit visible on road-side signs.

After approximately one minute of driving in the driving simulator, participants were queried for the occurrence of simulation sickness, and instructed to inform the experimenter should it arise later. At fixed locations in the scenario, participants were asked to rate both how difficult the driving task was, as well as how much mental effort they felt it took on a 7-point scale, with 0 being easy/low, and 7 being difficult/high.

Participants drove two scenarios on three separate days. This was done because physiological measurements can vary from day to day, and to avoid the occurrence of a fatigue effect from having participants drive six 10-15 minute scenarios consecutively on one day.

4. Analysis

This section describes how the physiological data was pre-processed and how the machine learning datasets were constructed.

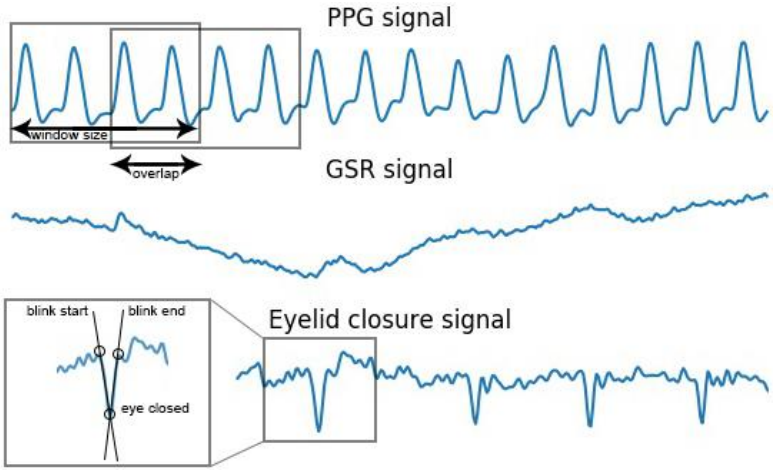

Figure 3: Examples of the PPG, GSR and Blink signals from a participant. An example of how the blink detection algorithm functions is included, as well as an illustration of the window size and window overlap factors.

4.1 PPG data

The raw heart-rate signal was analysed using an open-source Python algorithm [28], which identified the position of all R-peaks in the heart rate signal and calculated the required measures in the temporal and frequency domains. An example of the signal is shown in figure 3.

For the time-series data we calculated the beats per minute (BPM), the mean absolute deviation (MAD) from the BPM, the mean inter-beat interval (IBI), the standard deviation of intervals between heart beats (SDNN), the standard deviation of successive differences between heart beat intervals (SDSD), the root mean square of successive differences between heart beats (RMSSD) and the proportion of differences between successive heartbeat intervals greater than 20ms and 50ms (pNN20, pNN50).

In the frequency domain, we utilised the HF and LF frequency bands. These were calculated using a Fast Fourier Transform (FFT). The heart beat RR intervals are an inherently unevenly distributed time-series, since their position in time depends on the interval's magnitude. Because the FFT requires evenly spaced data, data was resampled using a cubic spline interpolation to create an evenly distributed time-series. After performing the

FFT, the LF and HF bands were extracted using a trapezoidal integration of the area under the 0.04Hz - 0.15Hz and 0.16Hz – 0.5Hz frequency bands in the squared FFT output.

4.2 GSR data

Skin response consists of a tonic and a phasic component [13]. The tonic component is the slow, long-term variation in the skin response signal, thought to reflect overall psycho-physiological activation [14]. The phasic component reflects short-term responses to external stimuli, and can generally be seen as short increases of skin conductance occurring with a latency of 1-3 seconds after stimulus onset [14]. We extracted the mean, the max-min difference, the mean absolute deviance (MAD) in the time-domain.

The frequency domain is also of interest in the analysis of skin response data, as reported by Shimomura et al. [15]. The skin response signal was transformed to the frequency domain using FFT with a Hamming Filter to reduce spectral leakage. Following this, we extracted the < 0.5Hz frequency component rather than the 0.03Hz-0.5Hz component, since the lower portion of the frequency spectrum would not be present in the short window sizes we used. An example of the signal is shown in figure 3.

4.3 Blink data

An algorithm was developed to detect blink data in the video. It extracted blink number, blink duration (mean and standard deviation) and interblink-interval (mean and standard deviation). The algorithm functioned by detecting 68 'facial landmarks' [29], displayed in figure 4. From these facial landmarks, the distances between top eyelid and bottom eyelid for both eyes were calculated for each video frame. Blinks were detected by detecting large slopes in the signal, indicating closing or opening motions of the eye. The start of the closing motion and the end of the opening motion were logged. From these, the described measures were calculated. The process along with an example of the signal is shown in figure 3.

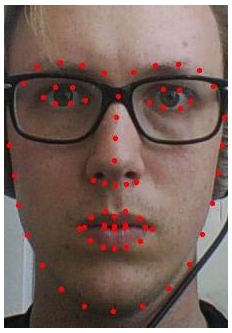

Figure 4: Facial Landmarks Detected on Face

4.4 Pre-processing

During all scenarios, participants were asked to rate the difficulty of the driving task and mental effort required on a 7-point scale. For the high workload scenario, participants rated both variables directly after one of the events (merging, ambulance, 20 questions and traffic jam). For the low workload scenario, participants were queried when they passed fixed locations on the road. Since interacting with the driver might influence workload, the scenarios were constructed in such a way that least one minute of data following a query would not be used in the analysis (e.g. would not contain an 'event').

An algorithm was developed to select corresponding data segments for each time the participants were queried from the physiological data, simulator data and facial landmarks datasets. Since the sampling frequency was different for the physiological data (100Hz), the blink data (30Hz) and the simulator data (50Hz), the resulting separate matrices were time-synchronised and stored in a hierarchical data structure.

4.5 Machine Learning Sets

The machine learning datasets were generated using the raw data for each event that was created in the pre-processing step. Sets were generated for all combinations of window sizes and overlap factors. The concepts of window size and overlap factors are illustrated in figure 2 on a portion of the PPG signal. We used window sizes of 5, 10, 20, 30, 40, and 60 seconds, overlap factors of 0, 25, 50, 75% and 95%, leading to a total of 30 sets. For the PPG and GSR measures, the difference between the window and baseline value was used. The data was standardised using the StandardScaler implementation in SciKit-Learn, as this is a requirement for Support Vector Machines (SVM) classifiers.

4.6 Classifier Development

We used two algorithms for the classification task: a Support Vector Machines (SVM) classifier and a Random Forest (RF) classifier. The RF classifier creates an ensemble (forest) of decision trees in which each tree is trained on a random subset of the features, effectively minimising the correlation between the individual trees to ensure each tree accounts for a unique portion of variance in the training data. When classifying, the data is

run through each tree, and the class predicted by the majority of the decision trees is output by the algorithm as the classification outcome. They have been used in for example [30].

Support Vector Machine classifiers function by mapping the data to a higher dimensional space, and solving an optimization problem to identify a set of hyperplanes that separate the training data into different classes. They have been used in for example [6], [25], [31]. SVM classifiers are generally insensitive to the curse of dimensionality and can handle large datasets efficiently. With the SVM classifier, we evaluated the Polynomial kernel (SVM(poly)), which uses a polynomial function to generate hyperplanes that separate the data into classes, and the Radial Basis Function kernel (SVM(rbf)), uses a radial basis function to generate hyperplanes separating the data into classes.

Using the generated datasets for each combination of window size and overlap factor, both individual and group models were generated. Models were evaluated using K-fold cross validation with K=5. For all models we considered RF, SVM(poly) and SVM(rbf) classifiers. Driver workload was predicted on a 7-point scale, with 0 being very low and 7 being very high. We used the Python SciKit-Learn package for the classification algorithms [32].

5. Results

5.1 Participant demographics

19 participants took part in the experiment. Data from one participant was excluded because not all tasks were understood due to a language barrier. This left 18 participants, 12 males and 6 females, with an average age of 34.56 years (SD: 10.09). 12 participants owned a car, while 6 did not. On average, those owning a car used it three to four times a week and estimated to travel between 2500 and 15000km annually. During the experiment, one participant mentioned slight simulator sickness while navigating sharp curves. Since most of the drive took place on a relatively straight highway, this participant expressed the wish to continue the experiment but chose not to navigate the exit of the highway at the end of the scenarios.

5.2 Self-reported workload

Participants reported their perceived mental effort and the perceived difficulty of the driving task on a 7-point Likert scale, with 0 being ‘low’/’easy’ and 7 being ‘high’/’difficult’. Reported mental effort and perceived difficulty correlated moderately with weather conditions ($r = .458$, $p < 0.001$, $r = .465$, $p < 0.001$), and with scenario type ($r = .328$, $p = 0.001$, $r = .404$, $p < 0.001$), respectively.

A repeated-measures ANOVA was performed with weather condition and scenario type as factors. A Greenhouse-Geisser correction was applied since the assumption of sphericity was violated. Results showed a significant effect for weather condition ($F(1.880, 22.564) = 9.127$, $p = 0.001$) and for scenario type ($F(12.863, 10.981) = 14.057$, $p = 0.003$). No significant interaction effect between weather condition and scenario type was present ($F(1.746, 20.948) = 0.109$, $p = 0.873$). This seems to indicate that our secondary tasks and weather conditions were successful in raising self-reported workload and task difficulty independently of each other, but that combinations of weather and workload conditions did not raise difficulty further.

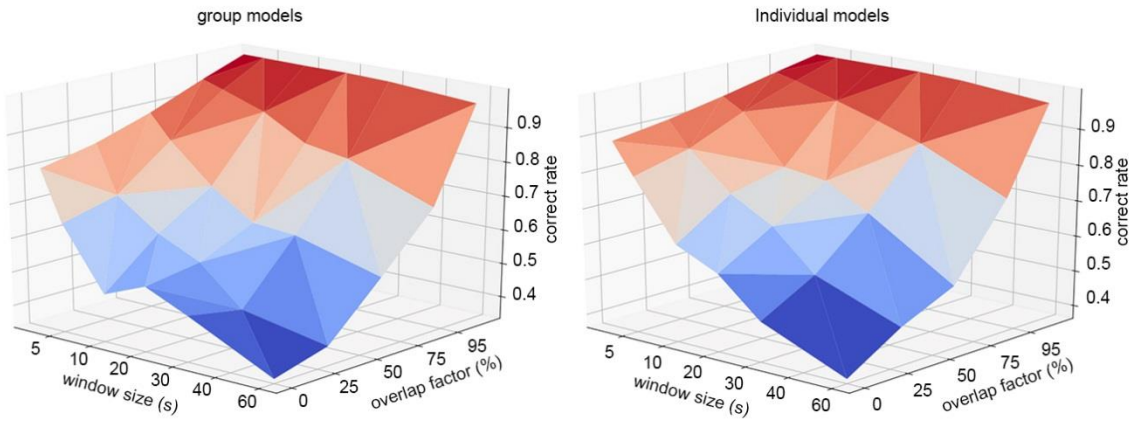

Figure 5: Plots displaying the average model correct rate for all combinations of window sizes and overlap factors for the RF classifier.

We were interested in performance of the classification models at the individual level as well as the group level. Using the machine learning datasets described earlier, we generated individual models and group models for all combinations of window sizes (5, 10, 20, 30, 40 and 60 seconds) and overlap factors (0, 25, 50, 75% and 95%). For each combination of overlap factor and window size, the classifier attempted to predict the self-reported workload.

5.3 Group models

This section briefly describes the classification procedure for the group models. The RF classifier outperformed the SVM(rbf) and SVM(poly) classifiers, especially at lower overlap factors. The left side of Figure 2 displays the average correct rate of the group models plotted against the used window size and overlap factors for the RF classifier. It shows an interaction effect between window size and overlap factor, generally with longer window sizes leading to a decrease in model performance, and increased window overlap leading to an increase in performance. The effect of decreasing performance with increasing window size is contrary to what has been reported by others [6], [33]. We suspect two main reasons for this. The first is that with 18 participants, a window size of 60 seconds and low overlap, the number of training instances available to the classifier becomes low, which can reduce performance. The second reason is that, when driving in naturalistic environments as in this experiment, driver workload may change on short timescales in response to short events (i.e. a lane change, a braking action, merging), and these short variations in measures are averaged out when the window size becomes larger.

Since we were also interested in what measures contributed most to model performance, feature importance was evaluated for the group models. The best performing features were the mean of the GSR signal, the frequency component of the GSR signal, the mean driving speed, the pNN20 heart rate variability measure and the IBI heart rate measure.

5.4 Individual models

This section briefly describes the classification procedure for the individual models. Similar to the group models, the RF classifier outperformed the SVM(rbf) and SVM(poly) classifiers, again especially at lower overlap factors. Especially at larger window sizes, there was considerable individual variation in model performance. This seems to indicate that for some, this type of workload prediction may work better than for others. Individual differences in physiological or behavioural responses to workload might be a cause of this.

The right side of figure 2 displays the average accuracy of the individual models plotted against the used window size and overlap factor values. Additionally, table 2 displays the accuracy of individual models with the RF classifier. We only included window sizes 5, 10 and 30 seconds for the first five participants (participant 2 was excluded) to keep the table reasonably sized. An average for all individual models is also included to give an overview of the result of the performance of all individual models.

Table 2: Correct rates for the RF Classifier, for window sizes of 5, 10 and 30 second and overlap factors 0, 25, 50, 75 and 95%

	Correct Rate Random Forest Classifier														
	5 second window					10 second window					30 second window				
pp.	0%	25%	50%	75%	95%	0%	25%	50%	75%	95%	0%	25%	50%	75%	95%
1	.93	.95	.98	.99	.99	.86	.91	.97	.99	1	.53	.66	.85	.98	1
3	.94	.94	.95	.99	1	.92	.93	.96	.98	1	.84	.90	.95	.97	1
4	.97	.98	.99	1	1	.93	.89	.98	.99	1	.71	.91	.94	1	1
5	.94	.94	.96	.99	.99	.90	.90	.95	.99	.99	.76	.88	.87	.99	1
6	.90	.94	.97	.99	1	.83	.87	.97	.99	1	.75	.76	.86	.96	1
all	.94	.95	.97	.99	.99	.89	.92	.96	.99	.99	.73	.80	.91	.98	.99

On average, both classifiers perform better at higher overlap factors and shorter window sizes. This result is similar to the result discussed for the group models, likely due to similar reasons of the number of training instances in the dataset, and the short timescale of workload variation.

Since we were also interested in what measures contributed most to model performance, feature importance was evaluated for the models. On average, the best performing features were the mean of the GSR signal, the frequency component of the GSR signal, the mean driving speed, the LF heart rate variability measure and the standard deviation of the driving speed.

5.5 Generalising to other drivers

We were also interested in how well the models would generalise to individuals outside of the experiment. To estimate real-world behaviour of the models, we evaluated how they would behave when generalising to unknown participants by using a k-fold method with individual participants as folds. For each model, one participant was held out, the model trained, and the workload for the unknown participant predicted. This was repeated until all participants had been held out once.

Results showed that, at least given our dataset, the models did not generalise well to unknown drivers. Results varied depending on the individual, ranging from chance level (approx. 0.14 with seven categories) to about 0.4. Although significantly above chance level, this is not a useful correct rate. This lack of generalising capacity of the models could have several causes. We suspect that, since we're working with self-reported workload, not all participants rate workload evenly. For example, assuming an objective load score of '5', one participant might score it as 5 on a 7-point scale, whereas another might rate the same perceived load as 7. Secondly, there may exist different patterns in physiological responses to workload that may differ across individuals. The simulator we used in the experiment may also lack the realism required to elicit a high enough physiological response from the drivers. This indicates that more research is required to reach the goals set in the paper, which we will continue working towards.

6. Conclusion and Discussion

In this paper, we have described our first steps towards the construction of an open-source driver workload prediction algorithm. Participants drove on three separate days, each day driving a high workload and a low workload scenario in randomly assigned weather conditions. We have recorded physiological measures and performance measures, and attempted to predict self-report workload measures from this data. We have shown that initial models function well on individual and group levels, indicating there is a common variance underlying the self-report workload measures and the physiological responses.

The main difficulty at present remains generalising the model to unknown drivers, as the models performed well at the individual and group level, but not when predicting data from a never-before-seen participant withheld from the training set. Our use of a self-report workload measure can be a contributing factor to this due to interpersonal differences in response tendencies. It may be that the physiological response to workload in the driving environment differs from person to person, in which case our sample size of 18 may be inadequate to account for all occurring categories. It may also be that the driving environment puts diverse types of demands on the driver (visual, cognitive, auditory) and that these types of driver demand have different physiological response patterns, or that the use of the artificial driving simulator environment did not induce large enough physiological responses in the participants. Further research using more naturalistic, on-road settings with a larger set of participants is required.

Our next steps are to explore workload measures other than self-report, gather on-road data to expand the current dataset and explore ways to increase the generalising performance of the models. We will explore alternative methodological approaches and additional feature generation techniques. The end-goal is an open-source workload model.

Acknowledgements

This research was performed in the Taking the Fast Lane project, which was funded by "Applied and Technical Sciences" (TTW), which is a subdomain of the Dutch Institute for Scientific Research ('NWO').

References

[1] D. de Waard, *The Measurement of Drivers ' Mental Workload*. Alphen aan den Rijn: Drukkerij Haasbeek, 1996.

[2] J. Aasman, G. Mulder, and L. J. M. Mulder, "Operator effort and the measurement of heart-rate variability," *Hum. Factors*, vol. 29, no. 2, pp. 161–170, 1987.

[3] G. Matthews, L. E. Reinerman-Jones, D. J. Barber, and J. Abich, "The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 57, no. 1, pp. 125–143, 2015.

[4] B. Mehler, B. Reimer, and J. F. Coughlin, "Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task: An On-Road Study Across Three Age Groups," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 54, no. 3, pp. 396–412, Apr. 2012.

[5] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-Physiological Measures for Assessing Cognitive Load," *Proc. 12th ACM Int. Conf. Ubiquitous Comput.*, pp. 301–310, 2010.

[6] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 340–350, 2007.

[7] B. Reimer, B. Donmez, M. Lavallière, B. Mehler, J. F. Coughlin, and N. Teasdale, "Impact of age and cognitive demand on lane choice and changing under actual highway conditions," *Accid. Anal. Prev.*, vol. 52, pp. 125–132, 2013.

[8] B. Mehler, B. Reimer, and Y. Wang, "Comparison of heart rate and heart rate variability indices in distinguishing single task driving and driving under secondary cognitive workload," *Proc. Sixth Int. Driv. Symp. Hum. Factors Driv. Assessment, Training, Veh. Des.*, pp. 590–597, 2011.

[9] E. Ferreira *et al.*, "Assessing real-time cognitive load based on psycho- physiological measures for younger and older adults," in *Computational Intelligence, Cognitive Algorithms, mind, and Brain (CCMB)*, 2014.

[10] N. Montano *et al.*, "Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior," *Neurosci. Biobehav. Rev.*, vol. 33, no. 2, pp. 71–80, 2009.

[11] N. Nourbakhsh, Y. Wang, and F. Chen, "GSR and blink features for cognitive load classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8117 LNCS, no. PART 1, pp. 159–166, 2013.

[12] J. A. Healey and R. W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, 2005.

[13] C. L. Lim *et al.*, "Decomposing skin conductance into tonic and phasic components," *Int. J. Psychophysiol.*, vol. 25, no. 2, pp. 97–109, 1997.

[14] M. Seitz, T. J. Daun, A. Zimmermann, and M. Lienkamp, "Measurement of Electrodermal Activity to Evaluate the Impact of Environmental Complexity on Driver Workload," *Proc. FISITA 2012 World Automot. Congr.*, pp. 245–256, 2012.

[15] Y. Shimomura, T. Yoda, K. Sugiura, A. Horiguchi, K. Iwanaga, and T. Katsuura, "Use of frequency domain analysis of skin conductance for evaluation of mental workload.," *J. Physiol. Anthropol.*, vol. 27, no. 4, pp. 173–177, 2008.

[16] S. Benedetto, M. Pedrotti, L. Minin, T. Baccino, A. Re, and R. Montanari, "Driver workload and eye blink duration," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 14, no. 3, pp. 199–208, 2011.

[17] S. W. Savage, D. D. Potter, and B. W. Tatler, "Does preoccupation impair hazard perception? A simultaneous EEG and Eye Tracking study," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 17, pp. 52–62, 2013.

[18] S. Kim, J. Chun, and A. K. Dey, "Sensors Know When to Interrupt You in the Car," *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. - CHI '15*, pp. 487–496, 2015.

[19] J. Mueller, L. Stanley, T. Martin, and D. Beach, "Driving Simulator and Scenario Effects on Driver Response," *Ind. Syst. Eng. Res. Conf.*, pp. 507–514, 2014.

[20] J. Jarvis, F. Putze, D. Heger, and T. Schultz, "Multimodal person independent recognition of workload related biosignal patterns," *Proc. 13th Int. Conf. multimodal interfaces - ICMI '11*, p. 205, 2011.

[21] H. Jae Baek, H. Bit Lee, J. Soo Kim, J. Min Choi, K. Keun Kim, and K. Suk Park, "Nonintrusive Biological Signal Monitoring in a Car to Evaluate a Driver's Stress and Health State," *Telemed. e-HEALTH*, vol. 15, no. 2, pp. 182–189, 2009.

[22] O. Nakayama, T. Futami, T. Nakamura, and E. R. Boer, "Development of a steering entropy method for evaluating driver workload," *Proc. JSAE Annu. Congr.*, no. 724, pp. 5–8, 1999.

[23] R. Hoogendoorn, S. Hoogendoorn, K. Brookhuis, and W. Daamen, "Mental Workload , Longitudinal Driving Behavior and Adequacy Of Car Following Models In Case Of Incidents In The Other Driving Lane," *Transp. Res. Board*, no. Idm, pp. 1–21, 2010.

[24] E. Teh, S. Jamson, O. Carsten, and H. Jamson, "Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 22, pp. 207–217, 2014.

[25] D. Pinotti, G. Francesco, B. Piccinini, and F. Tango, "Adaptive human machine interface based on the detection of driver ' s cognitive state using machine learning approach," vol. 8, pp. 163–179, 2014.

[26] D. de Waard, A. Kruizinga, and K. A. Brookhuis, "The consequences of an increase in heavy goods vehicles for passenger car drivers' mental workload and behaviour: A simulator study," *Accid. Anal. Prev.*, vol. 40, no. 2, pp. 818–828, 2008.

[27] A. L. Kun, A. Shyrokov, and P. a. Heeman, "Interactions between human-human multi-threaded dialogues and driving," *Pers. Ubiquitous Comput.*, vol. 17, no. 5, pp. 825–834, 2013.

[28] P. van Gent, "Analyzing a Discrete Heart Rate Signal Using Python," 2016. [Online]. Available: http://www.paulvangent.com/2016/03/15/analyzing-a-discrete-heart-rate-signal-using-python-part-1/.

[29] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2144–2151, 2011.

[30] M. Miyaji, M. Danno, H. Kawanaka, and K. Oguri, "Driver's cognitive distraction detection using adaboost on pattern recognition basis," *Proc. 2008 IEEE Int. Conf. Veh. Electron. Safety, ICVES 2008*, pp. 51–56, 2008.

[31] L. Moreira-matias and H. Farah, "On Developing a Driver Identification Methodology Using In-Vehicle Data Recorders," *IEEE Trans. Intell. Transp. Syst.*, vol. submitted, 2017.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[33] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler, "Classifying driver workload using physiological and driving performance data," *Proc. 32nd Annu. ACM Conf. Hum. factors Comput. Syst. - CHI '14*, pp. 4057–4066, 2014.