PHYSIOPRINT: A Workload Assessment Tool Based on Physiological Signals

Djordje Popovic, Maja Stikic, Chris Berka
Advanced Brain Monitoring Inc.
2237 Faraday Ave, Suite 100
Carlsbad, CA, 92008
{dpopovic,maja,chris}@b-alert.com

David Klyde, Theodore Rosenthal
Systems Technologies Inc.
13766 Hawtorne Blvd.
Hawtorne, CA
{dklyde, trosenthal}@systemstech.com

ABSTRACT
In this paper we introduce a novel workload assessment tool (called PHYSIOPRINT) that is based on the combination of two types of physiological signals: electroencephalography (EEG) and electrocardiography (ECG). The tool is inspired by a theoretical workload model developed by the US Army that covers a large number of different workload types relevant for driving scenarios, including auditory, visual, cognitive, and motor workload. The PHYSIOPRINT classifier was trained on the EEG and ECG data acquired during well-defined atomic tasks chosen to represent the corresponding types of workload. The trained model was validated on realistic driving simulator data from an independent study. The highest performance on the atomic tasks was achieved for visual workload, with precision of 91.8% and recall of 94.1%. The corresponding classification results in the validation study were: precision 78.3% and recall 80.6%. The utilized classification approach is not computationally expensive, so it can be easily integrated into automotive applications.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/Machine Systems – *human factors, human information processing, software psychology.*

General Terms
Algorithms, Measurement, Performance, Experimentation.

Keywords
Workload, electroencephalography, electrocardiography, driving simulator, physiology.

1. INTRODUCTION
Due to the rapid advances in technology and related changes in consumers' lifestyles and expectations, the motor vehicle industry will likely continue to integrate more sophisticated entertainment and information systems in new vehicles. The increasing complexity of interactions with in-vehicle equipment and the unprecedented amount of information streaming from these devices, however, create a palpable threat that drivers might find themselves overloaded with information that distracts them from the primary task of driving. This persistent and, in many cases, self-inflicted mental strain may cause driving performance decrements and lead to a substantial increase in the number of accidents with potentially grave consequences. One way to mitigate this issue is to study the driver's interactions with new in-vehicle technologies and use that knowledge to optimize system design and operating procedures. In order to accomplish this goal, we need an unobtrusive and objective measure of the driver's workload that not only quantifies average workload levels over long periods of time, but is also able to continuously capture workload variations throughout the task.

Workload is typically defined as the amount of mental or physical resources required to perform a particular task [19]. Its quantification is, unfortunately, difficult in practice because each individual's capacity of available resources varies greatly, as do the strategies for using them. The standard techniques used for workload assessment include self-report scales, performance-based metrics, and physiological arousal measures. Self-report measures are popular due to their low cost and consistency, though the latter quality assumes that the individual is cooperative and capable of introspection and accurate reporting of their perceived workload. Some of these scales are one-dimensional such as the Rating Scale of Mental Effort (RMSE) [24] and the Modified Cooper-Harper scale (MHC) [6], whereas some scales comprise subscales that measure specific mental resources, e.g., NASA Task Load Index (TLX) [9], Subjective Workload Assessment Technique (SWAT) [17], and Visual Auditory Cognitive Psychomotor method (VACP) [23]. The major drawback of these measures is that they cannot be unobtrusively administered during the task itself, but are assessed retrospectively at the conclusion of the task, which decreases accuracy of this technique. Furthermore, the inherent subjectivity of self-ratings makes across-subjects comparisons difficult. Self-report scales are, therefore, often complemented with an objective assessment of performance; this operates on the assumption that an increased workload diminishes performance. Performance measures include reaction time to different events, accuracy of responses, and overall driving performance such as steering wheel angle or lane position. The performance assessment is relatively unobtrusive and can be accomplished in real time at low cost as an indicator of actual workload level. Performance, however, is not sensitive enough to workload changes due to the complex relationship between the two variables. Performance is typically stable across a range of workload levels and deteriorates only near the extremes [18]. Moreover, performance measures cannot tap into all cognitive resources with comparable accuracy. Lately, there has been renewed interest in physiological measures as useful metrics for assessing workload. Their use was limited in the past by the obtrusive nature of earlier instrumentation, but this has changed with the advent of miniaturized sensors and embedded platforms capable of supporting complex signal processing techniques. Typically used physiological signals to derive measures of workload include: electrooculography (EOG) [7], electromyography (EMG) [22], pupil diameter [11], electrocardiography [20], respiration [5], electroencephalography [3], and skin conductance [19]. In some studies, physiological measures have been reported as being more sensitive to the initial changes in workload than performance-based measures, as they

AutomotiveUI'13, October 27–30, Eindhoven, The Netherlands.
Adjunct Proceedings.

show increased activation before the appearance of significant performance decrements [13]. This makes them more suitable for driving scenarios as they allow for an appropriate and timely intervention or mitigation. However, physiological workload measures have multiple drawbacks. First, the physiological workload scales are often derived empirically on a set of tasks assumed to represent different workload levels and selected ad hoc, without detailed consideration of their ecological validity and ability to tap into different mental resources (e.g., cognitive, visual, auditory, or motor workload). As a result, the models trained on such atomic tasks may not perform well when applied to the physiological signals acquired during other non-atomic tasks even though they seemingly require the same mental resources. Second, in spite of the well known fact of considerable between- and within-subject variability of nearly all physiological signals and metrics, the majority of physiological workload models have been developed and validated on a relatively small sample of subjects. Third, the classifiers used in the models introduced hitherto have typically lacked mechanisms for an adjustment of the model's parameters in relation to individual traits, which leads to models that do not generalize well. Finally, the models have mostly ignored the considerable amount of noise inherent in the acquired physiological signals. Thus, poor performance of some models could be attributed to their reliance on rather simple mathematical apparatus.

This paper introduces PHYSIOPRINT - a workload model based on the physiological measures of EEG and ECG that is built around a well defined and established theoretical workload model called Improved Performance Research Integration Tool (IMPRINT) [14]. The proposed model is able to distinguish between different workload types relevant for driving by incorporating complementary sensor modalities. The model is trained on a relatively large sample size, and it takes into consideration individual differences in physiological signals. The trained model is validated on an independent dataset recorded in a realistic driving simulator. Moreover, the utilized classification approach is not computationally expensive, so it is applicable in real time on a fine timescale.

The rest of the paper is organized as follows. In Section 2 we outline the experimental setting while Section 3 reports on the experimental results. Finally, in Section 4, we summarize our results and give an outlook on future work.

2. METHODS

In this section, we introduce the IMPRINT theoretical workload model that is used as a basis for the workload classes our PHYSIOPRINT workload model aims to classify. We also outline our study protocol, including both atomic and non-atomic tasks utilized for training and testing of the model, respectively. Lastly, we detail acquisition system together with the signal processing, data analysis, and evaluation procedures.

2.1 IMPRINT Workload Model

The IMPRINT Workload Model was developed by the Army Research Laboratory (ARL) [14] and it discriminates between seven types of workload: visual, auditory, cognitive, fine motor, gross motor, speech, and tactile. Each workload type is further quantified on a pertinent ordinal/interval scale, similar to the VACP scales [12]. Each of the seven scales is defined by a set of behaviors of increasing complexity that are associated with a numeric value between 0 and 7. Tasks and activities that mobilize more than one type of the mental resources receive separate independent scores on each of the relevant scales. Furthermore, for each point in time, IMPRINT produces a composite measure of the overall workload, which is defined as a weighted sum of the type-specific workload values calculated across all tasks that are being simultaneously performed. The weights in the formula describe the strength of all possible interactions (referred to as conflicts) between different workload types and/or different tasks. The IMPRINT model has been successfully applied to estimate mental workload in a number of settings of military relevance, including a strike fighter jet [4], a mounted combat system [16], and the Abrams tank [15]. As the model covers a large number of workload types, it is well-suited for the driving environment, which also employs distinct workload types. In this initial classifier development phase, we explored only a subset of the IMPRINT workload types: visual, auditory, cognitive, and fine motor.

2.2 Study Protocol

The PHYSIOPRINT workload classifier was developed and validated on the physiological data (EEG and ECG) acquired in two separate studies.

In the first study, physiological signals of 40 young healthy volunteers (17 females; age 26 ± 3 years) were recorded during four or five atomic tasks used for PHYSIOPRINT training to discriminate between the four IMPRINT workload types of interest: auditory, visual, cognitive, and fine motor. All subjects performed the auditory, visual, and cognitive tasks, while only a subset of 22 subjects also completed the fine motor control task. Each 1-sec segment of each task was assigned a score on each of the four scales. There was a dominant workload type in each atomic task, and the majority of 1-sec segments received a single non-zero workload score.

In the second study, six 10min scenarios were designed for the driving simulator by Systems Technologies Inc (STI). The physiological data recorded during these tasks were used for validation of the PHYSIOPRINT workload model. A total of 10 subjects took part in the experiment. The six test rides were taken in a random order, following a training ride at the beginning of the experiment. The rest period between the trials was 5min. The subjects also completed each atomic task once prior to the test rides. Again, all performed tasks were scored on each of the analyzed workload scales. Furthermore, the subjects provided self-reports of each driving scenario's difficulty after completion of the experiment.

2.2.1 Atomic Tasks

The atomic tasks were designed with the IMPRINT workload scales in mind. The goal for each task was to represent the corresponding workload type as closely as possible by engaging only the necessary mental resources to increase purity of the training data. The following atomic tasks were utilized:

Auditory Detection Task (ADET). The subject sits still for 5min in front of a blank computer screen and presses a button after hearing a beep.

Visual Detection Task (VDET). The subject sits still for 5min in front of a computer screen and presses a button whenever a geometrical shape appears on the screen.

Visual Discrimination Task (VDI). The subject sits still for 20min in front of a computer screen and presses a button if one target shape out of three possible geometrical shapes is shown on the computer screen. The shapes are randomly interspersed over time (the target shape is presented 70% of the time), and inter-stimulus interval ranges between 1.5sec and 10sec.

Figure 1. Driving Simulator at STI.

Figure 2. A subject wearing the wireless B-Alert sensor headset while performing an atomic task.

Forward/Backward Digit Span (FBDS). The subject sits still in front of a computer screen and memorizes sequences of 2 up to 9 digits that are shown on the computer screen and reproduces them by typing in the memorized sequence in the same or reverse order.

Fine Motor Control Task (FMCT). The subject holds a needle and inserts it into a target hole on a metal plate that is positioned at a 45º angle, and is instructed to keep the needle within the circular hole for 10sec without touching its perimeter. This was repeated for 5 holes whose diameters were 8, 7, 6, 5, and 4/32ths of an inch. The needle diameter was 3/64in.

Dominant workload types and the corresponding IMPRINT workload scores for the atomic tasks are as follows: ADET - auditory (1.0); VDET - visual (3.0); VDI - visual (5.0) and cognitive (3.7); FBDS - cognitive (5.3); and FMCT - fine motor (2.6) and visual (4.0).

2.2.2 Driving Simulator

The developed driving scenarios differed with respect to continuous visual-motor workload (related to the road curvature and a number of obstacles to be avoided) and the number of discrete events that were designed to cover a variety of activities on the visual, auditory, cognitive, and fine motor IMPRINT workload scales. There were three different sensory challenges during each scenario: (1) *auditory challenge* - honking (one of three possible patterns), (2) *visual challenge* - arrow signs pointing to one of the four possible directions, and (3) *cognitive challenge* - speed signs of two different colors (white and yellow) placed along the road requiring the subjects to add (if yellow) or to subtract (if white) the 3-digit numbers shown on the sign. The expected response to the visual and auditory challenges was a button press (*fine motor response*), verbal acknowledgment (*speech*) or no response (*cognitive action*), depending upon the arrow direction and the honking pattern. During the rides (Figure 1), the subjects sat on a gym bicycle whose front panel had been removed to avoid obscuring the view at the driving simulator screen. After the first 5min of the ride, the subjects were told to start pedaling till the end of the scenario (*gross motor* load).

The whole period from the onset of a particular stimulus (honking, arrow, and sign) till either its disappearance or the subject's response to it was considered a period with the dominantly auditory, visual, or cognitive workload, respectively. The 1-sec segments of the EEG and ECG that were completely or partially (>50%) covered by that period would consequently receive the same IMPRINT workload score (5.0 on the visual scale for the visual challenge, 6.6 on the auditory scale for the auditory challenge, and 7.0 on the cognitive scale for the mathematical challenge). The 2-sec periods centered around the subject's response (or, in case of the 'silent' response, the 2-sec period around the moment of the stimulus disappearance) received the appropriate score on the IMPRINT fine motor (score 2.2), speech (score 2.0), or cognitive (score 4.6) scales. The rest of the ride was scored with 4.4 on the visual ('visually track/follow') and 2.6 on the fine motor scale ('continuous adjustive control'). The portions with the pedaling also received a score of 3.0 on the gross motor scale.

2.3 Data Recording and Signal Processing

The wireless B-Alert sensor headset [1] (Figure 2) was used to acquire the EEG and ECG data of all subjects in the studies. The EEG data were recorded from 9 sites on the head (F3, F4, Fz, C3, C4, Cz, POz, P3, and P4 locations of the 10-20 international system), referenced to link mastoids. The ECG data were recorded from two electrodes placed on the left and right collar bone. All signals were filtered with a band-pass filter (0.1-70Hz, roll-off: 20dB/decade) before the analog to digital conversion (256Hz, 16 bits/sample), and transferred in real time via Bluetooth link to a nearby PC where the data was stored onto a disk. The sharp notch filters were applied to remove environmental artifacts from the power network. The algorithm [1] was utilized to automatically detect and remove a number of artifacts in the time-domain EEG and ECG signals, such as spikes caused by tapping or bumping of the sensors, amplifier saturation, or excursions that occur during the onset or recovery of saturations. Eye blinks and EMG were identified and decontaminated by an algorithm [2] based on wavelet transformation. Eye blinks and EMG bursts were also used as binary variables (present/absent) in the PHYSIOPRINT workload model.

From the filtered and decontaminated EEG signal, the absolute power spectral densities (PSD) were calculated for each 1sec epoch of data by applying the short-term Fourier transformation (STFT). The following PSD bandwidths were extracted: delta, theta slow, theta fast, theta total, alpha slow, alpha fast, alpha total, sigma, beta, and gamma. In order to account for individual differences in the EEG data, we also utilized relative PSD values by subtracting the logged absolute PSD values for each 1Hz bin

from the total logged PSD in the bandwidth of interest. Wavelet coefficients were also derived for each EEG channel in the exponential 0-2, 2-4, 4-8, 8-16, 16-32, and 32-64Hz bands. In some rounds of the model development, the same variables were extracted from the left-right and anterior-posterior differential EEG derivations that were constructed by subtracting the pertinent referential signals (i.e., Fz-POz, Cz-POz, F3-P3, C3-P3, F4-P4, C4-P4, F3-F4, C3-C4, and P3-P4). The proprietary physiological measure of alertness and mental fatigue (MF) was also calculated from the Fz-POz and Cz-POz derivations using our validated B-Alert algorithm [10]. The ECG signal was processed by a real-time algorithm that determined the inter-beat (R-R) intervals and heart rate. Measures of the heart rate variability (HRV) were derived from the R-R time series, such as NN50/NN20 (number of successive R-R intervals in the past 10sec that differ by more than 50ms and 20ms, respectively) and RMSSD (the square root of the mean squared difference of successive RR intervals). All the extracted variables were then also averaged over a 5sec sliding window in 1sec increments to include a short term history.

2.4 Data Analysis

The goal of the data analysis was to test four hypotheses:

H1: Classification results will be increased if a combination of complementary input signals (EEG and ECG) is relied on instead of a single modality (EEG).

H2: Classification results will be increased if multiple EEG channels from different areas of the scalp are utilized as opposed to reliance on only a few channels from adjacent regions.

H3: Classification results will be increased if concurrent measurement of levels of fatigue and alertness is performed and these measures are fed to the classifier.

H4: Classification results will be increased if the workload model relies on relative variables and descriptors of a period of time leading to the current moment and not only on descriptors of the current point in time.

The predictor variables were identified by the step-wise variable selection procedure on all available data. To test the hypotheses H1-H4, variable selection was repeated several times within four different feature spaces:

FS1 - the EEG variables derived only from the referential channels (EEG-REF);

FS2 - the EEG variables derived from both referential (EEG-REF) and differential channels (EEG-DIFF);

FS3 - the EEG-REF, EEG-DIFF and all ECG variables; and

FS4 - the EEG-REF, EEG-DIFF, ECG and mental fatigue scores (MF), i.e. all available variables.

Two separate rounds were conducted in each of the four feature spaces:

- 'No history' round, where the feature vectors included only variables calculated on the current segment, and

- 'Short-term history' round, where the feature vectors included averaged variables calculated for each of the 5sec prior to the current segment.

The selected variables were then used for building PHYSIOPRINT, which is a two-level classifier depicted in Figure 3. The first level outputted the dominant and second-dominant workload (WL) types: WLD and WLS, respectively. It included four independent classifiers, linear discriminant function analysis (L-DFA) [8] that fitted a multivariate normal density to each class

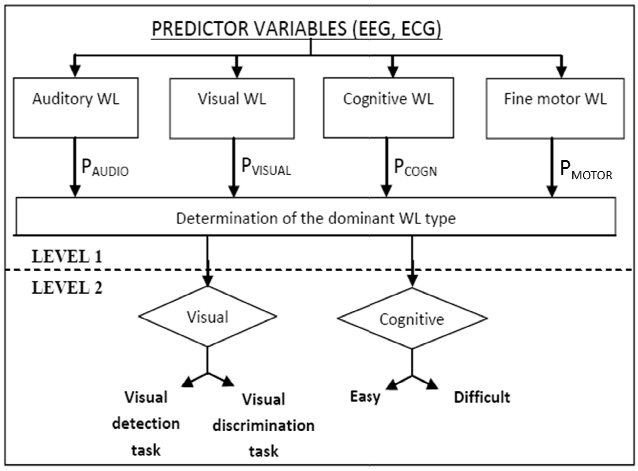

Figure 3. Two-level PHYSIOPRINT classifier.

with a pooled estimate of covariance. Based on the likelihood estimates and prior probabilities, the posterior probabilities (i.e., P_{VISUAL}, P_{AUDIO}, P_{COGN}, and P_{MOTOR}) that the given segment of data originated from a visual, auditory, cognitive, or fine motor workload task were calculated, respectively. The four classifiers were followed by a 'winner takes all' block that declared the WL type with the highest probability as the dominant type (WLD). In some cases, the second-dominant WL type was defined as the type with the second highest probability, but only if that probability exceeded a fixed threshold (PTH = 0.3). Level 2 of the PHYSIOPRINT classifier further quantified workload intensity within the dominant WL type. Level 2 comprised only two L-DFA classifiers: one that differentiated between the visual detection (score = 3.0) and visual discrimination task (score = 5.0), and one that further classified the cognitive task as easy (1-3 digits) or difficult (4-9 digits).

2.5 Evaluation Procedure

Once the predictor variables were selected for each combination of the feature spaces and history, the WL type-specific classifiers were evaluated using the leave-one-subject-out approach to assess the generalization capabilities of the classifier by testing it on the data that was not used for training. The model was first trained on all pertinent segments from 39 subjects (21 in the case of the Fine Motor WL classifier) and then tested on the remaining subjects. The procedure was repeated for all subjects in the study, and the results were averaged across all cross-validation rounds. This is a standard approach in the literature when dealing with relatively small sample sizes.

Furthermore, validation of the PHYSIOPRINT classifier was performed on the driving simulator data. Given the relatively small sample size in the driving simulator experiment, we limited our analyses to cross-validation of the Level 1 PHYSIOPRINT model. Only the best performing model was employed, and only its ability to recognize the dominant workload type was tested, i.e., the output of Level 2 was ignored at this stage.

3. EXPERIMENTAL RESULTS

In this section, we report on the most discriminative variables, PHYSIOPRINT classification results, driving simulator performance results, and cross-validation of the PHYSIOPRINT model on the driving simulator data.

3.1 Selected Variables

The variables selected in each round and feature space varied in number and type, but certain trends were observable in each round: (1) When the differential EEG variables were part of the feature space, they were dominantly selected, especially PSD bandwidth variables for the inter-hemispheric derivations (F3-F4, C3-C4, P3-P4); (2) When the ECG variables were part of the feature space, the heart rate (HR) variable was always selected as significant (but most of the HRV variables were not); (3) When the B-Alert measure of mental fatigue was included in the feature space it was selected; (4) The EEG variables mostly came from the theta (3-7Hz) and beta (13-32Hz) range; (5) The binary eye blink variable was selected for the Visual and Cognitive WL classifiers, but not the Auditory or Fine Motor WL classifiers; and (6) The EMG bursts derived from the EEG channels were never selected as significant.

3.2 PHYSIOPRINT Classification Results

In this section, we present classification results for both Level 1 (i.e., differentiation among different workload types) and Level 2 (i.e., differentiation between the workload levels on the same WL type scale) of the PHYSIOPRINT model.

3.2.1 Level 1 Classification Results

The summary results for all combinations of the feature space (FS1-FS4) and feature vector duration (1sec vs. 5sec) are shown in Table 1 for the visual, auditory, and cognitive workload type. As one can observe, the results confirm all four hypotheses (H1 - H4), and show that the multi-channel EEG and ECG signals successfully differentiate between the auditory, visual, and cognitive WL types (>80% precision/recall). Comparatively, the largest improvements were achieved with the addition of the differential EEG channels (~8% increase on average for the same feature vector duration), and with the increase in the feature vector duration (~4%-10% increase, depending on the WL type and feature space); the addition of ECG variables and EEG-based mental fatigue (MF) measures brought about moderate improvements (2-4% depending on the WL type). These variables may, however, be more important in situations when high stress is experienced (ECG), or when more complex visual or auditory tasks are pursued (MF).

The Fine Motor WL classifier's accuracy was not shown in the same table because this portion of the PHYSIOPRINT model could only be tested on a subset of subjects who had performed the FMCT task. The accuracy of this classifier showed similar trends (i.e., an increase with the addition of the differential EEG, ECG, MF and/or extension of the feature vector from 1sec to 5sec), but the values were relatively lower for any tested combination of the feature space and feature vector duration. The highest recall and precision – 62.7% and 68.3%, respectively – were obtained with all variable types (i.e., EEG-REF, EEG-DIFF, ECG, and MF) and 5sec long feature vectors. The data segments from the FMCT task were typically misclassified as 'Visual WL'. We attribute this, at least in part, to a substantial overlap between fine motor and visual workload during the execution of the fine motor (FMCT) task. Indeed, the Fine Motor WL was identified as the second-dominant WL type in 30% - 40% of the misclassified segments (the exact proportion varied with the feature space and feature vector duration). Therefore, the modest accuracy of identification of the Fine Motor WL type seems to be related to the impurity of the task that was nominally declared as the fine motor control task.

Table 1. Recall (REC, %) and precision (PREC, %) of the Level 1 PHYSIOPRINT model for the auditory, visual, and cognitive WL type and different combinations of the features.

	Visual WL		Auditory WL		Cognitive WL	
Fuature space	REC	PREC	REC	PREC	REC	PREC
EEG-REF, no-history (NH)	72.4	73.6	68.8	68.2	55.9	57.3
EEG-REF, 5-second history (5H)	74.6	78.6	72.4	73.0	63.2	61.7
EEG-REF, EEG-DIFF, NH	79.7	79.1	72.2	76.9	62.2	64.7
EEG-REF, EEG-DIFF, 5H	88.9	86.4	78.6	83.2	73.8	75.2
EEG-REF, EEG-DIFF, ECG, NH	81.9	80.0	75.5	80.1	64.9	67.4
EEG-REF, EEG-DIFF, ECG, 5H	91.3	89.7	81.2	87.4	76.3	77.8
EEG-REF, EEG-DIFF, ECG, MF, NH	85.1	80.3	76.2	81.6	67.3	69.2
EEG-REF, EEG-DIFF, ECG, MF, 5H	**94.1**	**91.8**	**83.1**	**89.9**	**79.1**	**80.1**

3.2.2 Level 2 Classification Results

Given the aforementioned findings, the classification accuracy at Level 2 was assessed only for the combination of the all-inclusive feature space (EEG-REF, EEG-DIFF, ECG, and FM) and 5sec long feature vectors. For the visual workload tasks, the recall and precision were (REC/PREC): 78.8%/76.4% for the visual detection task and 93.1%/93.4% for the visual discrimination task. For the cognitive tasks, the recall and precision were (REC/PREC): 75.4%/74.1% for the easy/short digit sequences and 76.8%/77.5% for the long/difficult digit sequences.

3.3 Validation on the Driving Simulator Data

In this section, we first present performance results on the driving simulator, and then validation of the PHYSIOPRINT model on the physiological data recorded during driving simulation scenarios.

3.3.1 Driving Simulator Performance Results

In order to test the validity of our simulated driving task, we analyzed the subjects' performance on the driving simulator. There were a total of 780 discrete visual challenges (57 per subject across all six rides) with an equal split of expected reactions (260 button presses, 260 verbal acknowledgments and 260 silent responses); a total of 780 auditory challenges (with equal split among the expected responses); and a total of 360 cognitive challenges (3-digit numbers, half of them positive, half negative). In general, the subjects responded accurately to visual and auditory stimuli (94.1% accurate responses to auditory and 90.5% to visual challenges), but had more problems with the mathematical (cognitive) task, as the subject arrived upon the correct result at the end of the ride in only 41 out of 60 rides (68.3%). The majority of the reported results were, however, within ±10 of the correct result (57 out of 60, or 95%), which we interpreted as a sign that the subjects adequately engaged their cognitive resources and aimed at responding to the challenge (addition and subtraction of 3-digit numbers), even though their affinity/talent for math varied. In general, more errors were made during the two most difficult rides (92.6% average accuracy of responses to auditory, 87.4% to visual, and 55% to cognitive challenges), while the performance was notably better on the other four rides (96.1% for auditory, 92.7% for visual, and 75% for cognitive challenges). There was no significant difference in performance between the portions of the ride without the pedaling and those while the subjects had to pedal. Self-reports corresponded to the objective findings: the subjects mostly complained about the mathematical task and reported the two objectively most difficult rides to be significantly more challenging than the other four.

3.3.2 Classification of Driving Simulator Data with PHYSIOPRINT Workload Model

The PHYSIOPRINT classification accuracy was in general slightly lower on the data from the driving simulator study than it had been on the atomic tasks. Recall and precision (REC/PREC) during the periods with dominantly visual workload were 80.6% and 78.3% across all subjects and rides. Recall and precision during the periods with dominantly auditory workload were 71.5% and 73.6%, whereas recall and precision during the periods with dominantly cognitive workload were only 64.7% and 62.1%. When the PHYSIOPRINT classifier was applied to the subject's atomic tasks (i.e., VDET, VDI, FBDS, and ADET), accuracy increased (REC/PREC: 85.2%/78.3% for the VDET+VDI tasks, 74.9%/77.3% for the ADET task, and 76.3%/75.7% for the cognitive FBDS task). The classification accuracy was, on average, ~5% worse during the portions with the pedaling, which suggested that changes in heart rate and heart rate variability have relatively modest effects on this version of the classifier. The drop in performance could not be attributed to an increased level of noise in the signals (asserted by visual inspection). The modest increase in the classification accuracy when the classifier was applied to the atomic tasks on which the model was trained (VDET, VDI, ADET, and FDBS) suggested that the between-subject variability played a role, but was not the only or major reason for the drop in classification accuracy in the driving simulator study. It is possible that the overlap between the different workload types throughout the majority of the ride confused the classifier, and that the results could improve once more sophisticated mechanisms for detection and resolutions of such conflicts are built into the classifier.

4. SUMMARY AND OUTLOOK

The current study sought to develop a physiologically-based method for workload assessment applicable in the challenging automotive setting. We addressed this need by designing a comprehensive, sensitive, and multifaceted workload assessment tool that incorporates the already established theoretical workload framework that both: (1) covers the different types of workload employed in complex tasks such as driving, and (2) helps define the necessary atomic tasks for building the model. The experimental results suggested that the classifier benefits from combination of complementary input signals (EEG and ECG), better coverage of the scalp regions by an increased number of EEG channels, inclusion of concurrent physiological measurement of fatigue and alertness levels, and short-term signal history. We aimed to overcome the individual variability inherent in the physiological data by including the relative PSD variables in the feature vector. The generalization capability of the trained model was tested by using leave-one-subject-out cross-validation, as well as testing the model on the independent driving simulator dataset. The proposed method demonstrated that integration of physiological monitoring into automotive settings holds great promise for real time assessment of the driver's workload.

In the future, we plan to extend the model to cover all workload types (visual, auditory, cognitive, fine motor, gross motor, speech, and tactile) together with the corresponding workload intensity level subscales from the IMPRINT workload model. In order to achieve this, we need to design new atomic tasks carefully. We must also refine the existing tasks, especially the FMCT task that proved not ideal for representing pure fine motor activity. Additional physiologically based inputs, such as EOG, EMG, respiration, and stress levels will also be included to enable better insight into activations of different workload types. Alternative classification algorithms such as multi-label learning [21] will be evaluated to facilitate the process of resolving the conflicts between different workload types. The final global workload score will be a composite measure of all seven resource-specific workload type scores (analogous to the overall IMPRINT workload score). The weights will be designed in a way that also considers the influence of environmental factors, workload management strategies, and other individual traits and their effect on the overall engagement level of mental resources. The classifier will be validated on a larger sample of subjects performing a variety of tasks in both laboratory and real-life environments (i.e., real car).

The ultimate PHYSIOPRINT workload assessment tool is envisioned as a flexible software platform that consists of three main components: (1) an executable that runs on a dedicated local (client) machine to acquire multiple physiological signals from one or more subjects, processes them in real time, and determines global and resource-specific workload on a fine time scale; (2) a large server-based database of physiological signals acquired during relevant atomic tasks from a large number of subjects with different socio-demographic and other characteristics (e.g., degree of driving experience); and (3) a palette of real-time signal processing, feature extraction, and workload classification algorithms. The platform will support a number of recording devices from a wide range of vendors (via the appropriate device drivers), and enable visualization of the workload measures. The users will essentially be able to build their own workload assessment methods from the available building blocks of feature extraction methods and implemented classifiers. Initially, the database will include 100-150 subjects, but we envision that the database will continue to evolve as the community grows in the following years.

5. ACKNOWLEDGMENTS

This work was supported by the Army Research Laboratory grant W91CRB-13-C-0007. The authors would like to thank Stephanie Korszen for her excellent editing advice.

6. REFERENCES

[1] Berka, C., Levendowski, D.J., Cvetinovic, M.M., Petrovic, M.M., Davis, G., Lumicao, M.N., Zivkovic, V.T., Popovic, M.V., Olmstead, R. 2004. Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset. *International Journal of Human-Computer Interaction.* 17, 151-170.

[2] Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L. 2007. EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. *Aviation, Space, and Environmental Medicine.* 78, 231-244.

[3] Berka, C. et al. 2005. Evaluation of an EEG-Workload Model in an Aegis Simulation Environment. In *Proceedings of SPIE Defense and Security Symposium, Biomonitoring for Physiological and Cognitive Performance during Military Operations.* SPIE: The International Society for Optical Engineering: Orlando, FL, 90-99.

[4] Brett, B.E., Doyal, J.A., Malek, D.A., Martin, E.A., Hoagland, D.G., Anesgart, M.N. 2002. *The Combat Automation Requirements Testbed (CART) Task 5 Interim Report: Modeling a Strike Fighter Pilot Conducting a Time*

Critical Target Mission. Technical Report. AFRL-HE-WP-TR-2002-0018.

[5] Brookings, J.B., Wilson, G.F., Swain, C.R. 1996. Psychophysiological Responses to Changes in Workload During Simulated Air Traffic Control. *Biological Psychology.* 42, 3, 361-377.

[6] Casali, J.G., Wierwille, W.W. 1983. A Comparison of Rating Scale, Secondary-Task, Physiological, and Primary-Task Workload Estimation Techniques in a Simulated Flight Task Emphasizing Communications Load. *Human Factors: The Journal of the Human Factors and Ergonomics Society.* 25, 6, 623-641.

[7] Galley, N. 1993. The Evaluation of the Electrooculogram as a Psychophysiological Measuring Instrument in the Driver Study of Driver Behavior. *Ergonomics.*36, 9, 1063-1070.

[8] Hardle, W., Simar, L. 2007. *Applied Multivariate Statistical Analysis.* Springer Berlin Heidelberg.

[9] Hart, S.G., Staveland, L.E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload*, Amsterdam, North Holland Press, 139-183.

[10] Johnson, R.R., Popovic, D.P., Olmstead, R.E., Stikic, M., Levendowski, D.J., Berka, C. 2011. Drowsiness/Alertness Algorithm Development and Validation Using Synchronized EEG and Cognitive Performance to Individualize a Generalized Model. *Biological Psychology.* 87, 241-250.

[11] Kun, A.L., Medenica, Z., Palinko, O., Heeman, P.A. 2011. Utilizing Pupil Diameter to Estimate Cognitive Load Changes During Human Dialogue: A Preliminary Study. *In Adjunct Proceedings of the Automotive User Interfaces and Interactive Vehicular Applications Conference (Salzburg, Austria, 2011).* AutomotiveUI'11.

[12] McCracken, J.H., Aldrich, T.B. 1984. *Analyses of Selected LHX Mission Functions: Implications for Operator Workload and System Automation Goals.* Research Note. ASI479-024-84B. Fort Rucker, AL.

[13] Mehler, B., Reimer, B., Coughlin, J.F., Dusek, J.A. 2009. The Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record.* 2138, 6-12.

[14] Mitchell, D.K. 2000. *Mental Workload and ARL Workload Modeling Tools*. Final Technical Report. ARL-TN-161. Army Research Laboratory, Aberdeen Proving Ground MD.

[15] Mitchell, D.K. 2009. *Workload Analysis of the Crew of the Abrams V2 SEP: Phase I Baseline IMPRINT Model.* Final Report. ARL-TR-5028. Army Research Laboratory, Aberdeen Proving Ground, MD.

[16] Mitchell, D.K., Samms, C.L., Henthorn, T., Wojciechowski, J.Q. 2003. *Trade Study: A Two- Versus Three-Soldier Crew for the Mounted Combat System (MCS) and Other Future Combat System Platforms.* Technical Report. ARL-TR-3026. U.S. Army Research Laboratory: Aberdeen Proving Ground, MD.

[17] Reid, G.B., Nygren, T.E. 1988. The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Advances in Psychology*. 52, 185-218.

[18] Son, J., Park, S. 2011. Cognitive Workload Estimation through Lateral Driving Performance. In *Proceedings of the 16th Asia Pacific Automotive Engineering Conference* (Chennai, India, October 6-8, 2011). APAC'11. SAEINDIA, India, SAE2011-28-0039.

[19] Son, J., Park, M. 2011. Estimating Cognitive Load Complexity Using Performance and Physiological Data in a Driving Simulator. *Adjunct Proceedings of the Automotive User Interfaces and Interactive Vehicular Applications Conference (Salzburg, Austria, 2011).* AutomotiveUI'11.

[20] Trutschel, U., Heinze, C., Sirois, B., Golz, M., Sommer, D., Edwards, D. 2012. Heart Rate Measures Reflect the Interaction of Low Mental Workload and Fatigue During Driving Simulation. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Portsmouth, NH, 2012).* AutomotiveUI'12.

[21] Tsoumakas, G., Katakis, I. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining.* 3, 1-13.

[22] Vilage, J., Frazer, M., Cohen, M., Leyland, A., Park, I., Yassi, A. 2005. Electromyography as a Measure of Peak and Cumulative Workload in Intermediate Care and its Relationship to Musculoskeletal Injury: An exploratory Ergonomic Study. *Applied Ergonomics.* 36, 5, 609-618.

[23] Yee, S., Nguyen, L., Green, P., Oberholtzer, J., Miller, B. 2007. *Visual, Auditory, Cognitive, and Psychomotor Demands of Real In-Vehicle Tasks*. Technical Report. UMTRI-2006-20. University of Michigan Transportation Research Institute.

[24] Zijlstra, F.R.H. 1993. *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. PhD Thesis. Delft University of Technology, Delft, The Netherlands, Delft University Press.