Multimodal Recognition of Cognitive Workload for Multitasking in the Car

Felix Putze, Jan-Philip Jarvis, Tanja Schultz
Karlsruhe Institute of Technology (KIT), Cognitive Systems Lab (CSL)
felix.putze@kit.edu, jan.jarvis@student.kit.edu, tanja.schultz@kit.edu

Abstract

This work describes the development and evaluation of a recognizer for different levels of cognitive workload in the car. We collected multiple biosignal streams (skin conductance, pulse, respiration, EEG) during an experiment in a driving simulator in which the drivers performed a primary driving task and several secondary tasks of varying difficulty. From this data, an SVM based workload classifier was trained and evaluated.

1 Introduction

Dialog systems have matured to a point where they make their way to real-world applications. However, especially in more dynamic environments like the car, dialog systems still struggle to provide a natural and fluid dialog experience. This is to a great extent due to the lack of understanding of human cognitive processes with changing user states, non-perfect users and influences of affect on the interaction. We have the goal to build a cognitive dialog system [7], i.e. a dialog system that has an understanding of human cognition and can adapt to varying types and states of users. One of the most important aspects of human cognition in the context of such systems is the level of cognitive workload the user experiences. To be able to classify the current workload level in a subject independent way, we collect multimodal biosignal streams from the driver from which we derive predictive features for classification. As measuring biosignals in a natural, non-constrained driving situation will produce noisy signals, we measure and combine multiple data streams to end up with a more reliable classifier.

2 Related Work

Liang, Reyes und Lee [6] developed a real-time workload classifier in the car using facial features, like pupil diameter or gaze direction, extracted from videos of the driver. The 10 participants followed a car with varying speed while performing a secondary memory and comparison task. Using Support Vector Machines, the authors achieved a recognition rate of 81.1% on average for the recognition of cognitive workload.

Healey and Picard [2] developed a classifier to monitor the stress levels in daily life car driving tasks. They collected data from twenty-four real-life drives of at least fifty minute duration and used the biosignals electromyography, electrocardiography and skin conductance for their system. Linear discriminant analysis (LDA) was used for dimensionality reduction and a classifier using a linear decision function was able to discriminate the three classes with accuracies of 100% (low workload), 94.7% (medium workload), and 97.4% (high workload).

Honal and Schultz [3] conducted various experiments to evaluate levels of task demand from EEG for 8 participants. Their focus was evaluation in a lecture and meeting scenario. Features are derived by Short time Fourier transform on overlapping windows of two seconds length and treated with a correlation-based feature reduction method. Finally, Support Vector Machines and Artificial Neural Networks are used for classification and regression. For the prediction of low versus high task demand during a presentation the authors obtained 92% accuracies in session dependent experiments and 80% in subject independent experiments.

Our described system differs from earlier work in its compilation of modalities, in the smaller size of available training data and in the systematic investigation of different secondary tasks.

3 The Experimental Setup

In order to collect data for training and evaluation of a workload classifier, we designed the following experimental setup: During the whole experiment, the participants sit in a realistic driving simulator (see [7] for details). In this simulator, the drivers perform the *Lane*

Change Task (LCT) [5], a standardized driving task in which the driver has to drive on a three-lane highway and has to change lanes as indicated by signs along the route. There are no other vehicles or junctions on the route and the drivers are told to keep a fixed speed.

Each experiment consists of several sections. Initially after introduction and initialization of the measurements, the driver performs a driving training run without any secondary tasks to get used to the characteristics of the driving simulator. Then, the driver goes through four sections of driving with a secondary task, each of three minutes length and seperated by pauses of the same length[1]. These four sections confront the driver with a secondary task of increasing difficulty. The first section is used as training section in which the driver gets used to the task characteristics and strategies. The following three sections present the same task with rising difficulty levels, to record sections of low, medium and high cognitive workload. For evaluation, we only use a window of one minute at the beginning of each section to extract the features. This duration gives a reasonable trade-off between feature robustness and flexibility. We can also assume the measured workload level to be relatively stable within this window while longer windows become prone to fatigue effects.

We employed two different secondary tasks: a visual search task and a mathematical cognitive task. The visual task presents screens with randomly arranged symbols of different forms (distractors) on a display in the cockpit and asks the driver to identify one symbol, which differs from the distractors in line thickness, and signal its type by pressing a button. After each decision, the next screen is presented. The difficulty level of this task is controlled by how evidently the target symbol is different from the distractors. For each run, we calculate an error score which rewards correctly solved screens while penalizing wrong answers to encourage both quick and correct answers. In the cognitive task, a prerecorded list of numbers is read to the driver who is asked to signal via button pressing whether the last number is divisible by a given divisor. The divisor controls the difficulty level: a divisor of 2 results in the easiest task while a divisor of 7 results in the highest task level. The time between two numbers is fixed and numbers are taken from the interval $[10d, 30d]$, where d is the current divisor. An error score for each run is derived as the relative number of wrongly solved items.

During the experiment, we gather a variety of signals from the driver in the car to get a reliable, continuous data stream without obstructing or distracting the user too much. We employ the following equipment to observe the user: Electroencephalography (EEG) data is recorded with a comfortable headband mainly from the prefrontal cortex (positions Fp1, Fp2, F7 and F8 in the international 10-20 positioning system). A light wireless sensor glove on the left hand of the driver measures skin conductance (SC) and heart rate via photoplethysmography (PPG). Finally, a respiration belt (RESP) on top of the clothes measures the respiration frequency and respiration strength. All sensors use the same recording interface by Becker MediTec and are either attached to a universal signal recorder or directly connected to the recording computer via Bluetooth, which reduces obstruction to a minimum.

	TLX	**ERR**	**LEV**	μ_{LCT}	σ_{LCT}
TLX	1.00	0.83	0.82	0.45	0.43
ERR	0.83	1.00	0.94	0.54	0.46
LEV	0.82	0.94	1.00	0.43	0.32
μ_{LCT}	0.45	0.54	0.43	1.00	0.80
σ_{LCT}	0.43	0.46	0.32	0.80	1.00

Table 1. Correlation coefficients of potential groudtruth scores for the visual task: Nasa TLX, Secondary Task Error (ERR), Task Diffculty Level (LEV), mean (μ_{TLX}) and std. dev. (σ_{TLX}) of deviation from ideal route.

To create classes for workload classification, there are several different thinkable approaches: Directly use the pre-defined task difficulty, use an objective error metric derived on the performance scores for primary and secondary tasks, or use subjective workload as for example measured via the standardized NASA Task Load Index (TLX) [1]. Table 1 shows that task difficulty, subjective workload and secondary task error measures are all strongly correlated using Spearman's Rank Correlation Coefficient[2]. The correlation between those three and the driving error measures is much less pronounced. This leads to the conclusion that the task difficulty level is a good indicator for both the experienced cognitive workload and the severity of distraction. It also shows that the drivers associate workload mainly with the secondary task, which may indicate that they always reserve the same cognitive capacity for the main driving task and assign only the remaining attention to the secondary task. Based on those observations, we chose to use task difficulty to form workload classes.

[1] We still record data during the pauses between the sections to generate a baseline for the measured biosignals.

[2] Table 1 contains data for the visual task only. However, correlation coefficients for the cognitive task are very similar.

4 Implementation

From the recorded streams, we extract an initial feature set for classification. Most features are well known physiological measures which are meaningful in the context of workload classification. We extract one feature vector for each experimental section and modality:

EEG: Before feature extraction from the EEG signal, we have to account for the frequent artifacts in the EEG signal (e.g. caused by eye movement). We employ a simple artifact detector which compares the signal amplitude with a fixed threshold and removes the affected segments from the data stream[3]. We then transform the signal to the spectral domain using Welch's method. The spectrum is smoothed by averaging k (here, $k = 2$) adjacent frequency bins to form the feature vector.

PPG: Preprocessing of the PPG signal simply consists of a subtraction of the mean amplitude in the baseline resting section to diminish interpersonal differences. From the preprocessed signal, we extract pulse frequency, pulse frequency variance and heart rate variability (which describes the ratio between the influence of the sympathetic and the parasympathetic nervous system on the driver's pulse). The feature extraction is similar to the one given by [2].

SC: For the skin conductance signal, we first perform the same preprocessing as for the PPG signal, then apply a moving average filter with a window size of 512 samples. In the smoothed signal, we look for rapid rises in the time domain, so called *startles* which we extract by inspecting the signal's slope [2]. For every startle, we extract its peak and its base point. From those points, we derive duration and height of the startle. As features for the whole window, we use the total number of startles, the sum of all startle durations, the sum of all startle heights and an approximation of the signal area covered by all startles. Those features are combined with the mean and variance of the smoothed signal.

RESP: For the respiration belt signal, we do preprocessing and startle extraction as we do for the SC signal. In addition to band-power spectral features, we use the number of startles to derive the mean respiration rate, the variance in the distance between two startles as the variance in respiration rate, and the average startle height as mean respiration strength.

The initial feature set consists of 164 features for EEG, 3 features for PPG, 6 features for SC and 8 features for RESP. We do not employ the whole set for classification as we try to avoid overfitting of the trained classifier, given the relatively small data set. Therefore, we use Forward Feature Selection (FFS) as a wrapper based feature selection approach. In the FFS scheme, we iteratively increase a tentative feature set by adding the feature which increases the recognition rate the most. Recognition rate is estimated by training and evaluating a classifier in a cross-validation. As FFS may be prone to overfitting the feature set to characteristics of the training data, we take some measures to counter this effect: Before the selection process is started, features are ordered by descending correlation to ground truth on the training data which lets the FFS prefer those features which are indicated as predictive by the correlation coefficient. The FFS greatly reduces the number of employed features to less than two features per modality on average.

For classification, we use a Support Vector Machine with kernels based on Radial Basis Functions. The parameters C and γ are tuned via cross-validation with the final feature set according to the grid search approach proposed in [4]. Before classification, features are normalized across drivers using z-normalization and range normalization to generate comparable and consistent feature values. To combine multiple modalities for a more robust classification, we use the following decision fusion (DF) scheme: We train and evaluate one classifier for each modality independently and to benefit from multiple modalities, we perform a majority vote among all single-modality classifiers to derive a single multimodal hypothesis. Votes are weighted based on training set recognition performance.

5 Evaluation

In the setup described in section 3, we collected an experimental corpus for which we recorded a total of 13 sessions from eight male and five female drivers. Each session yields six minutes of non-baseline data (six sections of one minute each), resulting in a total of 78 minutes. We use this data to train and evaluate a workload classifier. As our experimental corpus is small, we use the cross-validation scheme on session level[4] to estimate the recognition rate of the classifier on unseen data in a subject-independent way. As the cross-validation pattern is also applied for estimating SVM parameters and the feature set, we actually perform *nested cross-validation*, where each iteration of the outer cross-validation (to estimate the recognition performance on the hold-out set) performs a complete inner cross-validation to tune the recognizer parameters

[3] Note that it is not necessary to try to reveal information from obviously artifact contaminated data sequences to derive features for workload classification, as long as enough clean data remains.

[4] We work on session level to keep every participant either completely in the test set or in the training set. Partitioning on session level generates test sets containing six experiment sections all from the same driver.

Task	EEG	PPG	SC	RESP	DF
Vis_2	77 (5)	79 (3)	82 (0)	91 (4)	**95** (4)
Vis_3	58 (3)	50 (10)	66 (5)	**72** (7)	70 (9)
Cog_2	66 (18)	70 (14)	**76** (4)	55 (11)	73 (15)
Cog_3	41 (5)	**47** (17)	38 (10)	38 (6)	43 (5)
$AltCog_2$	72 (6)	67 (11)	87 (3)	74 (4)	**97** (2)
$AltCog_3$	45 (6)	36 (6)	62 (7)	49 (6)	**65** (4)

Table 2. Recognition rates in percent, std. dev. for window offsets in parantheses

on the outer training set. We separate the experiments in two dimensions: Firstly, we look at the visual task and the cognitive task separately. Secondly, we regard the three-class classification problem and a two-class problem where the middle class is dropped, such that the classifier has to only discriminate low and high workload.

Table 2 summarizes the cross-validation recognition rates. We average results across a range of different offsets for the one-minute window from which features are extracted. From the table, we see that our system achieves a multimodal recognition rate of 95% for the two-class visual task (Vis_2) which drops to 73% for the cognitive task (Cog_2). For the full three-class problem, we get recognition rates of 70% for the visual (Vis_3) and 43% (Cog_3) for the cognitive task. As table 2 shows, the standard deviation for recognition rates under variation of the offset are quite high, indicating that improvements in stability are still necessary.

We see that for all tasks, the fusion result yields the best recognition rate or is within 4% absolute of the best modality. Given that the best modality varies from task to task, doing multimodal recognition seems to be the best choice for a generic workload recognizer.

While the recognition rates for the visual task are very satisfactory given the small amount of training data, rates drop significantly for the cognitive task. Our hypothesis is that those tasks were too difficult for most drivers and that even the sections with low and medium task difficulty levels produced workload patterns which were already too similar to the ones observed for the high task difficulty[5]. To investigate this hypothesis, we trained another classifier with shifted classes: baseline vs. 'cognitive low' as two-class problem ($AltCog_2$) and baseline vs. 'cognitive low' vs. 'cognitive high' as three-class problem ($AltCog_3$). The two final rows in table 2 show that recognition rates improve over the original cognitive classification problems.

EEG	PPG	SC	RESP	DF
83 (3)	50 (8)	74 (7)	71 (4)	84 (3)

Table 3. Recognition rates for task type classification.

As we employed two different secondary tasks that need different cognitive resources, we also built a separate classifier to differentiate the task types using the same features as for cognitive workload. The results are given in table 3. It shows that the tasks can be discriminated with an accuracy of 84%, to which the EEG signal contributes the most by far. We assume this is due to eye movement artifacts which show different characteristics and frequency for a visual task than for the cognitive one.

6 Conclusions & Future Work

Our current system is able to reliably detect a stable level of workload on one-minute windows of data. During a real-life driving session, one will also encounter rapid or steady changes of parameters. Therefore, we need to improve on continously monitoring the driver's state by detecting those transitions and modeling them, for example in a dynamic model. For continuous prediction, it may also be helpful to re-formulate the task as a regression problem.

References

[1] L. Hart, S. & Staveland. *Human mental workload*, chapter Development of NASA-TLX (Task Load Index), pages 139–183. Amsterdam: North Holland B.V., 1988.

[2] J. Healey and R. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, Volume: 6, Issue: 2:156–166, 2005.

[3] M. Honal and T. Schultz. Determine task demand from brain activity. In *Proceedings of the 3rd International Conference on Bio-inspired Systems and Signal Processing*, 2008.

[4] C. Hsu, C. Chang, C. Lin, et al. A practical guide to support vector classification. 2003.

[5] F. Kuhn. Methode zur Bewertung der Fahrerablenkung durch Fahrerinformations-Systeme. *DaimlerChrysler AG Research & Technology*, 2005.

[6] Y. Liang, M. Reyes, and J. Lee. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):340–350, 2007.

[7] F. Putze and T. Schultz. Cognitive dialog systems for dynamic environments: Progress and challenges. In *Proceedings of the 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety*, 2009.

[5] This is also supported by the evaluation of the TLX questionnaire which shows that cognitive tasks of all difficulty levels get higher scores than their visual counterparts.