A MEASURE OF STRONG DRIVER FATIGUE

David Sommer[1], Martin Golz[1,5], Thomas Schnupp[1], Jarek Krajewski[2],
Udo Trutschel[3,5] & Dave Edwards[4]
[1]Faculty of Computer Science, University of Applied Sciences Schmalkalden, Germany
[2]Work and Organizational Psychology, University of Wuppertal, Germany
[3]Circadian Technologies Inc., Stoneham, Massachusetts, USA
[4]Product Safety and Compliance, Caterpillar Inc., Peoria, Illinois, USA
[5]Institute for System Analysis and Applied Numerics, Tabarz, Germany
Email: d.sommer@fh-sm.de

Summary: Strong fatigue during sustained operations is difficult to quantify because of its complex nature and large inter-individual differences. The most evident and unambiguous sign is the occurrence of microsleep (MS) events. We aimed at detecting MS utilizing computational intelligence methods. Our analysis was based on biosignal and video recordings of 10 healthy young adults who completed 14 sessions over two nights in our real-car driving simulation lab. Visual scoring by trained raters led to 2,290 examples of MS. Only evident events accompanied by prolonged eyelid closures, roving eye movements, head noddings, major driving incidents, and drift-out-of-lane accidents were regarded as MS. All other cases with signs of fatigue were regarded as dubious. The same amount of counterexamples (Non-MS) where continued driving was still possible were picked out from the recordings. Non-MS and MS examples covered only 15% of the whole time. Support-Vector Machines were utilized as classifiers and were adapted to these two classes of examples. If such classifiers were applied consecutively, then 100% of time is covered. Validation analysis demonstrated that the classifier gained high selectivity and high specificity. Based on this complete coverage, the percentage of MS in a predefined time span can be calculated. This measure was highly correlated to deteriorations in driving performance and to subjective self-ratings of sleepiness. We conclude that reliable detection of MS is possible despite large intra- and inter-individual differences in behaviour and in biosignal characteristics. Therefore, the percentage of detected MS gives an objective measure of strong driver fatigue.

INTRODUCTION

In their review on transport safety and sleepiness Philip and Åkerstedt (2006)stated that “one major obstacle to prevention of sleepiness behind the wheel is the lack of instruments for measuring absolute levels of sleepiness in field situations. Without such instruments enforcement of alertness will be extremely difficult and interventions may have to be restricted to public information campaigns, the results of which are unclear”. Currently there are more than twenty different instruments on the market which are based on a variety of different functional principles and which support different levels of interactions between the system and the driver. On a low level of interaction the estimated fatigue is displayed to the driver in order to give him a feedback and to support his own decision making. The resolution of such instruments (alertometer) should be at least as capable of discerning two or three gradations of fatigue. For higher levels of interaction where countermeasures, like audible or visual warnings are presented to the driver, much more accurate estimations of fatigue are required. If the false alarm rate would be too high, such systems would never be acceptable to the general public. Conversely, missing errors are also not acceptable especially during high-risk conditions where strong

fatigue-related errors could lead to significant life threatening consequences. Therefore, sensitivity, specificity, as well as temporal resolution must be high especially during strong fatigue and around MS events.

MS events are clearly observable behavioral states which appear as short intrusions of sleep into wakefulness under demands of sustained attention. Many authors have reported blink duration or related measures like PERCLOS as appropriate parameters of MS events. But,a recent contribution(Schleicher et al.,2008)reported onlarge inter-individual differences in oculomotoric parameters in a data set of 82 subjects. In addition to correlation analysis these authors investigated in detail blink duration immediately before and after MS whichwere defined as overlong eye blinks. The mean duration of overlong eye blinks (MS) is not substantially longer (269 ms) than of blinks immediately before MS (204 ms) and after MS (189 ms). These blink durations seems to be much shorter than the reported 700ms of (Summala et al.,1998). (Ingre et al.,2006) also reported large inter-individual variability of blink duration in a driving simulation study of 10 subjects after working on a night shift. In conclusion, only gradual changes and large inter-individual differences appear in this important parameter which is heavily used in fatigue monitoring instruments. Similar findings are also reported of other variables, e.g. delay of lid reopening, blink interval, and standardized lid closure speed (Schleicher et al.,2008).

Adaptive biosignal processing and modern pattern recognition techniques have been shown to be effective methods for detecting MS(Golz et al., 2007 a). Such non-parametric methodology is capable of handling the large inter-individual differences found in this kind of physiological data. The main concern of this contribution is as follows. Up to now, our MS detector has been based on biosignals fromwellobservableMS events and of the same amount of well observablecounter-examples, i.e. periods of Non-MS, where the driver is still able to drive. But there are many periods where such well observablevisual scoring of driver's state is simply not possible. It is during these times where it would be of interest to know how well the MS detector is performing. This paper aims at the processing of consecutively segmented biosignals. The detector output is expected to be high during MS and MS-like states of the driver and to be low at all other states. This way, it is possible to estimate a percentage of MS which provides a new measure of strong fatigue.

EXPERIMENTS

10 healthy young adults completed 7 overnight driving sessions (1 - 8 a.m.) in our real car driving simulation lab. Sessions started at the top of every hour, had duration of 40 min, and were preceded and followed by vigilance tests and responding to sleepiness questionnaires. Reports of vigilance tests will be given elsewhere. Time since sleep was at least 16 hours, checked by wrist actometry. Subjects have been prepared beforehand by simulator training.

Several biosignals were recorded: EEG (F1, F2, C3, Cz, C4, O1, O2, A1, A2, com.av.ref.), EOG (vertical, horizontal), ECG, EMG (m. submentalis). In addition, three video recordings (driver's head & pose, driver's eyes, driving scene) were stored. Also several variables of the car, like e.g. time series of steering angle and lateral position of the vehicle, were sampled. The standard deviation of the latter is abbreviated in this paper as sdlat.

Subjectively experienced sleepiness was rated every 4 min during driving following suggestions of (Åkerstedt et al., 2006). Subject's response was given orally using the Karolinska Sleepiness Scale (KSS) (Åkerstedt, 1990).
Further experimental details have been published elsewhere (Golz et al., 2007 a, b).

ANALYSIS

A. Scoring of MSE

A first judgment of ongoing MS was done immediately during the experiments by two operators who watched the video streams. Typical signs of MS are prolonged eyelid closures, roving eye movements, head noddings, major driving incidents and drift-out-of-lane accidents. Several other signs were observed, but it has been decided not to solely rely on them. Some examples are bursts of alpha and theta activity in the EEG, spontaneous pupil contractions and stare gaze. In all, we have found 2,290 MS events (per subject: mean number 229 ± 67, range 138 - 363).

For the detection of MS events on a second by second basis a careful determination of the point in time where MS is starting will be needed. Therefore, all recorded video material and biosignals underwent off-line scoring made by independent and trained raters. They refined results of online scoring into evident MS and Non-MS (Fig. 1, red and blue dots). Later, a third visual scoring wasperformed by another trained rater. He labeled all periods (every 30 seconds) by different scores of the drivers state (Fig. 1, green dots). This is needed as a sample set for validation of consecutive classification (see below).

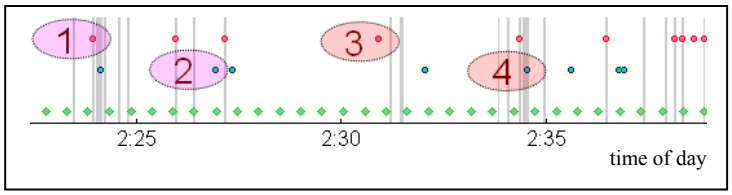

Figure 1.An example of the outcome of visual scoring during a time span of approx. 15 min. MS (red dots, upper row), Non-MS (blue dots, middle row) and time points of additional scores (green dots, lower row) are visual ratings of experts. The outcome of automatic MS classification is indicated by grey bars. Marked events are examples of conformities (1, 2) and non-conformities (3, 4) between subjective and objective detection, i. e. scoring and classification, respectively

B. Pre-Processing

Empirical investigations showed that segmentation is very sensitive to classification accuracy (Golz et al., 2007 a). Segment length should range between 4 and 12 seconds. Here we used 6 s and 0.1 s as step size of consecutive segmentation. This resulted in approximately 24,000 segments per driving session and 168,000 per night. Artifacts turned out to play a minor role when computational intelligence algorithms are applied for classification. Unpublished investigations utilizing Independent Component Analysis (ICA) to eliminate eye blink artifacts from EEG resulted in no significant improvements of MS detection compared to the case of no artifact elimination.

C. Feature Extraction

Several methods for extraction of signal features in time, spectral and wavelet domain as well as in state space were performed. It turned out that spectral power densities estimated by the modified periodogram method are most useful for MS classification (Golz et al., 2007 a). Logarithmic scaling and summation in narrow spectral bands (width 1 Hz, range 0.5 to 23 Hz) are necessary to minimize classification errors. The Delay-Vector Variance, which is a state space method, is useful as a complement, but is not highly important (Golz et al., 2007 a). Therefore, we only utilized band-averaged log power densities.

D. Classifier Setup

Support-Vector Machines incorporating radial basis functions as kernel were most optimal in terms of minimizing empirical errors of classification. But this has only been found correct if the hyperparameter and the regularization parameterwere optimized carefully. It must be emphasized that this preparation of a classifier was based solely on a set of clear examples of MS and Non-MS. They cover about 15 % of the whole time only. In contrast, the consecutive recall of the classifier covers 100 % of the whole time of driving.

RESULTS

A. Classificator performance

Conflict-free cases (true positives, true negatives), where SVM output is MS or Non-MS and rater's opinion is the same (Fig. 1: marked events 1 and 2, respectively), occurred most often. Conflicts arise when SVM output is Non-MS and rater judges MS (false negatives; marked event 3), or SVM output is MS and rater judges Non-MS (false positives; marked event 4).
This result is quantified by ROC analysis (Fig. 2, red). Lowest classification error rate (conflicts in test set) is 2.3 % in the mean. True positives and true negatives were found in 97.7 % of all events. Errors increased when data of the validation set were applied (Fig. 2, green). For a given specificity, sensitivity is lower. But note that this data set is unbalanced. That's why an optimistic bias due to prior probability has to be taken into account.

For comparison, results of other authors (Davidson et al., 2007) are presented (Fig. 2, blue). Their MS investigations were performed during a continuous tracking task and resulted in lower sensitivity, lower specificity, as well as in higher variance. This decrease in accuracy could be due to other methods of feature extraction and classification, but should mostly account to their definition of MS. These authors defined MS by large tracking errors which are due to performance decrements. But MS is only one cause for lowered performance among other psychophysiologic factors, such as lack of concentration, or aversions against the monotonous task. We believe that visual scoring of only evident examples of MS is more reliable because behavioral signs of strong central fatigue are complex and differ largely between subjects. Therefore, it is important to observe visually the temporal development of the many behavioral signs as well as to check the driving scene video for large performance decrements.

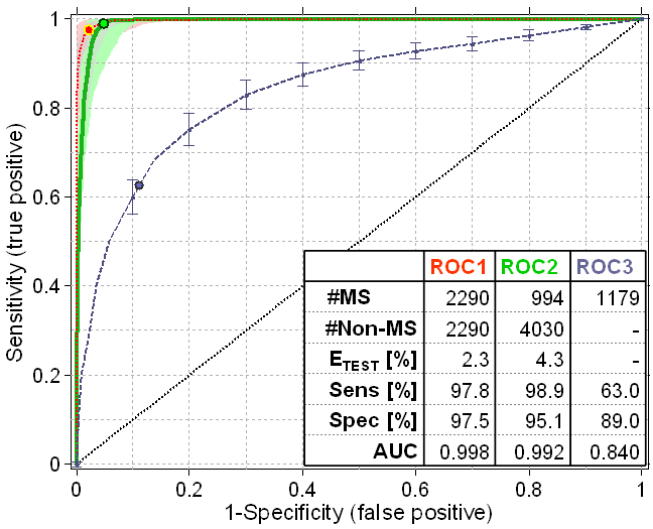

Figure 2.Receiver Operating Characteristic (ROC) of one subject. ROC1 was based on evident MS and Non-MS data, whereas ROC2 was based on visual scores given every 30 s. 95 % confidence intervals are marked (light red, light green). ROC3 represents results of (Davidson et al., 2007). Minimal test errors (E_{TEST}) are indicated (open circles). The table contains number of examples, sensitivity, specificity and area under ROC

B. Consecutive Classification

So far, we have shown that processing EEG and EOG and utilizing computational intelligence methods is successful in regard to classification of MS and Non-MS. But as mentioned above, this is true for clearly scored events. They cover only 15 % of the total length of driving.
For consecutive classification we utilized the optimized SVM classifier (Fig. 2, ROC1) in recall mode, which means that no further adaptation (training) was done. As described (Sect. III B) signals were segmented consecutively. Afterwards features were extracted and fed as input variables to the SVM classifier. This led to a binary output variable indicating MS or Non-MS at a sampling rate of 10 s^{-1}. From this variable MS percentage was calculated, which is the number of MS output samples to the total number of samples in a pre-defined interval (4 min).
All three independent variables: MS percentage, KSS, Sdlat,clearly confirm time-on-task and time-since-sleep effects(Fig. 3), because they all increase within and between driving sessions.

The only exception is that a slight decrease in MS percentage as well as in KSS arises between the 6th and 7th driving sessions. This could be caused by the time-of-day effect which is due to the habitual sleep-wake-cycle. Another effect could be a motivational one: subjects expect the end of experiments after the 7th session. Increasing MS percentages have two aspects: increasing frequency and increasing MS durations.

Mean and standard deviations of the purely subjective measure KSS is strongly correlated with MS percentage, which is a purely objective measure. Pearson's correlation coefficients are always greater than 0.95 except for the 7th session. Note that for both measures large standard deviations emerge which is mainly due to large inter-individual differences. Driving performance measured by Sdlat is also correlated with MS percentage, but not such high. Pearson's correlation coefficients range between 0.8 and 0.96.

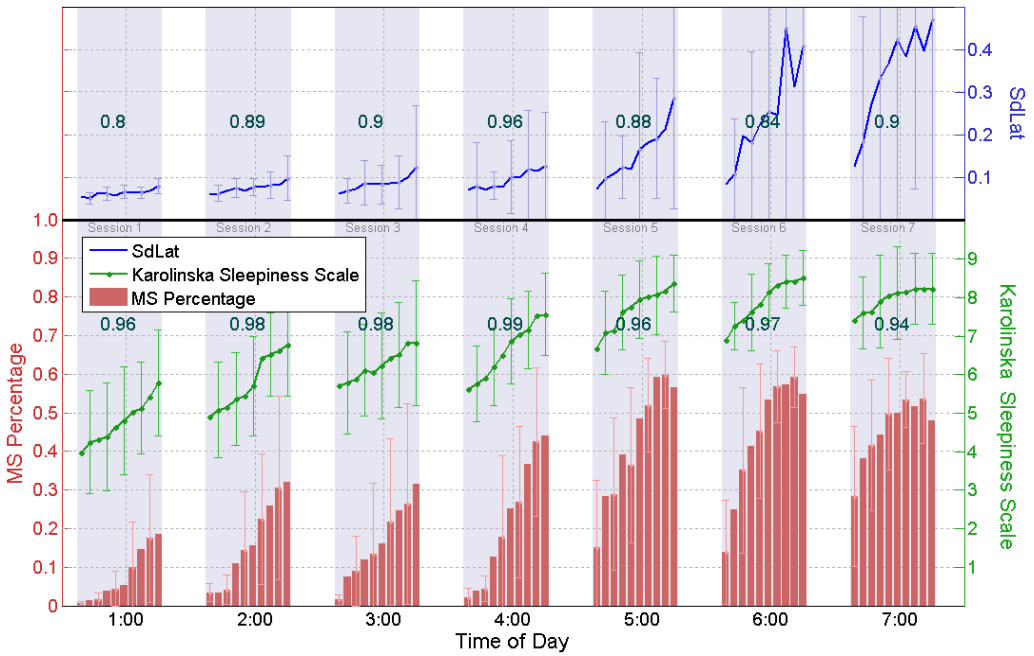

Figure 3.Mean and standard deviation of MS percentage (red), subjective sleepiness (green), and standard deviation of lateral position of the vehicle(blue). Averaging interval was 4 min. Strong correlations between both objective and subjective measures are indicated by Pearson's coefficients

CONCLUSIONS

We have shown that optimizing a classification algorithm empirically on a data set of only evident examples of MS and non-MS is successful for sensor applications where a consecutive sequence of signal segments has to be processed. The main difficulty of consecutive detection was the lack of validation because a lot of examples were to be processed where it was not clear how to label them. Many examples seemed to be Non-MS, but some behavioral signs gave reason to doubt, e.g. stare gazes or slow sliding head movements. Otherwise, not every prolonged eye-lid closure must be MS.

Two methods were proposed to validate the methodology. First, we engaged a further coworker to score visually at given points in time (every 30 sec) if MS or Non-MS appears. This way, a fully independent validation set was generated. ROC analysis (Fig. 2, ROC2) resulted in only slightly lower sensitivity and specificity compared to the test set of evident examples (ROC1). Results of a comparable study of other authors (ROC3) demonstrated that this match is not a matter of course. In conclusion, the MS detector based on EEG and EOG performs very well and is consequently validated by an independent measure (visual scoring).

Second, every consecutively detected MS event was counted and the percentage of MS was computed. This was considered as anew, objective measure of strong central fatigue. Well known effects in psychophysiology, like time-on-task and time-since-sleep, were confirmed by this measure. Moreover, a subjective measure, the self-reported sleepiness on the standardized Karolinska Sleepiness Scale, correlated always strongly to this objective measure. We have also demonstrated that MS percentage correlated to Sdlat, which is an objective driving performance measure (Åkerstedt et al., 2006).

Therefore, we conclude that consecutive and reliable detection of MS during periods of strong central fatigue is possible despite large intra- and inter-individual differences in behavior and in EEG and EOG characteristics (Ingre et al., 2006) (Golz et al., 2007 a). The question remains open if such detector application would also work during real driving. During strong fatigue and MS, such experiments would be too dangerous. Therefore, MS detection has to be verifiedunder controlled laboratory conditions. However, for applicative reasons this is valuable to develop driver monitoring technology. Their improvement and validation necessitates an independent reference standard of MS detection and strong central fatigue.

ACKNOWLEDGMENT

Part of this work was supported by German Federal Ministry of Education and Research as part of the program "Applicative Research and Development at Universities of Applied Sciences".

REFERENCES

Åkerstedt, T. (1990).Subjective and objective sleepiness in the active individual.*Int J Neurosci*, 52, 29-37.

Åkerstedt, T., Peters, B., Anund, A., Kecklund, G. (2005).Impaired alertness and performance driving home from the night shift: a driving simulator study. *J Sleep Res*, 14(1), 17-20.

Davidson, P.R., Jones, R.D. & Peiris, M.T. (2007). EEG-based behavioral microsleep detection with high temporal resolution. *IEEE Trans Biomed Engin*, 54, 832-839.

Golz, M., Sommer, D., Chen, M., Trutschel, U. & Mandic, D. (2007 a).Feature fusion for the detection of microsleep events.*J VLSI Signal Proc Syst*, 49, 329-342.

Golz, M., Sommer, D., Holzbrecher, M. & Schnupp, T. (2007 b). Detection and Prediction of Driver's Microsleep Events. In Gustafson, K. (Ed.), *Proc 14th Int Conf Road Safety on Four Continents*, (11 pages). Bangkok, Thailand.

Ingre, M., Åkerstedt, T., Peters, B., Anund, A. & Kecklund, G. (2006). Subjective Sleepiness, Simulated Driving Performance and Blink Duration: Examining Individual Differences. *J Sleep Res*, 15, 47-53.

Philip, P. & Åkerstedt, T. (2006).Transport and industrial safety, how are they affected by sleepiness and sleep restriction? *Sleep Medic Rev*, 10, 347-356.

Schleicher, R., Galley, N., Briest, S. & Galley, L. (2008).Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51 (7), 982-1010.

Summala, H., Häkkänen, H., Mikkola, T.& Sinkkonen, J. (1999). Task effects on fatigue symptoms in overnight driving, *Ergonomics*, 42(6), 798-806.