Martin Golz*, Sebastian Wollner, David Sommer and Sebastian Schnieder

EOG feature relevance determination for microsleep detection

Abstract: Automatic relevance determination (ARD) was applied to two-channel EOG recordings for microsleep event (MSE) recognition. 10 s immediately before MSE and also before counterexamples of fatigued, but attentive driving were analysed. Two type of signal features were extracted: the maximum cross correlation (MaxCC) and logarithmic power spectral densities (PSD) averaged in spectral bands of 0.5 Hz width ranging between 0 and 8 Hz. Generalised learning vector quantisation (GRLVQ) was used as ARD method to show the potential of feature reduction. This is compared to support-vector machines (SVM), in which the feature reduction plays a much smaller role. Cross validation yielded mean normalised relevancies of PSD features in the range of 1.6 - 4.9% and 1.9 - 10.4 % for horizontal and vertical EOG, respectively. MaxCC relevancies were 0.002 - 0.006% and 0.002 - 0.06 %, respectively. This shows that PSD features of vertical EOG are indispensable, whereas MaxCC can be neglected. Mean classification accuracies were estimated at 86.6±1.3% and 92.3±0.2% for GRLVQ and SVM, respectively. GRLVQ permits objective feature reduction by inclusion of all processing stages, but is not as accurate as SVM.

Keywords: Automatic relevance determination, microsleep, electrooculography, support-vector machines.

https://doi.org/10.1515/cdbme-2017-0053

1 Introduction

Some automatic relevance determination (ARD) methods have the advantage that a learning rule for relevance factors exists in parallel to the learning rule for a discriminant function. Relevance indicate how important each signal feature is for the discrimination task. Here, ARD is used to show feature reduction of two-channel EOG recordings for microsleep event (MSE) recognition. This is compared to support-vector machines (SVM), in which the feature reduction plays a much smaller role.

The potential of ARD in the context of MSE recognition should be important for many engineers developing devices of driver fatigue monitoring technologies, which are mainly contactless, camera-based and suffer somewhat from EOG in accuracy, reliability, and temporal resolution [1]. Once a prototypical, complex solution of a classification problem works, engineers are interested in cost reduction, for which they must distinguish the relevant factors from the less relevant. Particularly when assessing complex, human behavioral factors, the assessment of relevant factors is often difficult even for experts. In [2] it has been shown, for example, that oculomotor behavioral features under severe fatigue are clearly identifiable by experts (experimental psychologists), but inter-individual differences are considerable. Quantitative analysis yielded in some subjects positive and in other negative correlations of EOG features with independent fatigue measures. No single, oculomotor feature could be found that showed a uniform correlation for all subjects.

2 Material and methods

2.1 Experiments

25 young adults (11 females, 14 males, age: 23.8 ± 1.3 yrs.) drove in the driving simulator in 7 hourly sessions overnight with time-on-task of 7 x 40 min and time-since-sleep > 16 h. The study was conducted in 2016 in our laboratory. Here, first results of 7 subjects are presented. The time-consuming follow-up of the MSE findings is still ongoing for the rest of the subjects. On the experimental day, the subjects were checked with actimetry whether they were active at 8 am at the latest. They had to appear in the laboratory by 11 pm at the latest to start preparations. Among other things, their actometric recordings were checked whether they were active during the day and did not have any periods of nap sleep. The 40-minute driving sessions were arranged hourly, started at 1 am and ended at 7:40 am. During this time, the circadian

***Corresponding author: Martin Golz:** University of Applied Sciences Schmalkalden, Germany, e-mail: golz@hs-sm.de
Sebastian Wollner, David Sommer: University of Applied Sciences Schmalkalden, e-mail: {wollner, sommer}@hs-sm.de
Sebastian Schnieder: Institute of Experimental Psychophysiology GmbH, Düsseldorf, Germany, e-mail: s.schnieder@ixp-duesseldorf.de

trough was passed, so that with time-of-day a third crucial factor for high sleepiness was fulfilled. This daytime factor together with long time-on-task, long time-since-sleep and high monotony during driving simulation, the conditions were created to be able to observe a considerable number of MSE. MSE have always been defined using observable, behavioural features, especially prolonged eye closures and slow, roving eye movements. A total of 2,390 MSE was observed. Since as many MSE counterexamples (NMSE) should be included to represent a balanced data set, the sample size was ultimately at N = 4,780. NMSE are examples of sustained attention under alert and fatigued states. In addition to the vertical and horizontal EOG, also EEG and ECG as well as two videos (head, right eye region) were recorded. The recording device was a Somnoscreen EEG polygraph (Somnomedics GmbH, Germany). Only the analysis of the EOG is presented here.

2.2 Feature extraction

Assuming that the eye blinks are involuntarily controlled reflexes and are therefore relatively constant and only altered by fatigue, pattern matching is examined as the first type of feature extraction. An expert in the field looked for 10 typical examples of eye blinks in both the attentive and the MSE state. With the corresponding EOG patterns (2 s length), the cross-correlation (CC) function was estimated to analyse all blinks within 10 s immediately before MSE and before NMSE. Only the maximum CC amplitude (MaxCC) was extracted as a feature so that a total of 10 MaxCC features were extracted per event (MSE or NMSE) and per EOG channel.

Although the EOG represents a transient-rich process and apparently does not adequately satisfy the stationarity conditions in the first moment, the spectral power density (PSD) has been used as a second feature type. Segmenting both EOG signals (sampling rate: f_S = 128 Hz) was performed from 10 s to 0 s before the MSE/NMSE starting time. The two segmentation parameters, i.e. length and offset, resulted from empirical optimisation, not presented here in further detail. Logarithmic PSD were directly estimated using the modified periodogram method. Afterwards, feature reduction by averaging into spectral bands of 0.5 Hz width within the interval from 0 Hz up to 8.0 Hz. For both EOG channels together, this resulted in 32 PSD features. Together with 10 MaxCC features per EOG channel, a total of 52 features were obtained.

2.3 Automatic relevance determination

ARD aims to estimate the utility of each signal feature for the classification task. Of particular interest are methods which adapt relevance factors simultaneously when adapting a separation function between classes to discriminate. They represent a multivariate analysis and focus on the separability of classes and not on the moments of the class-conditioned probability density functions. Alternate methods that analyze the information content (entropies) or class-related moments of the univariate distribution density are disadvantaged.

Several ARD methods are available within the family of LVQ (learning vector quantization), all of which use an adaptive metric in the similarity estimation [3]. Similarity is a crucial concept of quantitative discriminant analysis. The adaptation of the metric by varying weight variables r_k is aimed at an improved separation function. Features with high weights r_k have a strong dominance in the similarity estimation and features with weight near zero affect the estimation little. The former are relevant and the latter irrelevant to classification.

One member of this family, GRLVQ, has the advantage to have clear mathematical background [4] and outperforms other ARD methods. GRLVQ is introduced in the following. Given a training set $\mathrm{S} = \{(\boldsymbol{x}_i, t_i) \subset \mathbb{R}^n \times \{1, \ldots, n_C\} | i = 1, \ldots ,\ N\}$ of pairs of feature vectors $\boldsymbol{x}_i = \left(x_{i,1}, \ldots, x_{i,n}\right)^{\mathrm{T}}$ and class labels t_i, where n_C is the number of classes. The present application has n_C = 2 (classes: MSE, NMSE) and N = 4,780. Furthermore, a set of n_P prototype vectors, which are also assigned to class labels d_i, has to be initialized randomly and data driven: $W = \{(\boldsymbol{w}_j, \tau_j) \subset \mathbb{R}^n \times \{1, \ldots, n_C\} | j = 1, \ldots, n_P\}$.

The following 4 steps are to be repeated until stop of training. ***Step* 1**: Select $(\boldsymbol{x}_i, t_i) \in S$ with random i and compute all distances between $\boldsymbol{x}_i$ and $\boldsymbol{w}_j \in W, j = 1, \ldots, n_P$ using the weighted, squared Euclidean metric:

$$\boldsymbol{d_{ij}^2} = \left\|\boldsymbol{x_i} - \boldsymbol{w_j}\right\|_{\boldsymbol{r}}^{\boldsymbol{2}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{r_k^o} \left|\boldsymbol{x_{i,k}} - \boldsymbol{w_{j,k}}\right|^{\boldsymbol{2}} \qquad (\mathbf{1})$$

with the normalised, uniformly initialised weight vector $\boldsymbol{r}^0 = (r_1^o, \ldots, r_n^o)^T$ containing the feature relevances. Find the smallest and the second smallest distance d_{ic}^2, d_{is}^2, respectively ($j = c$ for closest and $j = s$ for second-closest prototype vector). There are 4 cases for the assigned class labels: (a) $t_i = \tau_c$, $t_i \neq \tau_s$, (b) $t_i \neq \tau_c$, $t_i = \tau_s$, (c) $t_i = \tau_c$, $t_i = \tau_s$, (d) $t_i \neq \tau_c$, $t_i \neq \tau_s$. In case of (c) and (d) start again with the beginning of step 1. For (a) and (b) calculate the following two step size modulation factors:

$$\boldsymbol{\kappa_c} = \left(\frac{\boldsymbol{d_{is}}}{(\boldsymbol{d_{ic}}+\boldsymbol{d_{is}})^2}\right) \boldsymbol{sgd'}\left(\frac{\boldsymbol{d_{ic}}-\boldsymbol{d_{is}}}{\boldsymbol{d_{ic}}+\boldsymbol{d_{is}}}\right) , \boldsymbol{\kappa_s} = \left(\frac{\boldsymbol{d_{ic}}}{(\boldsymbol{d_{ic}}+\boldsymbol{d_{is}})^2}\right) \boldsymbol{sgd'}\left(\frac{\boldsymbol{d_{ic}}-\boldsymbol{d_{is}}}{\boldsymbol{d_{ic}}+\boldsymbol{d_{is}}}\right) \qquad (2)$$

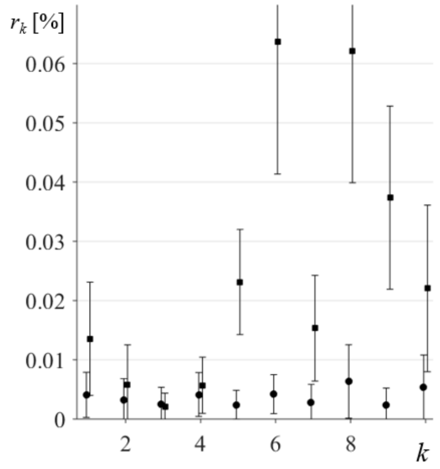

Figure 1: Mean and standard deviations of normalized MaxCC feature relevances from the horizontal (circles) and the vertical (squares) EOG channel versus target number of typical alert ($k = 1, ..., 4$) and typical drowsy patterns ($k = 5, ..., 10$).

***Step* 2**: Check if $\boldsymbol{x_i}$ lies in an hyperbolic window with width $\sigma \in [0{,}1;\ 0{,}4]$ around the perpendicular bisector of $\boldsymbol{w}_c$ and $\boldsymbol{w}_s$, otherwise go to step 1: $min\left(\frac{d_{ic}}{d_{is}},\frac{d_{is}}{d_{ic}}\right) > \left(\frac{1-\sigma}{1+\sigma}\right)^2$.

***Step* 3**: Update of the closest and second-closest prototype vectors using step size modulation factors of eq. 2:

$$\Delta\boldsymbol{w}_c = \pm\kappa_c\eta(t)(\boldsymbol{x}_i - \boldsymbol{w}_c),\ \Delta\boldsymbol{w}_s = \mp\kappa_s\eta(t)(\boldsymbol{x}_i - \boldsymbol{w}_s) \quad (3)$$

In case (a) use the upper sign, in case (b) the lower sign. The step size $\eta(t)$ controls the convergence rate and should be a slowly decreasing function of iteration index t .

***Step* 4**. Update of the relevance factors used in eq. (1)

$$r_k = r_k^0 \pm \eta_r\left(\frac{d_{ic}(x_{i,k} - w_{c,k})^2 - d_{is}(x_{i,k} - w_{s,k})^2}{(d_{ic}+d_{is})^2}\right) sgd'\left(\pm\frac{d_{ic}-d_{is}}{d_{ic}+d_{is}}\right) \quad (4)$$

In case (a) use the upper sign, in case (b) the lower sign. Negative relevances are prevented with $r_k = \max(r_k, 0)$. Finally, the relevance vector must be normalized: $\boldsymbol{r}^o = \frac{1}{\|\boldsymbol{r}\|_2}\boldsymbol{r}$. Then proceed to step 1. This loop of training steps must be terminated either when an accuracy limit is exceeded or at an a priori fixed number of iterations. The final vector $\boldsymbol{r}^o$ contains all feature relevances.

3 Results

GRLVQ performance was estimated using cross-validation, in particular multiple random sub-sampling, with random partitioning of the data into 80% training and 20% test data being repeated 50 times. This allows to estimate the classification accuracy based on the training set and the classification accuracy based on the test set. The first is an estimate of the adaptivity and the second is an estimate of generalizability.

Feature relevances are always estimated based on the training set, because their calculation is part of the training process. Relevances of MaxCC features were generally low and were in the interval from 0.002% to 0.006% for the horizontal EOG channel and in the interval from 0.002% to 0.06% for the vertical EOG channel (Figure 1). Relevances of PSD features were much higher and were in the interval from 1.6 % to 4.9 % for the horizontal EOG channel and in the interval from 1.9 % to 10.4 % for the vertical EOG channel (Figure 2). All relevances showed relatively large standard deviations.

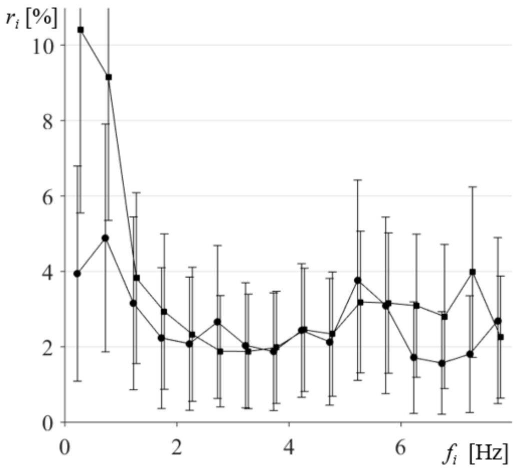

Figure 2: Mean and standard deviations of normalized PSD feature relevances from the horizontal (circles) and the vertical (squares) EOG channel versus centre frequency of the spectral bands.

Mean and standard deviations of the classification accuracy was estimated by both GRLVQ and SVM (Table 1). As expected, accuracies based on the training set was always higher than based on the test set, because the latter consists of feature vectors which were never presented during the training. Since relevances of the vertical channel were usually higher than that of the horizontal channel, it was investigated how much the classification accuracy decreases when only the vertical channel is used for the analysis. It turned out that GRLVQ results are not affected, but SVM results dropped by 1.6 % accuracy, but still outperformed GRLVQ.

Table 1: Mean and standard deviations of the classification accuracy based on both the training and the test set. Two different classification methods and two different data sets were compared (V = vertical EOG channel, H = horizontal EOG channel).

Method	EOG channels	a_{Train} [%]	a_{Test} [%]
GRLVQ	V + H	87.5 ± 0.7 %	86.6 ± 1.3 %
SVM	V + H	99.3 ± 1.6 %	92.3 ± 0.2 %
GRLVQ	V	87.4 ± 0.7 %	86.6 ± 1.4 %
SVM	V	96.5 ± 0.7 %	90.7 ± 0.1 %

4 Discussion and conclusion

It has been shown that the two states "fatigued, but still attentive" and "MSE, high accident risk" can be classified from the EOG with a mean average accuracy of 92.3%. Only two channels have to be recorded. If only one channel was used, i.e. the vertical EOG, then mean accuracy decreased by only 1.6 %. But for a number of reasons the results have a limited meaning validity. Firstly, driving conditions were ideal and had a reduced task complexity, compared to real driving. Secondly, the number of subjects was very low which led to a reduced complexity during data analysis. Thirdly, only a limited number of short segments were analysed; 13.28 h in sum. Total time of driving of the 7 subjects was 32.67 h, so that only 40,65 % of recordings was analysed. This fact cannot be easily changed to avoid biased results due to unbalanced data.

The ARD estimations carried out with GRLVQ showed that the relevances of the two feature types as well as of the two EOG channels clearly differed. Relevances of time domain features, i.e. amplitude of the main peak of the cross-correlation function, were smaller by three orders of magnitude than relevances of spectral domain features, i.e. band-averaged, logarithmic power spectral densities. Highest relevances were between 9 % and 11 % and affected the PSD at lowest frequencies in the vertical EOG. These spectral bands are dominated by the large and slow eyelid movements and to some extend also by transients due to saccadic eye movements. These two bands also reflect the changes caused by fatigue in velocity of lid opening and lid closure, which was also found by Schleicher and Galley [2]. New insights are the increased relevances from 5 Hz in both EOG channels, which, however, are weak and require a tighter control with more extensive data.

It is also noteworthy that the relevances of the horizontal EOG are not insignificant. However, this channel does not play any role in the GRLVQ results. The mean test accuracies are both the same at 86.6 %. This suggests a weakness of the methodology. A number of features may have a significant relevance and yet do not have an effect on the final outcome.

Nevertheless, the importance of ARD methods should be pointed out. Machine learning methods are able to perform model-free, multivariate analysis and to generate separation functions with high structural complexity. This complexity, however, makes interpretability difficult. Relevance determination is, at the very least, a simple attempt at automatic knowledge extraction by addressing the question of more relevant and less relevant features without elucidating the structural complexity.

GRLVQ displays less relevant features and thus permits an objective feature reduction. This is done in an empirical way by inclusion of all processing stages. Nevertheless, GRLVQ is not as accurate as SVM.

Acknowledgment: The authors would like to thank our lab engineer, Mr. Adolf Schenka, for technical support during the lab study and for data management.

Author's Statement

Research funding: The authors state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Golz M, Sommer D, et al. Evaluation of fatigue monitoring technologies. Somnologie 2010;14(3):187-199.

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

[3] Biehl M, Hammer B, Villmann T. Prototype-based models in machine learning. WIREs Cogn Sci 2016;7:92-111.

[4] Hammer B, Villmann T. Generalized relevance learning vector quantization. Neuronal Networks 2002;15:1059-1068.