24 Improving Human Performance in a Real Operating Environment through Real-Time Mental Workload Detection

Jens Kohlmorgen, Guido Dornhege, and Mikio L. Braun
Fraunhofer–Institute FIRST
Intelligent Data Analysis Group (IDA)
Kekuléstr. 7, 12489 Berlin, Germany

Benjamin BLankertz and Klaus-Robert Müller
Fraunhofer–Institute FIRST
Intelligent Data Analysis Group (IDA)
Kekuléstr. 7, 12489 Berlin, Germany

Technical University Berlin
Str. des 17. Juni 135
10 623 Berlin, Germany

Gabriel Curio
Department of Neurology, Neurophysics Group
Campus Benjamin Franklin, Charité University Medicine Berlin
Hindenburgdamm 30, 12200 Berlin, Germany

Konrad Hagemann, Andreas Bruns, Michael Schrauf, and Wilhelm E. Kincses
DaimlerChrysler AG
Group Research, HPC 50-G024
71059 Sindelfingen, Germany

24.1 Abstract

The ability to directly detect mental over- and under-load in human operators is an essential feature of complex monitoring and control processes. Such processes can be found, for example, in industrial production lines, in aviation, as well as in common everyday tasks such as driving. In this chapter, we present an EEG-based system that is able to detect high mental workload in drivers operating under real traffic conditions. This information is used immediately to mitigate the workload typically induced by the influx of information that is generated by the car's electronic systems. Two experimental paradigms were tested: an auditory workload scheme and a mental calculation task. The result is twofold. The

system's performance is strongly subject-dependent; however, the results are good to excellent for the majority of subjects. We show that in these cases an induced mitigation of a reaction time experiment leads to an increase of the driver's overall task performance.

24.2 Introduction

The detection of mental workload is considered an important issue in fields where operational alertness and elevated concentration is crucial, as it is, for example, for pilots, flight controllers, or operators of industrial plants. The output of such a workload detector could be integrated with existing systems to control the information flow to the operator in order to maximize the performance. One approach consists of creating a closed-loop system in which the system's interaction with the operator is adjusted according to the operator's mental workload measured by the workload detector. Another possibility consists of using the workload detector as an objective measure of mental workload to develop improved modes and organizations of human-machine interaction.

In this chapter, we follow the first approach and use a workload detector to reduce the imposed workload, thereby improving the operator's overall performance. We study the problem of workload detection and performance improvement in the context of driving a car while performing additional tasks that model interaction with the car's systems. The motivation for the present work was to obtain a system that is able to measure *and* mitigate mental workload (1) in real time and (2) in a real operational environment, ultimately to detect, or even to avoid, stressful and cognitively demanding situations for human operators in critical monitoring or control tasks.

Approaches to mental workload detection are largely based on the electroencephalogram (EEG) and have so far been investigated mainly under controlled laboratory conditions, for example, by using tasks that involve the subject's short-term memory (Gevins et al. (1997, 1998); Low et al. (1999); Schack et al. (2002); Stipacek et al. (2003); Howard et al. (2003)), by mimicking in-flight tasks of a pilot (Pope et al. (1995); Prinzel et al. (2000); Smith et al. (2001)), or by simulating air traffic control (Brookings et al. (1996)). Attempts to measure mental workload in real operational environments have so far been limited to an offline analysis after the recording (Sterman and Mann (1995); Hankins and Wilson (1998)), lacking the possibility of online feedback to actually control the workload as discussed above (see Scerbo et al. (2003) for a more comprehensive review of the field).

The utility of these studies for our current application is rather limited. While the studies have identified some neurophysiological effects of mental workload, the results do not provide clear evidence due to the heterogeneity of the studied tasks. In the works cited above, the workload is induced either visually or by memory tasks, and it is unclear if these observations carry over to the setting of car driving, a task that is rather visually demanding by itself.

Also, most of the results were obtained using laboratory experiments conducted under relatively controlled conditions, and it is unclear how the observations of these experiments translate to the more complex real-world setting. It is important to note that the analysis of EEG data under real operating conditions is significantly more challenging than under

controlled laboratory conditions. Besides uncontrollable sources of distraction and consequently a larger degree of uncertainty about the subject's true mental state, the EEG signals can be heavily contaminated by artifacts, primarily due to facial muscle activity.

Finally, previous work in the field of single-trial EEG analysis has shown large intra- as well as interindividual differences. Consequently, it does not seem realistic at this point to build a universally applicable detector with fixed parameters. It is our belief that any realistic workload detector currently must have some means of adaption to the individual under consideration.

Based on these considerations as well as on the results reported in the literature, we follow a flexible approach that takes into account the observed neurophysiological effects while at the same time addressing the uncertainty and variability of the experimental and physiological conditions. This is realized by designing a highly parameterized workload detector that can detect the reported neurophysiological effects, but is not restricted to a particular feature. The high dimensionality of the parameter set and the noisy nature of the EEG signals then pose the challenge of robustly estimating the parameters. This task is addressed by using methods from machine learning.

24.3 The Experimental Setup

The goal of the current study was to develop a system that is able to measure and mitigate mental workload in real time and in real operational environments. Operating a vehicle under real conditions, including the execution of secondary tasks not related to driving such as interacting with other vehicle occupants or with the electronic equipment of the vehicle, represents a complex operational task. We exemplarily used this task to develop our approach and prove its success.

Twelve male and five female subjects age 20 to 32 years old performed the experiment. The subjects were instructed to drive at approximately 100 km/h on the highway in moderate traffic conditions. Note, however, that the traffic intensity was not controllable. The experiments took place on the public German highway B10 (between Esslingen am Neckar and Wendlingen) during the usual daytime traffic (figure 24.1). The subjects were instructed not to speak during the experiment in order to avoid additional workload as well as a systematic activation due to muscle artifacts.

The subjects were instructed to perform three types of tasks: a *primary task* (driving the vehicle), a *secondary task*, and a *tertiary task*.

The secondary task was an auditory reaction time task mimicking the interaction with the vehicle's electronic warning and information system. It was important to choose a simple task that would most likely *not* impose any significant amount of additional cognitive workload on the driver. The task was used to measure the driver's performance in terms of reaction time: voice recordings of the German words *links* (left) and *rechts* (right) were randomly presented every 7.5 s via the car's audio system and had to be acknowledged as quickly as possible by pressing corresponding buttons mounted on both index fingers.

The tertiary task was designed to induce high mental workload. We studied two different types of workload. The first type was a *mental calculation task* (mimicking "thinking

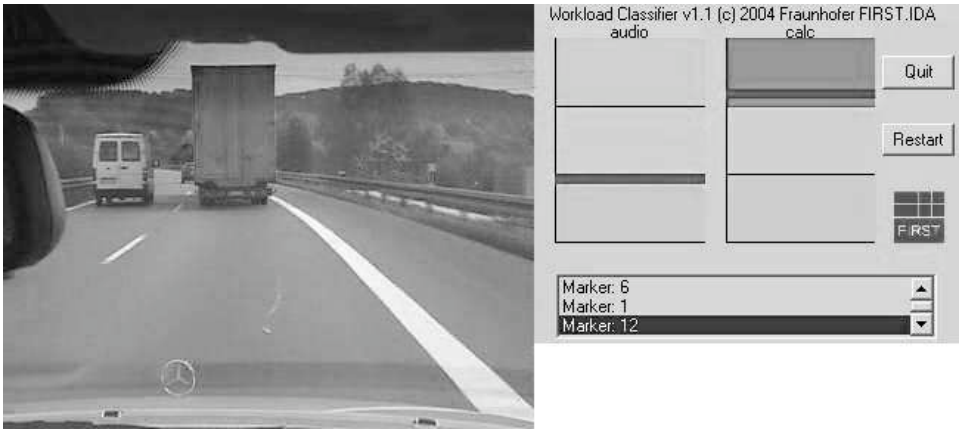

Figure 24.1 The mental workload detector in operation, during a mental calculation task performed by the driver (a scene from one of the experiments). Two gauges (right) separately indicate auditory and mental calculation workload. Each indicator bar can move over a green (lower), yellow (middle), and red (top) background, indicating the amount of detected workload. In the snapshot, the bars correctly indicate low auditory and high calculation workload.

processes") that stressed the driver's working memory. In this condition, the drivers are asked to silently count down in steps of twenty-seven, starting from an initially given three-digit random number (between 800 and 999). After two minutes, the subjects were stopped by the beep of a timer and verbally asked for the final result. The second type of workload-inducing task was an *auditory task* in which the drivers had to direct their attention to one of two simultaneously presented voice recordings, replicating a situation in which several vehicle occupants are talking at the same time: A female news reader and a male voice reciting from a book. The subjects were instructed to follow the latter. To verify whether the subjects were engaged or not, they had to answer related questions. To avoid artifact contamination of the EEG, the questions were presented during the turning points of the course, where EEG was not analyzed.

The entire experiment is organized in a block structure with a block length of two minutes each (see figure 24.2). A *high-workload block* of two minutes' length, comprising all three tasks, was alternated with a *low-workload block*, in which the subjects performed the primary task (driving) and the secondary task (reaction time task), but not the tertiary task. Experience from pilot testings shows that it is possible to perform the tertiary tasks for two minutes at the same attention level without getting tired.

One full pass of the experiment consisted of three pairs of high and low blocks in a row, with different initial three-digit numbers and with different parts of the story. Each pass is performed two times by each subject to get sufficiently many changes in workload level for the subsequent performance analysis of the detector.

A crucial purpose of the experiment is to investigate whether the output of the workload detector can be used to control the secondary task such that the performance of the subject is improved. This is accomplished by making the secondary task a controllable task interrupted by the workload detector each time the system identifies a high workload

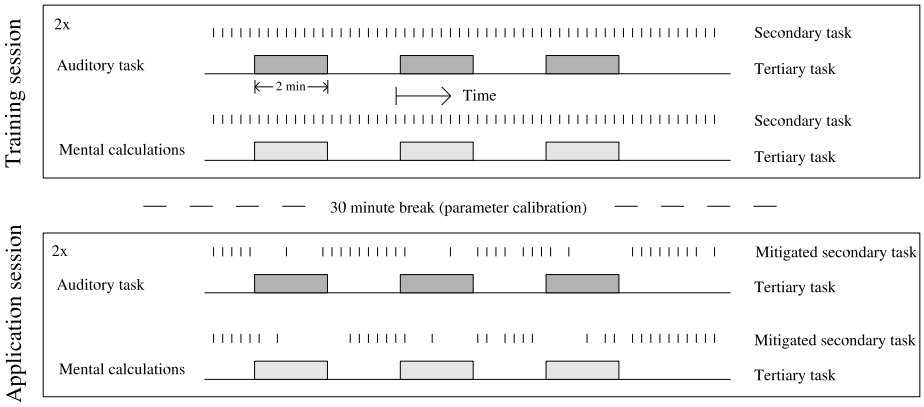

Figure 24.2 Illustration of the experimental procedure. The experiment consists of two consecutive sessions, *training* and *application*, joined by a short break in which the parameters of the workload detector are computed from the training session data. Each session consists of two runs for each tertiary task, and a run consists of three high workload blocks of two minutes length, followed by a low workload block of the same length. In the training session, the secondary task (reaction time experiment) is performed throughout the session, whereas in the application session it is controlled by the workload detector. If a high workload condition is detected, the secondary task is suppressed in order to improve the performance for this task as measured by the average reaction time.

condition. This serves to mitigate the workload imposed on the driver. The performance is measured by the average reaction time over the course of the experiment.

This experiment provides two measures for our method: the accuracy of the prediction of a high workload condition, and the performance increase as measured by the reaction times.

As stated, our workload detection method is highly parameterized to be able to adapt to the environmental conditions, the task, and the driver. To estimate these parameters, one experiment consists of two sessions: a training and an application session. The training session is performed without running the detector. Immediately after the training session, the recorded EEG data is used to train the detector, that is, its parameters are computed from the data. In the subsequent application session, the trained workload detector is applied to continuously analyze the ongoing EEG measurement in real time and to output a high or low workload indication. In case of a high workload indication, the secondary task automatically gets suppressed without external intervention until the detector indicates low workload again (mitigation strategy). Both sessions are performed on the same day with an intermediate break of roughly thirty minutes in which the detector is trained.

24.4 Online Detection of Mental Workload

In this section, we describe the workload detector and the procedure for parameter calibration. To test whether it is possible to distinguish types of workload, we use two independent

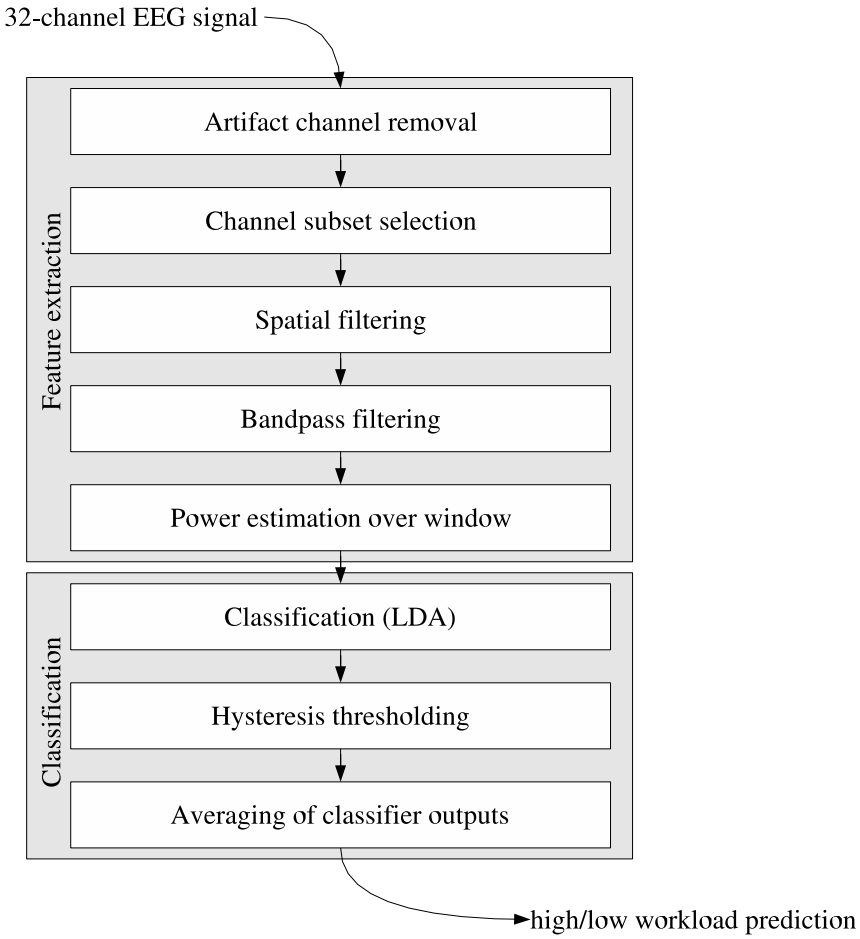

Figure 24.3 The workload detector predicts two different workload conditions in real time from an ongoing 32-channel EEG measurement. The detector consists of two stages: Feature extraction and classification.

workload detectors per subject, one for each paradigm. As mentioned, we observed a large inter- and intrasubject variability of the EEG in precursory investigations. We therefore adopted a rather general approach and designed a workload detector with subject- and task-specific parameters.

24.4.1 The Workload Detector

Each detector consists of two parts: feature extraction and classification. The feature extraction component extracts neurophysiologically interesting features, which are then used by the classifier to predict the workload (see figure 24.3).

The feature extraction consists of the following four steps: (1) removal of artifact contaminated EEG channels, (2) selection of a subset of the remaining channels, (3) spatial filtering, and (4) computing the power in a selected frequency band. The possible

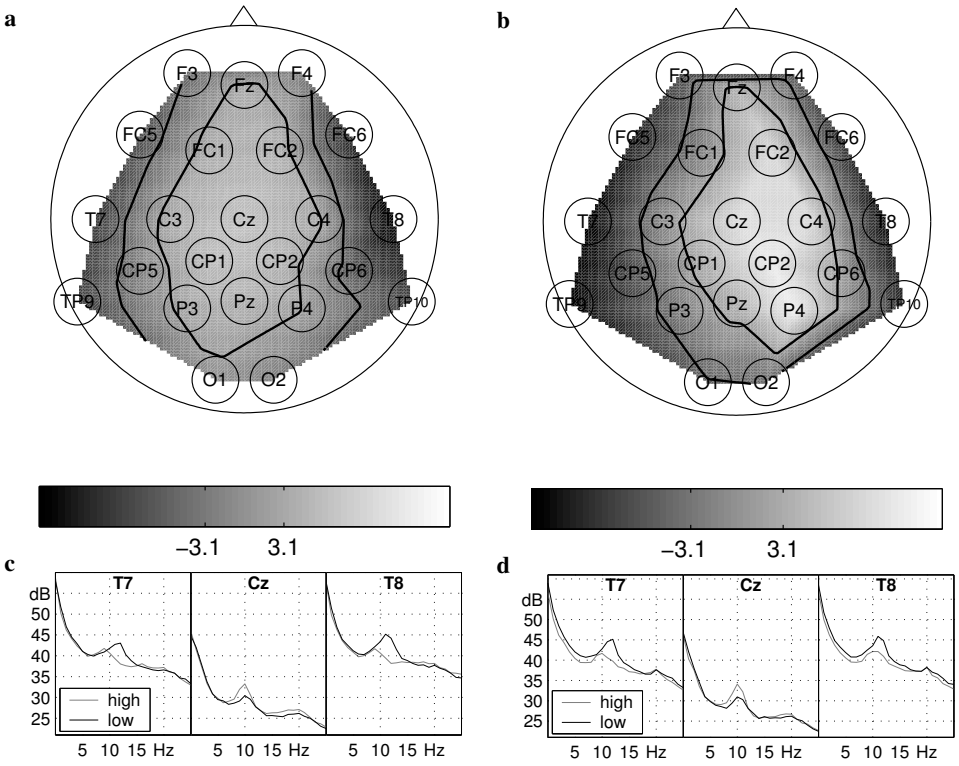

Figure 24.4 Spectral differences in EEG between low and high workload condition (on the average). Left: for auditory workload. Right: for mental calculation workload. (a), (b): t-statistics of 8–12 Hz and power, interpolated between the electrode positions of subject *ps*. The contour lines denote $P = 0.001$, i.e., bandpower differences in central and temporal areas are significant ($P < 0.001$). The central locations exhibit *more* bandpower under the high workload condition, the temporal locations *less*. (c), (d): power spectra of three discriminative EEG channels of subject *ps*. Clear differences are found at the 10 Hz α peaks.

parameters for each of these steps (and also for the classification stage) are listed below. The sets of candidate parameters were designed on the basis of neurophysiological findings that have been reported in the literature. A specific parameterization is chosen after the initial training session and kept constant for the entire application.

EEG channels:[1] (F# denotes the whole F-row, see figure 24.4a and b)
{FC#, C#, P#, CP#};
{F#, FC#, C#, P#, CP#, O#};
{F#, FC#, C#, P#, CP#, O#, T7, T8};
{FC#, C#, P#, CP#, T7, T8}.

Spatial filter: common median reference or none.

Frequency band: 3–15, 7–15, 10–15, 3–10 Hz.

Window lengths and integrate values: 10 s and 10; 15 s and 5; 30 s and 1.

Classifier parameters: real number weight for each remaining channel.

Hysteresis thresholds: two real numbers m_l and m_h.

Starting with 32 recorded EEG channels, the first step of data processing is the exclusion of channels that are contaminated with artifacts during the training session. More precisely, channels containing muscle or eye-movement artifacts that are correlated with a particular workload condition are identified based on their frequency spectra and excluded. This is a crucial step and prevents the classifier from being driven by artifacts rather than neurophysiological effects. Therefore, frequencies above 20 Hz and below 6 Hz are scanned for significant broadband differences between the two workload conditions. Such broad-band differences are characteristic for muscle artifacts ($>$ 20 Hz) or eye artifacts ($<$ 6 Hz). The channels that exhibit those differences are excluded.

Next, a subset of the remaining EEG channels is selected for further processing. This subset is one of four candidate sets that potentially include frontal, occipital, and temporal scalp positions. By using these sets, a rough preselection of EEG channels is achieved. Each of the selected channels is then optionally normalized by the common median reference signal (the median of all channels is subtracted from each channel), which is a variation of the commonly employed common average reference filter. We choose the median because it is more robust than the mean with respect to measurement outliers, which we expect to occur more often in the given real-world setting.

The signal is then processed through a subject- and task-specific bandpass filter using one of the bands listed above. The actual input to the workload classifier is the power of each bandpass-filtered channel in a time window of specific length (within 10–30 s), sampled every 200 ms. The use of time windows shorter than 10 s typically leads to a clear degradation of the classifier performance, which reflects the difficulty of distinguishing between the high and low workload class. Indeed, for the shorter window lengths we use an average of the classifier output for a predefined number of successive predictions to get a more robust result (i.e., 10 successive predictions for 10 s. window length, 5 for 15 s). Interestingly, in 82 percent of the cases, a 10 s window was finally chosen by our method. From this feature extraction stage, every 200 ms, we obtain a feature vector which is then fed into the classifier.

For classification, we use a linear model whose parameters are computed by standard linear discriminant analysis (LDA) of the feature vectors obtained from the high and low workload conditions of the training session (Fisher (1936)). Nonlinear methods, such as regularized kernel ridge regression (Poggio and Girosi (1990)) or support vector machines (Vapnik (1998)), produced comparable results but no improvements in offline analyses.

The output of the classifier is a scalar value representing the estimated degree of low workload (values below zero) or high workload (values above zero). We then map this real-valued output to a binary quantity that indicates the two states, high and low workload, by means of a threshold scheme that employs a hysteresis, which makes the classification substantially more robust. It consists of two thresholds, $m_l < m_h$, such that switching to a *high workload* indication takes place once the output exceeds m_h, while switching to a *low workload* requires that the output falls below the lower value m_l. The values m_l and m_h are subject- and task-specific, and therefore are calibrated on the training data also.

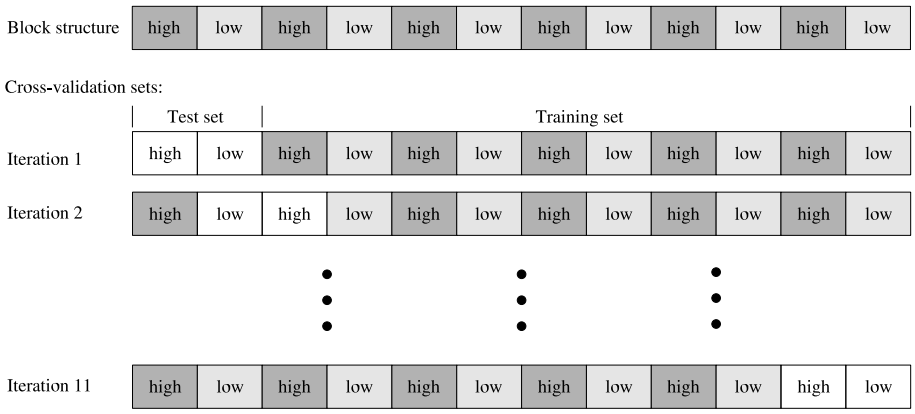

Figure 24.5 Recorded EEG signals typically are highly correlated locally. Therefore, the cross-validation scheme has to be set up properly to obtain a realistic estimate of the true generalization error. This can be accomplished by splitting the datasets based on the block structure of the experiment. For each of the eleven iterations, the workload detector is trained on ten blocks (dark and light grey). The detector is then applied to the two remaining blocks (white) to obtain an estimate of the performance on new, unseen data. The eleven individual performance estimates finally are averaged to obtain a robust estimate of the generalization error.

24.4.2 Parameter Calibration

The set of possible parameters, as specified in the last section, results in a controlled flexibility of the workload detector. The detector can adjust itself to the most discriminative features, individually for each subject and each task (and it thereby accounts for the known inter- and intrasubject variability). On the other hand, this adaptation is limited to a scope that is neurophysiologically reasonable.

This flexible approach poses the problem of robustly identifying the most suitable parameter set in each experiment. Therefore, to find suitable values for all the previously mentioned subject- and task-specific parameters, we use the well known cross-validation technique (Cover (1969)), taking into account the particular block structure of our experiment. The rationale of this technique is to find parameters that *generalize well*, that is, lead to good performance on new, unseen data, given just a fixed training dataset. For the training data, the class labels are known, in this case *high* or *low workload*, whereas for new data they must be inferred by the model (i.e., the workload detector). To avoid *overfitting* the training data, resulting in inferior performance on new data, new data is simulated in this approach by splitting the training dataset into two sets: One is used to fit the model to the data, and the other one, the *validation set*, is used to assess the quality of the model.

It is important to note that for time series data like EEG signals, some care has to be taken to perform the split such that the estimated generalization error is realistic. Since data points that are close in time are likely to be highly correlated, the split cannot be performed by selecting a random subset. This would result in many almost identical data points in both sets, such that the training and validation sets would be very similar and thus useless for testing the generalization performance. Instead, the validation set should

be a single block of consecutive data points (see figure 24.5). For the same reason, if there is a block structure of the class labels, the two split points should be at the class label boundaries. Finally, to make the estimate of the generalization error more robust, it is useful to perform the split in two subsets several times in different ways and then average over all individual generalization errors. We therefore perform the splits by leaving out two consecutive high- and low-workload blocks for validation and repeat the estimation of the generalization error for all subsequent pairs of blocks, which thus results in an *elevenfold cross-validation*. This procedure is performed for each possible combination of parameter candidates in the feature extraction part (EEG channel subset, spatial filter, frequency band, window length): the corresponding features are extracted and a classifier is trained on the extracted features by using LDA.

For each classifier obtained in this way, the hysteresis thresholds are then determined using the workload predictions of the classifier for the data in the training set. Recall that these are real numbers below or above zero, representing the estimated degree of low or high workload, respectively. The idea behind using a threshold scheme is to identify an uncertainty interval by a lower and upper threshold, m_l and m_h, in which outputs are generated almost equally likely from data of both classes. In this region of uncertain predictions, the system should stick to its previous class decision, exploiting the fact that changes in workload are slow in comparison to the frequency at which predictions are made.

The thresholds lie in the interval spanned by the smallest and largest classifier output. In this range, there exists a decision threshold m_0 that attains maximum classification accuracy on the training set (without hysteresis). A candidate pair for m_l and m_h is then given as the smallest and largest threshold such that the classification accuracy is still larger than $\eta\, m_0$, with η being a value between 0.9 and 1.0. Such candidate pairs are generated for a number of η-values. The pair that maximizes the training set accuracy resulting from classification *with* hysteresis is ultimately selected.

The workload detector fully specified in this way then predicts the workload on the two left-out blocks, resulting in a cross-validation error. Each complete set of parameters is evaluated in this fashion, and the winning configuration is finally chosen among the candidates with the smallest cross-validation error.

24.5 Results

We discuss the results of our experiment with respect to three different criteria: neurophysiological interpretation of the results, the accuracy of the workload detector, and the performance increase obtained by using the workload detector to mitigate the workload in high workload conditions.

24.5.1 Neurophysiological Interpretation

A comparison of the relevant discriminating quantities, the channel-wise power spectra, reveals a strong intersubject variability. Only some subjects exhibit clear α peaks (8–12 Hz). There also is no clear unique neurophysiological effect that can be observed.

The best performing subject *ps* not only has very pronounced α peaks, but also displays clear differences in the amount of α power for the two workload conditions (at least in the overall average), which explains the good performance (figure 24.4). Remarkably, there is an *increase* in α power under the high workload condition (at and around Cz in figure 24.4), which also can be observed in eleven other subjects (mainly parietal). This is somewhat in contrast to the work of others, where α generally decreases (Scerbo et al. (2003))—an effect that we find in only about half of the subjects (and also in figure 24.4, e.g., at T7 and T8). A reason for this difference could be the complexity and real-world nature of our experiment, but most likely it is because our workload-inducing tasks are not visual, as opposed to a large number of experiments reported in the literature.

In summary, the large intersubject variability justifies our highly adaptable approach, which automatically adjusts the workload detector to these neurophysiological variations.

24.5.2 Accuracy of the Workload Detector

The quality of the workload detector is assessed by comparing the indicated workload with the high/low block structure of the experiment. The presented results reflect the performance of the subject- and task-specific workload detectors after training, that is, in the (real-time) application session.

As an example, the exact time course of the workload detector output for the best performing subject, *ps*, is depicted in figure 24.6. For this subject, the *continuous* classifier output (lower line) already exhibits a remarkable correlation with the block structure of the experiment. For the other subjects, this correlation was less prominent and there the hysteresis mechanism, which finally yields the binary high/low workload indication, significantly improved the classification performance.

The results for all subjects are shown in figure 24.7 as the percentage of correctly classified time points. The first few seconds of each task block that amount to the window length (i.e., typically 10 s) were excluded from the assessment, since this is the potential response time of the system. One can see that the intersubject variability is very large, but nevertheless a classification accuracy of more than 70 percent for eight out of seventeen subjects for the auditory task and for eleven out of seventeen subjects for the mental calculation task was achieved. The best performing subject, *ps*, achieved classification accuracies greater than 90 percent for both the auditory and the classification task.

24.5.3 Performance Improvement

The binary detector output was used to mitigate the workload of the subject by suppressing the auditory reaction task when the workload indication is *high*. By comparing the (unmitigated) training session with the (mitigated) application session, we see that the mitigation

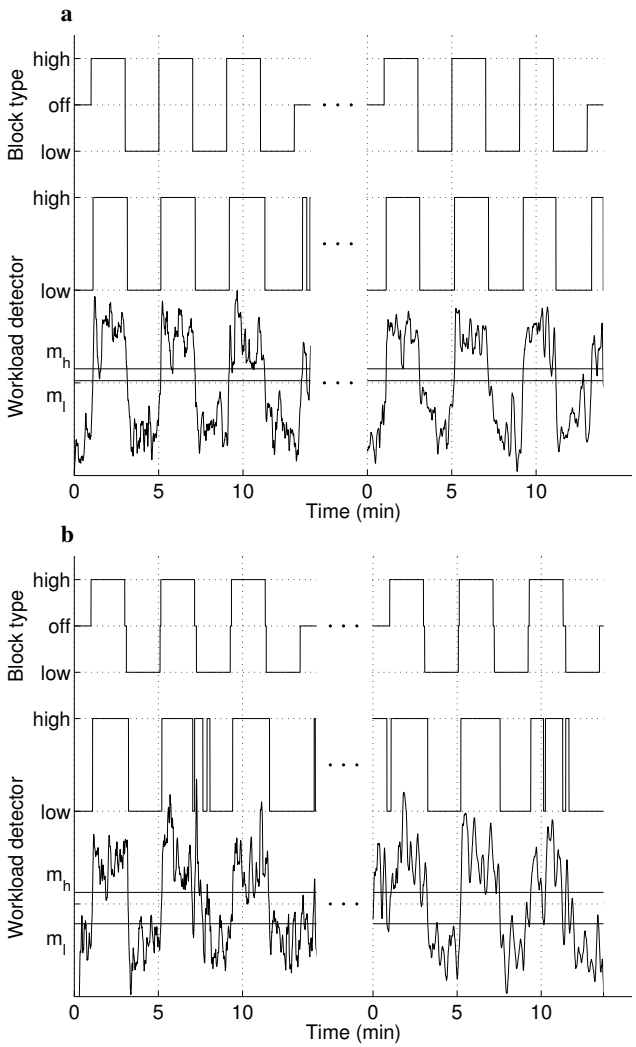

Figure 24.6 The exact time course of the classifier output for the best performing subject, *ps*, and the corresponding binary high/low workload indication used to control the mitigation, in comparison with the true high and low workload conditions. (A) For auditory workload (95.6% correct); (B) For mental calculation workload (91.8% correct).

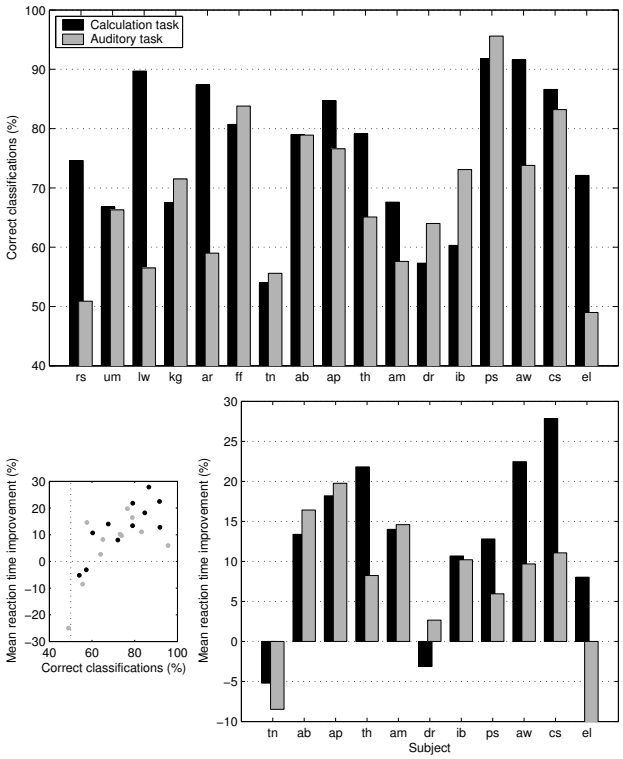

Figure 24.7 Experimental results. Top: Percentage of correctly classified workload for each subject (in chronological order). In some cases the classifier could not find sufficiently discriminating features in the EEG since even a random classification would already yield a rate of 50%. Bottom right: The average improvement of the reaction time of each subject due to the workload mitigation strategy (10% typically corresponds to an improvement of about 100 ms). Mitigation was introduced after the first six subjects and consists in the temporary suppression of the reaction task. Degradation of average reaction times happens only in cases where the classifier does not perform well. Bottom left: The correlation between classification performance and mean reaction time improvement.

strategy leads to significantly better reaction times on average (figure 24.7, bottom right). This is due mainly to the circumstance that reaction times are typically longer in the high workload phase. That clearly degrades the average performance in the training session, but it almost does not affect the performance in the application session, where the subject is largely exempted from reacting during the high workload condition because of the successfully activated suppression of the reaction task. Thus, the mitigation strategy effectively improves the performance of the subject by circumventing the periods of potentially long reaction times.

24.6 Discussion

In conclusion, we showed that mental workload detection in real-time and in real operating environments is possible and can lead to an improved performance of a subject by mitigating the workload in high workload conditions. In the context of driving, the mitigation of high mental workload can be of vital importance since a reaction time improvement of 100 ms, as achieved in the experiments, reduces the braking distance by 2.8 meters when driving at 100 km/h. This could be enough to prevent a collision.

We also have seen that the strong intersubject variability of the EEG makes the use of a highly adaptable system necessary. The performance of the workload detector is nevertheless strongly subject-dependent and depends on the existence of differences in the EEG power spectra for different workload conditions. These differences can be localized in different frequency bands and channels. Therefore, it seems unlikely that one can obtain good results by using a fixed neurophysiological feature. Instead, a system is required that can select from a number of neurophysiologically sensible features in a robust fashion. The presented feedback system is based on a *parameterized* EEG analysis, in which the parameters are adapted to the subject and the task in an initial training session. For future research, one major challenge is to reduce the amount of data necessary for the adaptation of the workload detector.

Acknowledgments

We gratefully acknowledge S. Willmann and S. Rothe for helpful support. The authors Konrad Hagemann, Andreas Bruns, Michael Schrauf, and Wilhelm E. Kincses were partly supported by DARPA grant NBCH 3030001.

Notes

* The first three authors contributed equally to this work.
E-mail for correspondence: jek@first.fhg.de

(1) The EEG signals are recorded from Ag/AgCl electrodes at positions according to the international 10-20 system. The actual sampling frequency of 500 Hz was downsampled to 100 Hz.