Pers Ubiquit Comput (2013) 17:229–239
DOI 10.1007/s00779-011-0466-1

ORIGINAL ARTICLE

Monitoring of mental workload levels during an everyday life office-work scenario

Burcu Cinaz · Bert Arnrich · Roberto La Marca · Gerhard Tröster

Received: 22 January 2011 / Accepted: 20 June 2011 / Published online: 4 October 2011

Abstract Personal and ubiquitous healthcare applications offer new opportunities to prevent long-term health damage due to increased mental workload by continuously monitoring physiological signs related to prolonged high workload and providing just-in-time feedback. In order to achieve a quantification of mental load, different load levels that occur during a workday have to be discriminated. In this work, we present how mental workload levels in everyday life scenarios can be discriminated with data from a mobile ECG logger by incorporating individual calibration measures. We present an experiment design to induce three different levels of mental workload in calibration sessions and to monitor mental workload levels in everyday life scenarios of seven healthy male subjects. Besides the recording of ECG data, we collect subjective ratings of the perceived workload with the NASA Task Load Index (TLX), whereas objective measures are assessed by collecting salivary cortisol. According to the subjective ratings, we show that all participants perceived the induced load levels as intended from the experiment design. The heart rate variability (HRV) features under investigation can be classified into two distinct groups. Features in the first group, representing markers associated with parasympathetic nervous system activity, show a decrease in their values with increased workload. Features in the second group, representing markers associated with sympathetic nervous system activity or predominance, show an increase in their values with increased workload. We employ multiple regression analysis to model the relationship between relevant HRV features and the subjective ratings of NASA-TLX in order to predict the mental workload levels during office-work. The resulting predictions were correct for six out of the seven subjects. In addition, we compare the performance of three classification methods to identify the mental workload level during office-work. The best results were obtained with linear discriminant analysis (LDA) that yielded a correct classification for six out of the seven subjects. The k-nearest neighbor algorithm (k-NN) and the support vector machine (SVM) resulted in a correct classification of the mental workload level during office-work for five out of the seven subjects.

Keywords Personal and ubiquitous healthcare · Mental workload · Office-work · Heart rate variability · Stress

B. Cinaz (✉) · B. Arnrich · R. La Marca · G. Tröster
ETH Zurich, Electronics Laboratory, Gloriastrasse 35, 8092 Zurich, Switzerland
e-mail: Burcu.Cinaz@ife.ee.ethz.ch

1 Introduction and motivation

Recently, the European Foundation for the Improvement of Living and Working Conditions called the attention on work-related stress that was associated with an increasing number of mental disorders [8]. Work-related stress occurs when there is a mismatch between job load and the capabilities of the worker [23]. Since in the developed countries, the workplace has changed due to globalization, use of new information, and communication technology, mental workload is the dominant element in most jobs. If high level of mental workload cumulates and recovery fails, health problems such as chronic stress, depression, or burnout can occur.

Continuous monitoring of mental workload offers new opportunities to support preventing mental disorders and maintaining mental health. Most of the existing studies try

to discriminate a state of mental load from a resting condition in a laboratory setting. In [1] and [21], two stress factors were investigated under laboratory conditions: high cognitive load under time pressure and social-evaluative threat. In both studies, mild cognitive load was discriminated from a constant high-stress level. In [22] a mental arithmetic task was used to induce mental workload and the recovery patterns of physiological responses as indicators of stress were investigated. Kim et al. [12, 13] studied heart rate variability (HRV) features of subjects under chronic stress. Subjects were divided into a high-stress group and a low-stress group based on their self-reporting stress scores. Henelius et al. [10] investigated the ability of short-term HRV metrics to discriminate between low and high level of mental workload.

Continuous monitoring of work-related stress or mental workload is still in an exploratory stage. One example is the ambitious research project "Mobile Heart Health," which aims to detect early signs of stress triggered by physiological or contextual changes [18]. The authors used HRV as stress indicator and since individuals vary dramatically in their HRV values, they addressed the importance of an individually calibrated and adaptive system. It was proposed that each subject's baseline and stress threshold should be established in a laboratory setting using a protocol to alternately evoke stress responses that can then be used to discriminate between stress and non-stress in everyday life. However, an experimental evaluation about the feasibility of discriminating mental workload levels in everyday life scenarios by incorporating individual calibration measures is missing.

In our previous work [4], we already presented our first steps toward monitoring of mental workload in daily life. In this work, we present how mental workload levels in everyday life scenarios can be discriminated by incorporating individual calibration measures. Since for an "everyday life application," a minimal sensor setup is desired for comfort reasons, we employ a single sensor modality: a mobile system to measure heart rate (HR). The analysis of the heart rate variability (HRV) was chosen, because it represents a sensitive stress and mental load measure by providing information about the activity of the sympathetic and parasympathetic nervous system. In addition to the above-mentioned works, numerous studies reported the reliability of psychophysiological responses induced by mental workload tasks [15, 19, 24, 25]. In this work, we investigate HRV features in the time as well as in the frequency domain.

1.1 Research contribution

The present study enhances the state of the art in two ways. First, compared to other studies that mostly tried to discriminate mental stress from a baseline condition, we are investigating different levels of mental workload occurring in everyday life. Second, we target the variation of individual's response to stress by calibration measures. The reason behind is that recently the need to address individual differences was highlighted. Morris et al. [18] proposed to establish each subject's baseline and stress threshold in a laboratory setting by evoking sympathetic and parasympathetic responses. In the presented study we have actually implemented this proposal by designing and performing a calibration procedure to measure each subject's sympathetic and parasympathetic responses during three different levels of mental workload (low, medium, and high) in a laboratory experiment. By doing so, each subject's baseline and workload heart rate features were established in a controlled laboratory setting. Afterward, we have investigated whether the data collected in our calibration session were appropriate to discriminate the low, medium and high mental workload levels occurred during a daily life scenario, i.e., office-work. For this, we used the individual HRV responses of each workload level to train our models and test the trained models on the data collected while the subjects performed normal office-work.

In the following we first give an overview about the measurement system. Then we describe our experiment design to induce three different levels of mental workload in calibration sessions and to monitor mental workload levels in everyday life scenarios. Afterward we introduce the data processing methods and finally we present and discuss our results.

2 Data collection

2.1 Mobile ECG measurement

The physiological responses were measured with the Zephyr BioHarness chest belt as depicted in Fig. 1. The monitoring belt consists of three smart fabric sensors to acquire cardiac activity, breathing rate and skin temperature [27]. The ECG data was sampled with 250 Hz. In addition to ECG data, the chest belt provides RR intervals by measuring the duration between two consecutive R waves of the ECG.

2.2 Experiment

Seven healthy subjects participated in this study (age between 25 and 34 years). Due to the effects of oral contraceptives and menstrual cycle phase on HRV, we decided to restrict the sample to male subjects as it is common practice in many biomedical studies related to stress or cognitive load [14, 20].

Fig. 1 Zephyr BioHarness monitoring system

In a first step, a calibration setting was designed to measure individual responses when confronted with three levels of mental workload in a laboratory setting. In a second step, mental workload levels in an everyday life scenario were investigated. The purpose of the overall experiment was to estimate each subject's perceived mental workload level occurred during a daily office-work by employing the data obtained in the laboratory calibration setting. Therefore, the overall experiment consisted of four sessions: the first three sessions were designed to induce three levels of mental workload in order to conduct an individual calibration (the *calibration conditions*); in the fourth session, subjects were monitored during 1 h of normal office-work (the *office-work condition*) that contained working activities such as programming, and reading or writing research papers. Subjects performed each session in different days. The whole experiment ends up with 4.5 h of data for each and 31.5 h of data for all subjects (calibration condition lasts 1 h, and the office-work session takes one and half hour including questionnaires and cortisol collection). The experimental procedure can be seen in Fig. 2.

Directly after each workload period in the calibration and the office-work conditions, each subject was asked to indicate his perceived workload by completing the NASA Task Load Index (TLX) [9]. First, the subject had to rate each workload phase with 6 items on a scale from 1 to 20 that best indicate his experience in the task. The rating consists of the following items: mental demand, physical demand, temporal demand, own performance, effort, and frustration. Next, the subject was asked to indicate which of the items represents the most important contributor to the workload. Based on these ratings, the total workload was computed as a weighted average. In addition to subjective workload, saliva samples were repeatedly collected with salivettes (Sarstedt, Sevelen, Switzerland), in order to measure cortisol, an important stress hormone indicating the activity of the hypothalamus–pituitary–adrenal (HPA) axis [16]. Subjects had to chew the salivettes for 1 min, immediately before and after each workload period, during the office-work, and 15 min after the completion of each condition (Fig. 2). Saliva samples were stored at −20°C, before biochemical analysis was conducted (Biochemical Laboratory, Dept. of Clinical Psychology and Psychotherapy, University of Zurich, Zurich, Switzerland). Saliva samples were centrifuged for 5 min at 3000 rpm and analyzed using an immuno-assay with time-resolved fluorescence detection [7].

2.2.1 *Calibration conditions: investigation of mental workload levels*

Since individual's response to stress can vary to a huge extend, Morris et al. [18] proposed to establish each subject's baseline and stress threshold in a controlled laboratory setting. In this section, we present our implementation of such a controlled calibration procedure. We have induced three levels of mental workload and measured the individual responses with a mobile ECG system, NASA-TLX, and saliva samples. Three sessions with low, medium, and high workload were defined, while each session consisted of a "baseline," "workload," and "recovery" period. Subjects performed each session on separate days in the afternoon, in order to control for circadian rhythms, while the different sessions were randomly assigned for each subject, in order to avoid sequence effects and, therefore, to counterbalance learning effects. Additionally, we recorded the individual performance during each task. The baseline and recovery periods were the same for the three sessions: the subjects watched a relaxing documentary film in order to calm down. The workload phases differed in the amount of induced mental workload. We used three variants of the Dual N-Back Task [2, 11] to induce low, medium, and high mental workload as outlined in the following:

1. **Position 1 Back (Low workload; very easy task with visual stimuli):** A square appears every 4.5 s in one of eight different positions on a regular grid on the screen. By using the keyboard, the subject has to indicate, if the position of the currently shown square is the same as the one that was presented just before (1-back task). This kind of workload is comparable to monotonous monitoring tasks, where the subject has to sustain his attention at the same level.
2. **Arithmetic 1 Back (Medium workload; easy task with combined visual and auditory stimuli):** An integer number between 0 and 9 appears every 4.5 s on the screen. For each number, a math operator (add, subtract, multiply, or divide) is presented via an audio message. The subject has to apply the math operation on the currently shown number and the one that was presented before (1-back task). The result of the calculation has then to be entered on the keyboard. This task reflects medium cognitive load, since the

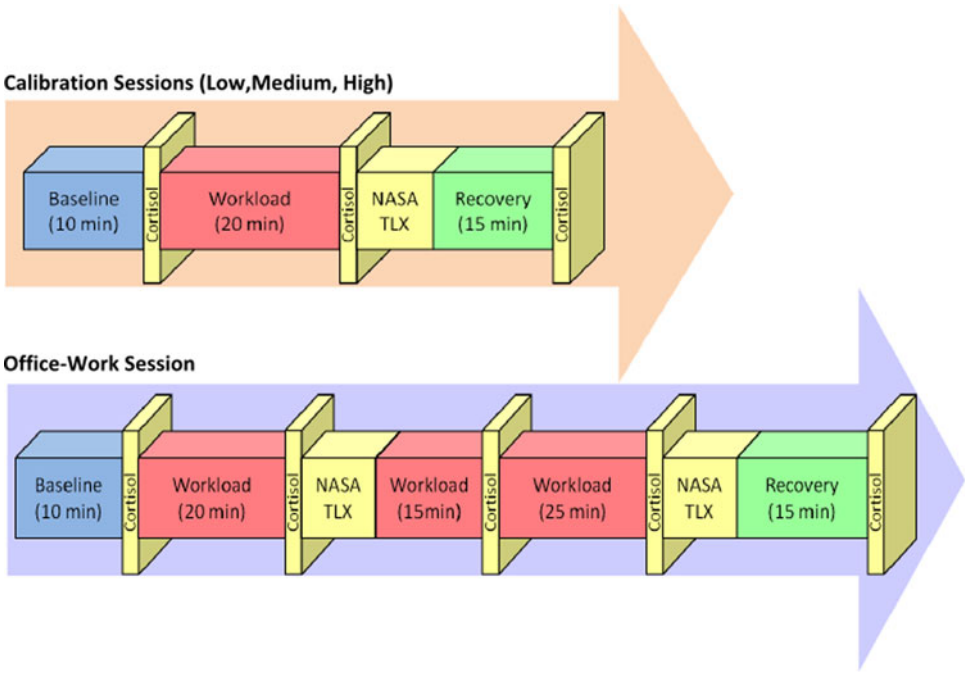

Fig. 2 Experiment procedure for calibration and office-work sessions. A total of three calibration sessions were conducted which differed in the level of induced workload: low, medium, and high. The office-work condition consisted of 1 h of normal office working activities. The subjective rating of perceived workload was assessed with the NASA-TLX, whereas an objective measurement was assessed by collecting salivary cortisol at particular points in time

subject has to memorize one number and to perform a math task in the given time.

3. **Dual Arithmetic 2 Back (High workload; demanding task with combined visual and auditory stimuli):** In this mode, the two former position and arithmetic tasks are combined. An integer number between 0 and 9 appears every 4.5 s in one of eight different positions on a regular grid. For each number, a math operator (add, subtract, multiply, or divide) is presented via an audio message. The subject has to respond if the position of the currently shown number is the same as the one that was presented two positions back (2-back task). In addition, the subject has to apply the math operation on the currently shown number and the one that appeared 2 positions back. The result of the calculation has then to be entered on the keyboard. An example of this task is shown in Fig. 3. This task represents a high cognitive load, since the subject has to memorize the position of a prior value, compare it with a current value, and has to perform a math task under time pressure.

2.2.2 *Office-work condition: monitoring of mental workload during office-work*

During the office-work condition, the subjects performed their daily office tasks for 1 h. In the baseline and recovery periods, the subjects watched a relaxing documentary film in order to calm down. After 20 min of workload and directly after the completing the workload period, subjects were asked to indicate their perceived workload by completing the NASA Task Load Index.

3 Data analysis

This section describes the employed data analysis methods. In a first step, we preprocessed the ECG data and extracted relevant time and frequency features from the RR interval data. Afterward, we evaluated subjective and objective measurements of mental workload and applied statistical methods on the extracted features. Figure 4 illustrates the complete data processing chain comprising the steps of preprocessing, feature extraction, and application of methods.

3.1 Preprocessing and feature extraction

For the analysis of the cardiac data, we first removed RR intervals that differed more than 20% from their predecessors in order to remove artifacts. Due to the high data quality, for each subject less than 1% of the RR intervals were removed. In the next step, we extracted time and frequency domain features that were recommended by the Task Force of the European Society of Cardiology and North American Society of Pacing and Electrophysiology [17]. In the present work, we calculated the following time and frequency domain features following the guidelines of the European Task Force:

Time Domain Features: The following eight commonly used time domain features were calculated: mean value of the heart rate (*Mean HR*), standard deviation of the heart rate (*STD HR*), mean value of the RR intervals (*Mean RR*), standard deviation of the RR intervals (*SDNN*), root mean square of successive difference of the RR intervals

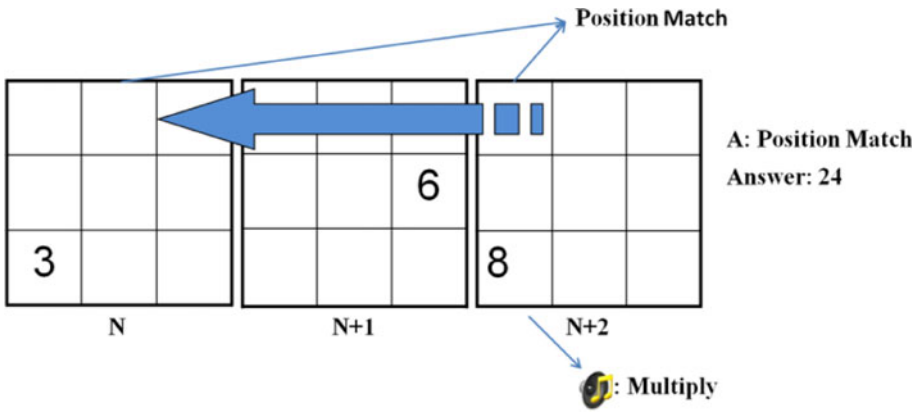

Fig. 3 Dual Arithmetic 2 Back Task was used to induce high mental workload on subjects. An integer number between 0 and 9 appears every 4.5 s in one of eight different positions on a regular grid. In each step, a math operator (add, subtract, multiply, or divide) is presented via an audio message. The subject has to respond if the position of the currently shown number is the same as the one that was presented two positions back. In addition, the subject has to apply the math operation on the currently shown number and the one that appeared 2 positions back

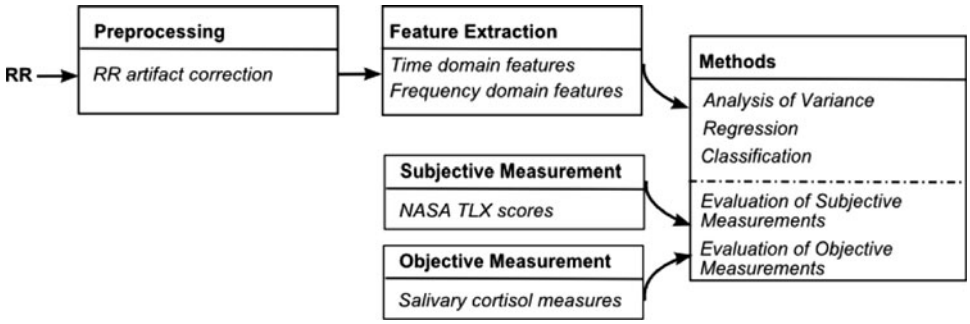

Fig. 4 Block diagram showing the preprocessing, feature extraction, subjective and objective measurements, and mental workload evaluation steps

(*RMSSD*), the percentage of the number of successive RR intervals varying more than 50 ms from the previous interval (*pNN50*), the total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s (*HRV triangular index*), and triangular interpolation of RR interval histogram (*TINN*).

Frequency Domain Features: The extraction of HRV features in the frequency domain was done using the Lomb periodogram since it does not require resampling of unevenly sampled signals such as RR data [5]. We used two frequency bands defined as follows: low frequency (LF): 0.04–0.15 Hz and high frequency (HF): 0.15–0.4 Hz. Next, we calculated the normalized values of LF, HF, and LF/HF, which represents the relative value of each power component in proportion to the total power minus the very low frequency (VLF) component. In this work, we used the ratio of LF and HF (*LF/HF*) as the frequency domain feature of the HRV signal. The LF/HF ratio is known to be an indicator for sympathovagal balance. High values indicate the dominance of sympathetic activity, whereas low values indicate a switch toward a dominance of parasympathetic activity.

3.2 Methods

In a first step, we investigated the subjective ratings of the total workload obtained with the NASA Task Load Index (subjective measure). We compared the individual ratings of each calibration period to see, if the participants perceived the induced workload levels as intended from the experiment design. Next, we examined the relation between each calibration period and the salivary cortisol measures (objective measure). In addition, we analyzed the individual task performance.

After evaluating the subjective and objective measures, we divided the recordings of each subject and each experiment condition (calibration and office-work) into the experiment phases "baseline," "workload," and "recovery." Next, we calculated all HRV features for each phase of the experiment. In order to test whether different workload conditions (i.e., low, medium, and high) had any effects on the outcome of HRV parameters, we compared extracted features by using the analysis of variance (ANOVA) test. As significance level, $p < 0.05$ was considered.

After statistical analysis, we created data segments each containing 2 min of data with 50% overlapping for

"baseline" and "workload" phases. In all segments, the above-mentioned HRV features were computed. Since each subject performed each experiment condition on four different days (i.e., 3 days for low-, medium-, and high-workload calibration, and 1 day for office-work), we divided the features obtained during the workload periods by the corresponding mean value of the baseline feature in order to control for daily variations. In the following, we denote these features as "relative features."

Our next goal was to develop a model based on the calibration data that for a given 2-min RR signal (a) predicts the corresponding subjective workload score by using relevant HRV features and (b) identifies the mental workload class (low, medium, or high) to which the new observation belongs. For the first problem, we employed multiple regression analysis to model the relationship between HRV features and the subjective ratings of NASA-TLX. In this work, the predictor variables are non-correlated HRV features and the response variable is NASA-TLX score. For the second problem, we employed and compared the performance of three classification methods: linear discriminant analysis (LDA), k-nearest neighbor algorithm (k-NN), and SVM (with linear kernel). LDA and k-NN algorithms were applied using MATLAB. The classification results of the support vector machines (SVM) were obtained using MATLAB Arsenal toolbox [26] that encapsulates various classification algorithms and machine learning packages such as WEKA or libSVM [3]. For the SVM classification, we used the libSVM implementation of the MATLAB Arsenal package with a linear kernel and the default cost factor 1. For the multiple regression and all three classification models, we used the entire "calibration" data as training set and "office-work" data as test set. This means, the model parameters were estimated using the "calibration" data as observed data, and the predictions of the "office-work" session has been done using these model parameters for each subject.

4 Results

In the following, we first present the results of subjective and objective measurement of mental workload. Then, we present the achieved results of analysis of variance, multiple linear regression, and classification methods.

4.1 Subjective measurement of mental workload

Figure 5 shows subjective workload scores for each subject. It can be seen that all subjects perceived the induced load levels by the three variants of the N-Back as intended from the experiment design (ANOVA, $p < 0.001$). Compared to the calibration sessions, subjective workload scores of the office-work session were ranked either between low and medium (subjects 1, 3, 5, and 6) or between medium and high (subjects 2, 4, and 7). A multiple comparison test between each group of workload sessions revealed that subjective workload of the office-work session differ significantly from low and high workload ($p < 0.001$) but not from the medium workload session ($p = 0.88$). The visualization of differences between each group can be seen in Fig. 5 (right).

In order to see the variation of the perceived subjective workload over time, we actually have asked the subjects to fill out the self-assessment NASA questionnaire twice (after 20 min and at the end) during one-hour office-work. However, we applied the methods described in the previous section using the NASA results obtained at the end of the working session since the subjective assessments after 20 min were nearly the same like the ones obtained at the end of the working session. This can be seen in Fig. 6.

Afterward, in order to assign the workload score of the office-work into one of three classes (low, medium, and high), we first defined individual boundaries for low-, medium-, and high-workload levels according to the subjective workload scores collected during the N-Back calibration sessions. The workload score of the office-work session for each subject was assigned according to the following equations,

$$\text{low} < (\text{low}_c + \text{medium}_c)/2$$

$$(\text{low}_c + \text{medium}_c)/2 \leq \text{medium} \leq (\text{medium}_c + \text{high}_c)/2$$

$$\text{high} > (\text{medium}_c + \text{high}_c)/2$$

where low_c, medium_c, high_c represent the subjective scores of low-, medium-, and high-workload periods of the calibration session for a particular subject. Individual boundaries for low-, medium-, and high-workload classes and the subjective rating for the office-work session are depicted in Fig. 7.

4.2 Objective measurement of mental workload

For the analysis of salivary cortisol measurement, we normalized the workload cortisol levels by dividing the last measured cortisol value obtained directly after the recovery phase with the cortisol value obtained after the baseline phase. This enabled us to compare cortisol measurements taken at different days, since we considered baseline differences. Figure 8 shows the normalized salivary cortisol levels of each subject for the different workload periods. It can be seen that with increasing workload levels, four subjects (2, 4, 5, and 7) show increasing levels of cortisol, while two subjects (1 and 6) show decreasing levels of cortisol. In contrast, subject 3 shows the highest cortisol value for the office-work session. ANOVA revealed that no

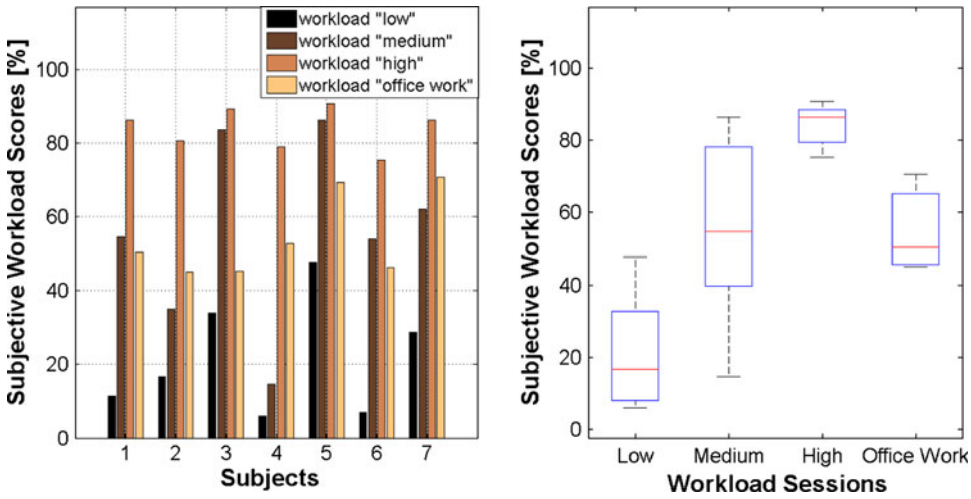

Fig. 5 Subjective workload scores obtained from the NASA Task Load Index for each session and each subject (*left*). Comparison of the workload sessions for all subjects using boxplots (*right*)

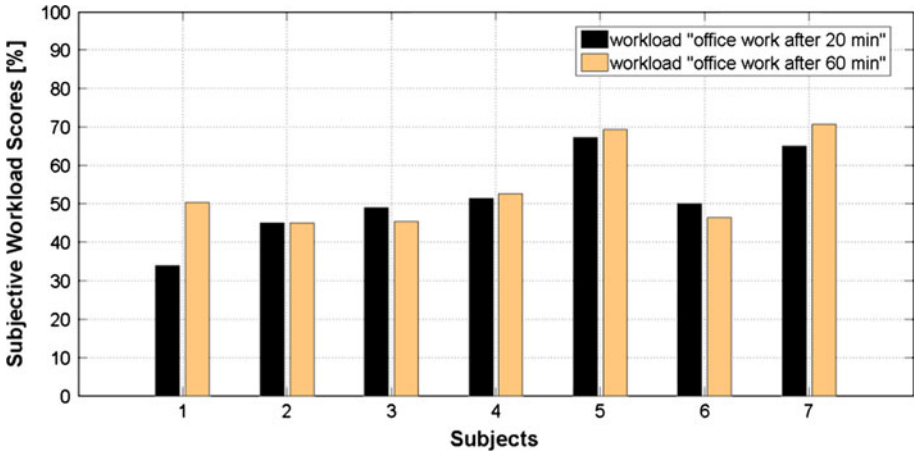

Fig. 6 Comparison of the NASA results from two particular points in time (after 20 min and at the end of the working session)

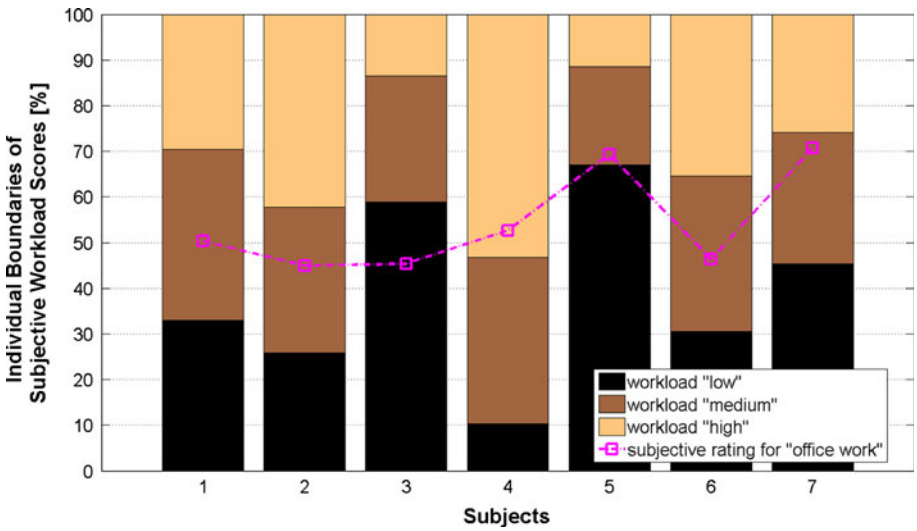

Fig. 7 Individual boundaries for low-, medium-, and high-workload classes and subjective rating for the office-work session

groups have means significantly differ from each other ($p = 0.47$). Varying effects of cortisol responses might be explained by the findings that both uncontrollable and social-evaluative stressors are associated with the largest cortisol changes [6]. In our case, the stressor was a continuous performance task that was controlled and not characterized by social-evaluative threat. By adding social-evaluative threat such as judging the subject about his performance by others during the experiment might increase cortisol levels.

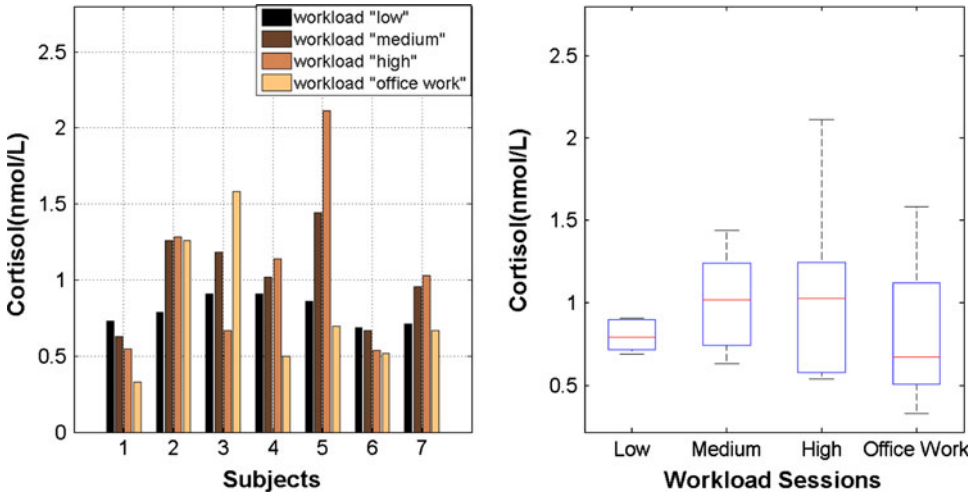

Fig. 8 Normalized salivary cortisol levels of each subject for different workload periods (*left*). Comparison of the workload sessions for all subjects using boxplots (*right*)

4.3 Performance results

In each calibration session, the individual task performance was recorded. In Fig. 9, it is shown that the individual performance reflects the three different workload levels. As can be seen from the figure, there is a significant difference between workload sessions (ANOVA, $p < 0.001$).

4.4 Analysis of variance

We compared the HRV features obtained from the three workload periods in the calibration condition by applying ANOVA tests. The mean values including standard errors of all HRV features extracted for the workload phases are listed in Table 1. It can be observed that the HRV features can be classified into two distinct groups. Features in the first group show consistently a decrease in their values with increased workload. A statistically significant decrease can be observed for the features *RMSSD* and *pNN50* ($p < 0.05$), while *STD HR*, *Mean RR, SDNN, HRV Index*, and *TINN* show a consistent but non-significant decrease. In contrast, features in the second group show an increase in their values with increased workload. A statistically significant increase can be observed for the *LF/HF* ratio ($p < 0.05$).

4.5 Correlation-based feature selection

Before applying regression and classification, we employed a feature selection using a filter approach: since some of the features are expected to be correlated, we investigated the correlation coefficients of the relative HRV features in the 2-min segments of all workload phases. *Mean HR, STD HR*, and *TINN* were excluded from the analysis, because of the high correlations between *Mean HR* with *Mean RR*, *STD HR* with *SDNN*, and *TINN* with *SDNN* ($r > 0.9$).

4.6 Multiple linear regression

We examined the relationship between subjective workload scores and HRV features. Multiple linear regression analysis was performed with NASA-TLX as the response variable. For each subject, the multiple linear regression coefficients are shown in Table 2. Please note that the regression coefficients in the table were computed by fitting the linear regression using the calibration data. The NASA-TLX scores of the office-work session were then predicted based on this model. Figure 10 shows the predicted workload scores of the individual office-work sessions.

In order to evaluate the regression results, we considered the following evaluation metrics:

(a) **Predicted class:** The class to which the majority of predicted values falls into.
(b) **Accuracy:** The percentage of predicted values that falls into the correct class.

By using these metrics, we can transform the regression problem into a classification problem using the majority rule.

Table 3 shows the actual workload scores of the office-work session, their actual class, and the results of the proposed metric. As seen in the table, the assigned class of the office-work session was correct for all but the third subject.

4.7 Classification

Table 4 shows the classification results for each subject. As in multiple linear regression, the class to which the majority of predicted values fall into is considered as classification result. It can be observed that the maximum accuracy is achieved by LDA (correct classification for 6

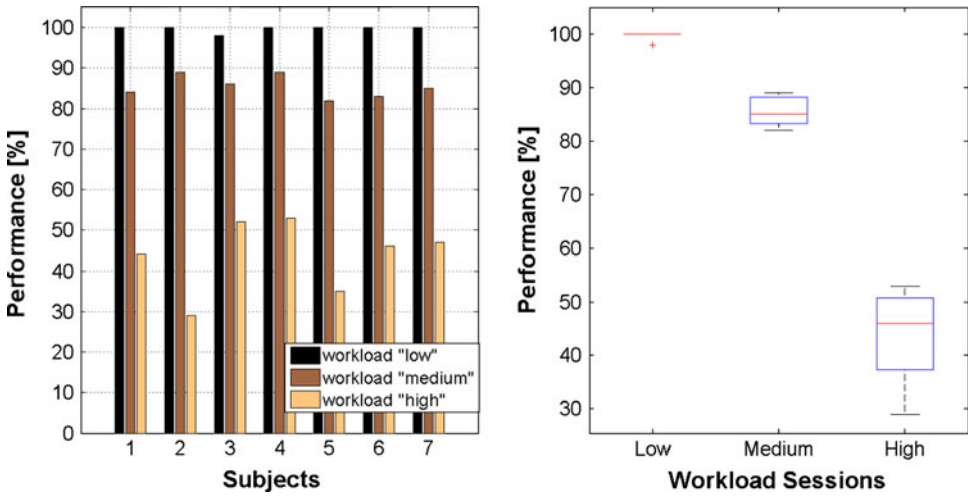

Fig. 9 Performance scores of each subject for each N-Back session (*left*). Comparison of the workload sessions of the calibration condition for all subjects using boxplots (*right*)

Table 1 Comparison of mean HRV features ± standard error during low, medium, and high workload in the calibration condition

HRV features	Low workload	Medium workload	High workload	F; p
Mean HR (1/min)	69.6 ± 2.5	76.1 ± 3.9	80.2 ± 5.5	1.62; 0.22
STD HR (1/min)	5.8 ± 0.7	5.4 ± 0.5	5.2 ± 0.4	0.29; 0.75
Mean RR (ms)	875.3 ± 32.2	803.2 ± 36.5	769.1 ± 43.0	2.09; 0.15
SDNN (ms)	72.2 ± 8.4	58.7 ± 7.8	51.5 ± 6.4	1.89; 0.18
RMSSD (ms)*	51.6 ± 5.2	38.7 ± 4.4	31.2 ± 4.6	4.65; 0.02
pNN50 (%)*	30.7 ± 4.8	19.3 ± 3.6	12.4 ± 3.2	5.48; 0.01
HRV index	19.5 ± 2.4	14.9 ± 1.8	13.0 ± 1.5	2.86; 0.08
TINN (ms)	462.8 ± 45.7	385.7 ± 53.1	385.1 ± 53.7	0.77; 0.48
LF/HF (n.u)*	1.9 ± 0.2	2.5 ± 0.3	4.6 ± 1.0	4.59; 0.02

Mean ± standard error

* $p < 0.05$

Table 2 Summary of multiple regression coefficients: NASA-TLX as dependent variable and HRV features as independent variables

Features	Subj1	Subj2	Subj3	Subj4	Subj5	Subj6	Subj7
Mean RR	11.57***	15.048***	5.983	52.794***	5.863*	6.13	3.893
SDNN	−4.512	−3.758	3.064	18.052*	7.385***	−2.839	−12.15*
RMSSD	5.023	−2.957	5.131	−61.16***	−9.768*	−4.292	−3.887
pNN50	6.067	−0.033	−2.964	−4.399	8.444**	12.731*	23.013***
HRV index	5.018*	15.285**	9.713**	1.544	6.062**	7.315	5.885
LF/HF	15.04***	6.697	4.251	24.966***	3.947*	11.499***	8.238***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

subjects), whereas KNN and SVM worked successfully for 5 subjects.

5 Conclusion and future work

In this work, we have presented how mental workload levels in everyday life scenarios can be discriminated with data from a mobile ECG logger by incorporating individual calibration measures. We have presented an experiment design to induce three different levels of mental workload in a calibration session and to monitor mental workload levels in everyday life scenarios. Seven healthy male subjects participated in this study. Besides the recording of ECG data, subjective rating of the perceived workload was collected with the NASA Task Load Index, whereas an objective measurement was assessed by collecting salivary cortisol. According to the subjective ratings and the performance of the participants in the calibration conditions, we could show that all participants perceived the induced load levels as intended from the experiment design. In accordance, the performance decreased with increasing workload. Compared to the calibration conditions, subjective workload scores of the office-work session were ranked either between low and medium or between medium and high. In order to assign the workload score of the office-work into one of three classes (low, medium, and high), individual boundaries according to the subjective

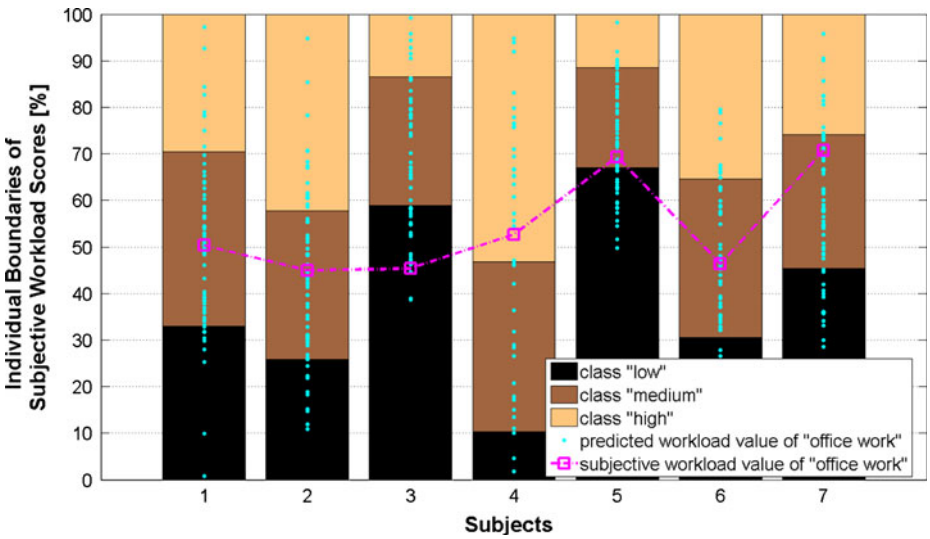

Fig. 10 Predicted workload scores of the office-work session based on linear regression model

Table 3 Workload score, actual workload class, and estimated class with corresponding accuracy

Subjects	NASA score (office-work) (%)	Actual class	Predicted class	Accuracy (%)
1	50.33	Medium	Medium	69.8
2	45	Medium	Medium	59.4
3	**45.33**	**Low**	**Medium**	**34.9**
4	52.66	High	High	38.1
5	69.33	Medium	Medium	51.5
6	46.33	Medium	Medium	66.7
7	70.66	Medium	Medium	61.3

False identified classes are indicated in bold

Table 4 Classification results for each subject

Method	Subj1	Subj2	Subj3	Subj4	Subj5	Subj6	Subj7
True class	M	M	L	H	M	M	M
LDA	M (55.55)	M (37.50)	**M (33.33)**	H (49.20)	M (54.54)	M (50.79)	M (37.09)
KNN	M (57.14)	M (46.87)	**M (30.15)**	**L (23.80)**	M (51.51)	M (44.44)	M (48.38)
SVM	M (47.61)	**L (32.81)**	L (41.26)	**M (19.04)**	M (43. 93)	M (53.96)	M (43.54)

False identified classes are indicated in bold

Predicted class (Accuracy %)

L low, *M* medium, *H* high

workload scores collected during the calibration conditions were defined. By applying ANOVA tests, the HRV features from the calibration conditions could be classified into two distinct groups with respect to their response: with increasing workload, features in the first group showed a decrease in their values, while features in the second group showed an increase in their values. The features *RMSSD* and *pNN50* showed a statistically significant decrease while *LF/HF* ratio showed a statistically significant increase with increased workload. The remaining features showed a consistent but non-significant increase or decrease, what might be explained by the limited number of subjects. We employed multiple regression analysis to model the relationship between relevant HRV features and the subjective ratings of NASA-TLX. Thereby the model parameters were estimated using the calibration data in order to predict the mental workload levels during office-work. The resulting predictions were correct for six out of the seven subjects. In only one subject, there was a confusion between low and medium workload. In addition, we employed and compared the performance of three classification methods to identify the mental workload class (low, medium, or high) to which a new observation belongs. As in multiple regression analysis, the classification models were trained using the calibration data in order to predict the mental workload levels during office-work. The best results were obtained with linear discriminant analysis (LDA) that yielded a correct classification for six out of the seven subjects. The only confusion between low and medium workload occurred for the same subject as in multiple regression analysis. The k-nearest neighbor algorithm and the support vector machine (SVM) resulted in a

correct classification of the mental workload level during office-work for five out of the seven subjects. In conclusion, we were able to discriminate the perceived mental workload level during an office-work scenario by modeling the relationship between relevant HRV features and the subjective ratings in calibration settings.

In future work, we are going to extend the amount of monitoring periods in daily life to several days or weeks. In addition, we have to increase the number of subjects to obtain a more balanced collective, e.g., regarding subject's age. In order to minimize the disturbance of the participants, we will restrict ourselves to mobile ECG logging and 3–5 questionnaires for self-assessment per day. Such a data basis would allow investigating daily variations of perceived and objectively measured mental workload. In addition, we are going to target a broader variety of everyday life scenarios. Up to now, we have investigated office-work in front of a computer. In future work, we will target other activities like giving lectures. In particular, we will investigate whether the presented calibrations method (3 levels of N-Back tasks) is appropriate or which modifications are necessary to model different kinds of real world workload.

References

1. Arnrich B, Setz C, La Marca R, Tröster G, Ehlert U (2010) What does your chair know about your stress level? IEEE Trans Info Technol Biomed Affect Perv Comput Healthc
2. Brain workshop—a dual n-back game. http://brainworkshop.sourceforge.net/
3. Chang C–C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol 2:27:1–27:27
4. Cinaz B, La Marca R, Arnrich B, Tröster G (2010) Monitoring of mental workload levels. In: Proceedings of IADIS eHealth conference
5. Clifford GD (2002) Signal processing methods for heart rate variability analysis. PhD thesis, St Cross College
6. Dickerson SS, Kemeny ME (2004) Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychol Bull 130:355–391
7. Dressendörfer RA, Kirschbaum C, Rohde W, Stahl F, Strasburger CJ (1992) Synthesis of a cortisol-biotin conjugate and evaluation as a tracer in an immunoassay for salivary cortisol measurement. J Steroid Biochem Mol Biol 43(7):683–692
8. European Foundation for the Improvement of Living and Working Conditions (2007) Work-related stress. http://www.eurofound.europa.eu/
9. Hart SG, Stavenland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Human mental workload, chapter 7. Elsevier, Amsterdam, pp 139–183
10. Henelius A, Hirvonen K, Holm A, Korpela J, Muller K (2009) Mental workload classification using heart rate metrics. Conf Proc IEEE Eng Med Biol Soc 1:1836–1839
11. Jaeggi SM, Buschkuehl M, Jonides J, Perrig WJ (2008) Improving fluid intelligence with training on working memory. Proc Natl Acad Sci USA 105:6829–6833
12. Kim D, Seo Y, Salahuddin L (2008) Decreased long term variations of heart rate variability in subjects with higher self reporting stress scores. Perv Healthc
13. Kim D, Seo Y, Cho J, Cho C-H (2008) Detection of subjects with higher self-reporting stress scores using heart rate variability patterns during the day. Conference Proceedings IEEE Engineering in Medicine and Biology Society, pp 682–685
14. Kirschbaum C, Kudielka BM, Gaab J, Schommer NC, Hellhammer DH (1999) Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. Psychosom Med 61:154–162
15. Kramer AF (1991) Physiological metrics of mental workload: a review of recent progress. Multiple-task performance, pp 279–328
16. Marca RL, Waldvogel P, Thorn H, Tripod M, Wirtz PH, Pruessner JC, Ehlert U (2010) Association between cold face test-induced vagal inhibition and cortisol response to acute stress. Psychophysiology
17. Marek M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, Schwartz PJ (1996) Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation 93:1043–1065
18. Morris M, Guilak F (2009) Mobile heart health: project highlight. IEEE Perv Comput 8(2):57–61
19. Riener A, Ferscha A, Aly M (2009) Heart on the road: Hrv analysis for monitoring a driver's affective state. In: AutomotiveUI '09: proceedings of the 1st international conference on automotive user interfaces and interactive vehicular applications. ACM, New York, NY, USA, pp 99–106
20. Sato N, Miyake S, Akatsu J, Kumashiro M (1995) Power spectral analysis of heart rate variability in healthy young women during the normal menstrual cycle. Psychosom Med 57:331–335
21. Setz C, Arnrich B, Schumm J, La Marca R, Tröster G, Ehlert U (2010) Discriminating stress from cognitive load using a wearable EDA device. IEEE Trans Inf Technol Biomed Person Health Syst
22. Soga C, Miyake S, Wada C (2007) Recovery patterns in the physiological responses of the autonomic nervous system induced by mental workload. In: SICE, 2007 annual conference, pp 1366–1371
23. van Daalen G, Willemsen TM, Sanders K, van Veldhoven MJPM (2009) Emotional exhaustion and mental health problems among employees doing people work: the impact of job demands, job resources and family-to-work conflict. Int Arch Occup Environ Health 82:291–303
24. Wilson GF (2002) An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int J Aviat Psychol 12:3–18
25. Wilson GF, Eggemeier FT (1991) Psychophysiological assessment of workload in multitask environments. Multiple-task performance, pp 329–360
26. Yan Rong (2006) MatlabARSENAL: a MATLAB package for classification algorithms. School of Computer Science, Carnegie Mellon University, Pittsburgh
27. Zephyr. http://www.zephyr-technology.com/