Diss. ETH No. 21091

Monitoring of Cognitive Load and Cognitive Performance using Wearable Sensing

A dissertation submitted to

ETH Zurich

for the degree of

Doctor of Sciences

presented by

Burcu Cinaz

Dipl. -Inf., University of Bremen, Germany
born April 27, 1981
citizen of Turkey

accepted on the recommendation of

Prof. Dr. Gerhard Tröster, examiner
Prof. Dr. Mike Martin, co-examiner

2013

Burcu Cinaz

Monitoring of Cognitive Load and Cognitive Performance using Wearable Sensing

Diss. ETH No. 21091

First Edition 2013
Published by ETH Zürich, Switzerland

ISBN 978-1-304-29249-0

Printed by Lulu.com

Acknowledgments

First of all, I would like to thank my supervisor Prof. Dr. Gerhard Tröster for giving me the opportunity to write this dissertation at the Electronics Laboratory. I thank him for all the support and encouragement he gave me during these past four years. He provided me with an excellent research infrastructure at the Wearable Computing Lab. It has been a honor to graduate from his Lab.

I would also like to thank Prof. Dr. Mike Martin for co-examining and reviewing my PhD thesis. At this place, I would like to thank Dr. Roberto La Marca, Nathan Theill and Vera Schumacher from the Department of Psychology at the University of Zurich for their valuable discussions, inputs and advices from a psychological perspective. This collaboration enabled an inspiring interdisciplinary research.

I owe a huge thank you to Dr. Bert Arnrich for his incredibly support on my dissertation, professional and personal issues, even during tough times in the PhD pursuit.

I am grateful to the members of the Electronics Laboratory that provided a great environment to work. Especially, I would like to thank my office mates Sebastian, Amir and Franz. Special thanks go to Mirco for the interesting teamwork on the wearable electronics and the realization of a fashion show. Thanks also to Alberto, Bernd, Christian, Christina, Christoph, Daniel W., Giovanni, Julia, Lars, Long-Van, Luisa, Martin W., Michael, Niko, Rolf, Sinziana, Thomas K., Thomas S., Tobias, Ulf and Zack.

I would also like to thank my semester and master students Philip Dieringer, Sebastian Steiner, Thomas Steiner, Christian Vogt and Yi Li for their contribution to several experiments and some of the results of this thesis. Special thanks go to our test subject Sonja Zagermann who allowed us to monitor her daily working routine throughout 15 days.

Furthermore, I would like to thank Ruth Zähringer for her professional help for any kind of problems. Thanks also to Fredy Mettler for his help with any technical problems.

Most importantly, I want to thank my family for supporting me during my PhD studies.

Zurich, May 2013 Burcu Cinaz

Contents

Abstract

In recent years there is more and more evidence for a significant increase of work-related stress burden and disease in the Western civilization. If high level of work demands cumulate and recovery fails, serious mental health problems such as chronic stress or depression can occur. Monitoring of cognitive load would allow supporting the prevention of mental disorders and maintaining mental health. The first part of this thesis is directed towards paving the way for a continuous monitoring of cognitive load in daily life scenarios.

Cognitive tests allow measuring the cognitive performance of a person. Basic performance measures are capacity of remembering, reaction time and attention. Conducting cognitive assessment tests throughout daily life offers opportunities to early detect changes in cognitive performance. In most studies, cognitive performance is measured with computerized tests which are not well suited to measure cognitive performance in daily life. In the second part of this thesis, a wearable reaction time test is developed in order to allow obtaining continuous measurements of cognitive performance in daily life.

This thesis comprises six scientific publications that address five aims: (1) to investigate the applicability of heart rate variability (HRV) features to discriminate different levels of mental workload in a mobile setting, (2) to target individual differences in HRV responses by incorporating individual calibration measures, (3) to develop a wearable reaction time (RT) test which can be operated throughout everyday life in order to obtain a continuous measurement of speed of processing by means of RT, (4) to evaluate the feasibility of the wearable RT device with empirical studies, (5) to examine how common daily activities affect the reaction times of young and elderly subjects.

In the first part of this thesis, the applicability of HRV features was investigated in a controlled experiment designed to induce three levels of mental load. According to the subjective ratings of the participants, it was shown that all participants perceived the induced load levels as intended from the experiment design. The investigated HRV features obtained from a mobile ECG logger showed significant differences between the three load levels.

The knowledge gained from the controlled laboratory experiment was then transferred to an office-work setting. Since each individual's physiological response to high mental load can vary depending on certain factors, a calibration procedure was introduced. It was examined whether the data collected in the calibration session were appropriate to discriminate low, medium and high mental load levels occurred during a daily life office-work scenario. The overall results showed that in 6 out of 7 participants the self-reported load levels perceived during office tasks could be modeled by incorporating individual calibration measures.

In the second part of this thesis, a wearable RT test was designed and implemented in order to allow a continuous operation throughout everyday life. The watch-like RT device combines the generation of a haptic stimulus and the recognition of subject's hand movement as response gesture. The feasibility of the wearable RT test was evaluated with two empirical studies. The first study showed that the wearable RT tests are suitable to measure factors that influence length and variability of reaction times. In the second study, a long-term monitoring in an unrestricted real working environment of a graphic designer was conducted. Reaction time data under different workload factors such as stress, sleep deprivation, night shift and moderate alcohol were continuously collected throughout 15 working days. The results showed that the wearable watch-like RT test can be operated without interrupting the working routine of a graphic designer. The investigated RT features showed significant correlations with the workload factors and with the self-reported ratings on mood and perceived workload.

Finally, the thesis investigated how common daily activities affect the reaction times of 14 young subjects (mean age 26 years) and 12 elderly subjects (mean age 70 years). The results showed that RT and RT variability were significantly affected by the type of activity. The increase in mean RT and RT variability from idle to cognitive load condition was significantly higher for the older participants compared to the younger ones. A context-aware wearable reaction time test which considers the type of activity achieved an accuracy of 87.5% when discriminating between idle and load conditions. It was concluded that a wearable RT test combined with an activity recognition system is feasible to detect changes in RT performance and variability during common daily life activities.

Zusammenfassung

In den letzten Jahren konnte in vielen Studien gezeigt werden, dass arbeitsbedingte Stressbelastungen und Erkrankungen in der westlichen Welt signifikant zunehmen. Wenn bei sehr hohen Arbeitsbelastungen keine Erholung mehr möglich ist, können ernste mentale Gesundheitsprobleme wie chronischer Stress oder Depression entstehen. Die Beobachtung von kognitiver Belastung würde die Prävention von mentalen Erkrankungen unterstützen und die mentale Gesundheit erhalten helfen. Der erste Teil dieser Arbeit befasst sich mit der kontinuierlichen Überwachung von kognitiver Belastung im täglichen Arbeitsleben.

Kognitive Tests erlauben die Messung der kognitiven Leistung einer Person. Wichtige Leistungsmasse sind Erinnerungsvermögen, Reaktionszeit und Aufmerksamkeit. Die Durchführung kognitiver Tests im täglichen Leben eröffnet Möglichkeiten, um frühzeitig Veränderungen in der kognitiven Leistung zu erkennen. In den meisten bisherigen Studien wurde die kognitive Leistung mit computerbasierten Tests gemessen. Solche Tests erlauben in der Regel keine Messung in der natürlichen Umgebung der Menschen. Der zweite Teil dieser Arbeit befasst sich mit der Entwicklung eines tragbaren Reaktionszeit-Tests, der eine kontinuierliche Messung der kognitiven Leistung im täglichen Leben erlauben soll.

Diese Arbeit beinhaltet sechs wissenschaftliche Veröffentlichungen, die insgesamt fünf Ziele anstreben: (1) Unterscheidung von verschiedenen mentalen Belastungsstufen mittels Herzraten-Variabilitäts-Analyse in einem mobilen Szenario, (2) Einbeziehung individueller Kalibrationsmasse, um individuelle Unterschiede in der Herzraten-Variabilität zu kompensieren, (3) Entwicklung eines tragbaren Reaktionszeit-Test, der eine kontinuierliche Messung im täglichen Leben erlaubt, (4) Evaluierung des tragbaren Reaktionszeit-Tests mit empirischen Studien, (5) Untersuchung, wie tägliche Aktivitäten die Reaktionszeiten von jungen und älteren Probanden beeinflussen.

Im ersten Teil dieser Arbeit wurden zunächst Charakteristiken der Herzraten-Variabilität bei drei mentalen Belastungsstufen in einem kontrollierten Experiment untersucht. Gemäss der subjektiven Beurteilung der Probanden konnte gezeigt werden, dass alle Probanden die induzierten mentalen Belastungsstufen wie vom Experiment beabsichtigt empfunden haben. Es konnte gezeigt werden, dass sich die von einem mobilen EKG-Rekorder aufgezeichneten Charakteristiken der Herzraten-Variabilität signifikant zwischen den drei Belastungsstufen unterscheiden.

Die aus dem kontrollierten Experiment gewonnenen Erkenntnisse wurden anschliessend auf Büroarbeit übertragen. Da bekannt ist, dass individuelle Unterschiede die physiologische Reaktion auf kognitive Belastung beeinflussen, wurde eine Kalibrations-Prozedur entwickelt. Es wurde untersucht, inwieweit die erhobenen Kalibrationsdaten geeignet sind, um zwischen geringer, mittlerer und hoher kognitiver Belastung während der Büroarbeit zu diskriminieren. Die Ergebnisse haben gezeigt, dass bei 6 von 7 Probanden die Selbsteinschätzung hinsichtlich kognitiver Belastung während der Büroarbeit durch Einbeziehung der individuellen Kalibrationsdaten modelliert werden konnte.

Im zweiten Teil dieser Arbeit wurde ein tragbares Reaktionszeit-Testgerät entwickelt, um eine kontinuierliche Datenaufnahme im täglichen Leben zu ermöglichen. Das Reaktionszeit-Testgerät kann wie einer Uhr getragen werden. Es werden haptische Stimuli erzeugt und eine Handbewegung wird als Reaktion auf den Stimulus erkannt. Die Eignung des tragbaren Reaktionszeit-Testgeräts wurde in zwei empirischen Studien evaluiert. In der ersten Studie wurde gezeigt, dass das tragbare Reaktionszeit-Testgerät geeignet ist, um Veränderungen in der Dauer und Variabilität von Reaktionszeiten zu messen. In der zweiten Studie wurde eine Langzeitmessung in der Arbeitsumgebung einer Grafik-Designerin durchgeführt. Innerhalb von 15 Tagen wurden kontinuierlich Reaktionszeiten bei verschiedenen Arbeitsbelastungen wie Stress, Schlafentzug, Schichtarbeit und moderatem Alkoholeinfluss gemessen. Die Ergebnisse haben gezeigt, dass das tragbare Reaktionszeit-Testgerät verwendet werden kann, ohne das die Arbeitsroutine unterbrochen werden muss. Die Analyse der Reaktionszeitdaten zeigte signifikante Korrelationen zwischen den untersuchten Arbeitsbelastungen und den subjektiven Einschätzungen zu Gemütslage und Arbeitsbelastung.

Im letzten Teil dieser Arbeit wurde untersucht, wie tägliche Aktivitäten die Reaktionszeit von 14 jungen Probanden (mittleres Alter 26 Jahre) und 12 älteren Probanden (mittleres Alter 70 Jahre) beeinflussen. Die Ergebnisse haben gezeigt, dass Reaktionszeit und Reaktionszeit-Variabilität signifikant von der Art der Aktivität beeinflusst wurden. Der Anstieg der Reaktionszeit und der Reaktionszeit-Variabilität zwischen normaler und kognitiver Belastung war bei den älteren Probanden signifikant stärker. Unter Berücksichtigung der Aktivität konnte eine Genauigkeit von 87.5% bei der Diskriminierung zwischen normaler und kognitiver Belastung erreicht werden. Schlussfolgernd wurde aufgezeigt, dass ein tragbares Reaktionszeit-Testgerät die Aktivitäten einer Person berücksichtigten sollte, um Veränderungen in der Reaktionszeit und der Reaktionszeit-Variabilität hinsichtlich kognitiver Belastung messbar zu machen.

1

Introduction

1.1. Motivation

The European Foundation for the Improvement of Living and Working Conditions has been calling attention on work-related stress as a workplace health issue since 1993 [14]. According to the Fourth European Working Conditions Survey in 2005, stress is one of the most common work-related health problems affecting 22% of European workers. Work-related stress is the response people may have when presented with work demands and pressures that are not matched to their knowledge and abilities and which challenge their ability to cope [54]. The changing nature of work due to globalization has made increased cognitive demands on workers resulted in high workload and time pressure. 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. Monitoring of mental workload in everyday life is still in an exploratory stage. Most scientific knowledge about the effects of mental load is based on laboratory research or inferred from retrospective reports, although monitoring of mental load in everyday life is clearly the principal aim. It is known that the subject's response to real-life mental load can be different from the ones observed in the laboratory. Even the most sophisticated experimental design and measurement may not accurately reflect subject's response to mental load. The main limitations of laboratory research are the time scales able to capture and the inadequate mapping of different workload characteristics found in real life. These considerations make it apparent that without monitoring mental load in real life scenarios, one simply cannot be sure of the relevance of laboratory assessment.

Cognitive tests allow to measure the cognitive performance of a person. Basic performance measures are capacity of remembering, reaction time and attention. In most studies, cognitive performance is measured with computerized tests which require the full attention of the test person for several minutes. This restriction prohibits the measurement of cognitive performance during daily routine tasks. In addition, such dedicated tests are not well suited to measure the cognitive performance in daily life since most people are not willing to interrupt their primary task for several minutes. There exists only few approaches to measure the cognitive performance continuously in normal daily routine. Assessing the cognitive capabilities of individuals continuously throughout everyday life activities provides several opportunities: (1) early detection of cognitive decline such as dementia or Alzheimer's disease in elderly, (2) identifying decrease in cognitive performance in order to determine the ability to conduct mental demanding tasks such as driving, piloting or search and rescue, (3) management of mental illness (e.g. whether a patient is responding to a medication), (4) identifying of children with intellectual disabilities such as Attention Deficit Hyperactive Disorder (ADHD).

In this thesis new opportunities of mobile and wearable computing are investigated aiming at identifying changes in cognitive load and cognitive performance in daily life scenarios.

1.2. Assessment of Cognitive Load

In the literature, there are several studies about the measurement of cognitive load, mental load, mental workload or mental effort. Often, the different terms are used as aliases. Psychologists provided theoretic constructs on how the different terms can be conceptualized. For instance, Paas et al. [34, 35] defines cognitive load as a multidimensional construct that represents the load that a particular task imposes on the individual. In other words, it represents the interaction between task demands and subject's capabilities. Theoretically, cognitive load can be conceptualized in the dimensions of mental load, mental effort, and performance and can be measured by assessing these factors. There are four promising approaches in the literature to estimate individual's cognitive load: methods that use subjective ratings, task performance-based indices, behavioral and physiological measures. In the following these four approaches are introduced.

1.2.1. Subjective Assessment

In subjective rating techniques the subject is directly asked to assess his experienced mental effort. Depending on the application scenario, there exist various rating scales for assessing the subjective load. For example, the Cooper-Harper Scale [10] was mainly designed to assess pilot's cognitive ability to steer an aircraft. In this rating scale, the pilots have to assess their experienced demands when performing certain flight tasks. The Subjective Workload Assessment Technique (SWAT) [39] uses the three dimensions of time load, mental effort load, and psychological stress load to assess workload. Compared to SWAT, the

NASA Task Load Index (TLX) [17] uses the six dimensions of mental demand, physical demand, temporal demand, own performance, effort and frustration. The total workload is computed as a weighted average of all scales. In the literature, NASA-TLX has been used in a large variety of contexts. An evaluation of different subjective workload assessment techniques can be found in [43].

A main drawback of subjective assessment methods is the difficulty to perform questionnaire surveys in real life scenarios. For practical reasons, often the subjects are asked to rate their experienced level of cognitive load in a retrospective way. This can lead to the problem of forgetting important aspects concerning the experience depending on the time between the self-report and the cognitive load itself. Besides, the most recent events at the time of the recall can affect the overall rating at the time the questionnaire is asked [60].

1.2.2. Performance Based Assessment

The second approach to assess cognitive load is the measurement of task performance. Thereby, two performance measures are considered: primary task performance as a measure of the actual task performance and secondary task performance for assessing the performance of a task that is operated simultaneously to the primary task [35]. In this context, the performance of the secondary task is influenced by the cognitive load caused by the primary task. The primary task performance can be measured directly while performing a continuous performance task that measures certain aspects such as attention, vigilance, speed or working memory. On the other hand, secondary task performance can be used to evaluate the cognitive load in more real life scenarios such as driving or flying. As a secondary task, usually subjects are asked to recognize and respond to a visual or auditory stimulus while performing the primary task. The performance variables such as the number of errors, accuracy or reaction time obtained from the secondary task is then used as the level of cognitive load experienced by the subject. For instance, Patten et al. [38] used visual stimulus as a secondary task to estimate the levels of cognitive workload of experienced and inexperienced drivers. The visual stimulus was integrated into the peripheral area of the driver's line of sight and the reaction times were recorded as performance variable. Thereby, the impact of cell-phones on driving performance was investigated in a driving simulator by measuring several categories including response time to a pop-up stop sign [6]. Riener et al. [41] compared the driving performance in terms of cognitive load between real driving and simulated driving. Reaction time to visual, auditory and vibro-tactile stimuli was used as performance measure. Visual notifications were placed on the dashboard, auditory information was provided via headphones and vibro-tactile stimulus was given through the driver's seat.

Most of these studies have in common that the secondary task is integrated into the surrounding area of a well-defined and restricted task such as driving. The same tools cannot be used to assess the cognitive load of the subject during other daily life activities which are generally performed in different settings. These examples illustrate that there is a need for cognitive load estimation independent of the current task and surrounding.

1.2.3. Behavioral Measurements

The third approach aims for detecting variations in behavior as a criterion of high workload. Compared to task performance approaches which mostly use a secondary task to assess the performance, in this approach it is investigated how the primary task itself is accomplished by the subject. Similar to task-performance approaches, this method has been studied mostly in the domain of driving. Thereby, certain aspects of the driving behavior such as lane keeping, lane changing, or driver steering performance were measured and evaluated [6, 44–46, 62]. In these studies, mostly the effects of distraction such as using cell-phones during driving were examined. The driving performance was measured by monitoring the steering and speed control during navigating a roadway in a driving simulator. It was found out that there were high correlations between distracting activities such as using a cell-phone and the deviations in speed control or lateral vehicle control. As a consequence, nowadays several car companies are currently involved in research on driver behavior monitoring systems. For instance, recently Mercedes Benz introduced the so-called "Attention Assist" [1, 2] to detect deviations in driving behavior which can be the result of drowsiness or fatigue. The system employs several sensors to record the driver behavior in terms of steering movements, steering speed or pedal use. Based on the recording, the system creates an individual driver profile and in case that a deviation is detected, the system warns the user. A review about similar systems can be found in [12]. Similar to performance based assessments, this type of measurement is specific for the task being studied, e.g. the monitoring of driving behavior was rendered possible by using particular driving simulators during the studies.

1.2.4. Physiological Measures

There exists numerous studies which investigate physiological measures as markers of cognitive load. For instance the effect of varying task difficulties on pupillary responses, eye movements or blink intervals have been studied in various scenarios. For example, Klingner et al. [25] investigated remote video eye trackers for measurements of task-evoked pupillary responses. Subtle changes in pupil size were observed while subjects performing different mental tasks. In addition, effects of cognitive demand on eye movements were systematically investigated in driving scenarios. Victor et al. [57] collected eye movement measures while subjects were performing different in-vehicle tasks with incremental levels of difficulty. Tests were conducted both in a driver simulator and in field trials using an instrumented vehicle which contained a camera based eye tracking system. The results showed that increases in task difficulty produced both a gaze concentration to the road center area and an increase in display viewing time. Similar results were achieved by Engström et al. [13] who showed that the cognitive load resulted in increased gaze concentration towards the road center. Palinko et al. [36] presented a study in which subjects were involved in spoken dialogues while driving a simulated vehicle. The cognitive load of the driver was estimated by measuring the pupil size from a remote eye tracker mounted on the dashboard in front of the driver. In a simulated flight study, Papadelis et al. [37] reported a decrease in eye blink duration during the time of increased mental effort.

The relevance of galvanic skin response (GSR) as an indicator of user's cognitive load was investigated in a traffic control scenario where the investigated tasks had different levels of difficulty [49]. Results showed that mean GSR across users increased as cognitive load increased. The application of GSR in combination with other sensor modalities was mostly used to detect emotions and stress [7, 18, 48, 58].

Measures of surface electromyography (EMG) and electroencephalogram (EEG) were employed to investigate correlations with task difficulty and performance metrics. For example, in a flight simulation study of Papadelis et al. [37] the muscular activity of the subjects were recorded in addition to other physiological parameters. The results showed that there was a significant correlation of EMG measures with the difficulty levels of the task. Berka et al. [8] recorded participants' EEG while they were performing different cognitive tests. The results indicated that EEG measures correlated with both subjective and objective performance metrics.

Heart rate variability (HRV) is one of the most common studied physiological measures for the assessment of mental load. There exists numerous studies which investigated HRV responses induced by mental workload tasks [3, 27, 33, 53, 61]. HRV is a measure of the variability in heart rate, i.e., variations of time intervals between consecutive heart beats [31]. HRV is known to provide information about the activity of the autonomic nervous system (ANS). The ANS is divided into two branches: sympathetic and parasympathetic. The sympathetic nervous system mediates activities associated with stressful situations. As a result, physiological changes such as increased heart rate and blood pressure or increased respiration rate occur. On the other hand, the parasympathetic system counterbalances the sympathetic nervous system by returning the body in a calming state. The analysis of HRV allows to assess sympathetic and parasympathetic activities of the ANS.

HRV measures can be obtained by extracting features in time and frequency domain. Time domain measures are mostly preferred due to simplicity of calculation. Commonly used time domain features include mean value of the RR intervals, standard deviation of the the RR intervals, or the number of interval differences of successive RR intervals greater than a threshold. The extraction of HRV features in the frequency domain provides information on the power distribution as a function of frequency. Power spectral density (PSD) analysis is applied on RR data in order to calculate three main spectral components: very low frequency (VLF), low frequency (LF), and high frequency (HF) components. 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. More detailed explanations for the calculation of the parameters both in time and frequency domain can be found in [31].

Most of the recent studies on HRV try to discriminate a state of mental load from a resting condition in a laboratory setting. In [50] 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. [23, 24] studied 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. Subjects in the high stress group showed a decreased HRV compared to subjects in the low stress group. Henelius et al. [19] investigated the ability of short-term HRV metrics to discriminate between low and high level of mental workload. Field studies were conducted especially in the driving application domain. For example, Riener at al. [42] employed HRV metrics for assessing the driver's affective state. The subjects were monitored while driving between home and work place during a period of two weeks. Higher arousal levels were observed at roads of higher traffic volume. Reimer et al. [40] monitored heart rate together with skin conductance during real highway driving. A secondary cognitive memory task was presented to the user simultaneously to generate different levels of cognitive load. The results indicated that increased cognitive demand was highly associated with the pattern of change in heart rate. The level of mental stress during a simulated office work was investigated in [52]. Subjects performed different levels of mental tasks in which they had to indicate the correct answers to mental arithmetic operations. It was shown that heart rate features allow to differentiate between active and rest condition.

Continuous monitoring of work-related stress or mental workload in daily life is still in an exploratory stage. One example is the ambitious research project "Mobile Heart Health", which aims to detect early signs of stress by identifying physiological and contextual changes [32]. The authors investigated HRV as a stress indicator. Since HRV responses vary between individuals, the authors highlighted 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 which 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.

1.3. Cognitive Assessment with Reaction Time Tests

A reaction time (RT) test measures how rapidly a person can initiate a response to a triggering signal [30]. In other words, RT is the elapsed time between a designated stimulus and the individual's reaction to it, usually measured in milliseconds [22]. The subject being tested is instructed to react to a stimulus (e.g. light turns on). Depending on the experimental setting, the response to a stimulus is defined as a certain action which has to be performed by the subject, e.g. pressing or releasing a button [22]. In addition, the subject has been made aware that the reaction speed is the major focus of the experiment and thus the response to a stimulus should be performed as fast as possible. In contrast, response time test refer to testing conditions in which timing and speed are not explicitly mentioned in experimenter's instructions to the subject [22, 30]. Nevertheless, in the literature the two terms response time and reaction time are often used with the same meaning.

The first experimental investigation of reaction times aiming at understanding the structure of mental activity was presented in the paper "On the Speed of Mental Processes" by F.C. Donders (1868) (translated in English first in [11]) [4, 22, 51]. In his experiment, he recorded RT data from the same subject during two different tasks. The second task was identical to the first one with the exception that an additional mental operation had to be solved. Donders subtracted the reaction times recorded during the simple task from those collected during the more complex one in order to measure the time needed for the additional mental operation. Following Donders, RT has been extensively investigated by experimental psychologists since the middle of the 19th century.

1.3.1. Computerized Reaction Time Tests

There exist mainly three kinds of RT tests: simple, recognition and choice RT tests [5, 26]. Simple RT tests consist of a single stimulus and a single intentional response. For instance, the subject has to press or release a key as fast as possible whenever a particular light or symbol appears on the screen. Differently, in recognition RT tests, there are multiple non-target stimuli but only one dedicated target stimulus. This task is commonly called "go/no-go" reaction time task. In computer-based go/no-go tests the target stimulus and the non-target stimuli are temporary shown in a random sequence on the screen. Lastly, choice RT tests include multiple stimuli and multiple responses. The subject has to respond to each stimulus with a corresponding response, e.g. by pressing a certain key whenever a corresponding letter appears on the screen. A detailed series of recommendations on how to conduct laboratory experiments using RT tests and how to analyze the collected data can be found in [5, 22, 30, 59].

1.3.2. Factors Influencing Reaction Times

Since RT has been extensively investigated for many years, there exists a broad variety of clinical studies where researchers have identified different factors that influence RT. A literature review on factors that affect reaction times can be found in [22, 26]. For example Jensen [22] reports that RT to a visual stimulus is longer than to an auditory stimulus and RT to a haptic stimulus lies in between. Intensity of the stimulus has also a considerable effect on RT. In general, lower stimulus intensities causes longer reaction times. However, as the stimulus intensity becomes very large, the RT approaches asymptotic values, i.e. does not change any more [22, 30].

One of the most important factors that influence length and variability of reaction times are increasing age and age-related diseases like cognitive impairment. It has been known that with increasing age, reaction times become more variable and longer. Hultsch et al. [20] examined age differences in RT variability with four different RT tests. RT variability was increased in older participants in comparison to younger adults. Similarly, Gorus et al. [16] explored RT performance and variability across different complexity levels in young and elderly groups. Older participants showed a general slowing in the overall RT while reaction time variability increased with age and task complexity. In another study of Gorus et al. [15] the authors investigated the reaction times and performance variability as potential markers for the early detection of Alzheimer's disease (AD). The RT data from cognitively healthy elderly, patients diagnosed with mild cognitive impairment (MCI) and patients with Alzheimer's disease were analyzed. Subjects with cognitive impairment demonstrated more intra-individual performance variability and slower reaction times than cognitively healthy elderly. In a similar direction, Braverman et al. [9] showed that the test of variables of attention (TOVA) is an accurate predictor of early attention complaints and memory impairments in a clinical setting. The standard measures of attention disorders and memory function showed high correlations with a decreased TOVA performance.

Another application area of RT tests is the diagnosis of Attention Deficit Hyperactive Disorder (ADHD). Patients diagnosed with ADHD have in general difficulties in focusing on specific tasks. In a study of Vaurio et al. [56] children with and without ADHD were investigated using go/no-go tasks with differing levels of cognitive demand. The findings showed that one of the most consistent findings to characterize the children with ADHD was increased variability in RT.

The effect of psychosocial stress on RT was investigated in a controlled experimental setting by Scholz et al. [47]. Subjects were randomly assigned to the Trier Social Stress Test (TSST) versus a rest condition. After the stress test, each subject performed a go/no-go task. Subjects in the stress condition revealed slower reaction times compared to subjects in the rest condition.

Van den Berg examined the effect of a night's sleep loss on RT performance [55]. Subjects missed significantly more stimulus signals and had slower reaction times when sleep deprived compared to well rested subjects. The effects of mental fatigue on RT performance caused by prolonged time on monotonous simple RT task was investigated in [28]. Significant slowing of mean RT over time was observed.

Other factors which are known to influence reaction times are beside others gender, arousal level, personality type or intelligence level [22, 26].

1.3.3. Wearable Reaction Time Tests

Empirical studies which aim to explain the relationships between RT and independent variables have been mostly conducted in carefully controlled laboratory settings. Most of the studies have in common that the employed RT tests are operated with a traditional desktop-based test, which requires the full attention of the subject for several minutes. Since the user has to interrupt his daily activities to perform the computerized tests, most of these techniques are not feasible to be used during everyday life activities. There exist only a few applications which investigate the feasibility of measuring reaction times continuously during everyday activities. Lieberman et al. developed a wrist-worn watch like device which assess reaction time, vigilance and memory. The device consists of visual stimuli (3 LEDs), auditory stimuli (a miniature speaker) and two push buttons [29]. The reaction time is assessed by presenting a series of auditory stimuli at random times and measuring the response time until a button is pressed. In the second generation of the device the authors integrated an additional vibratory stimuli which can be used in noisy environments where auditory stimulus might be masked. Ivorra et al. [21] implemented a haptic stimulus into a watch-like device in order to interrogate the subject. Subject's response was detected by means of recognizing a subtle hand movements using an accelerometer. The authors conducted a feasibility study with ten subjects. The subjects used the device on their normal activities. The total number of RT interrogations for a 8 h period was 33. During the study the subjects were asked to fill out a questionnaire about the usability of the device. However, the obtained reaction times were not further analyzed and relationships between reaction time and influencing factors such as age, gender, cognitive demanding tasks or performed activities were not conducted.

1.4. Aims of the work

The aim of this thesis is to investigate mobile and wearable technologies for continuous monitoring of cognitive load and cognitive performance in daily life scenarios. In the first part of the thesis, physiological sensing is investigated to assess different levels of mental load occurring in everyday life scenarios. In the second part, a wearable reaction time test device is presented which enables the measurement of reaction times during daily life activities. The thesis then investigates the feasibility of the device for detecting variations in reaction times occurred due to influencing factors such as cognitive load, age and daily activities. Specifically, the thesis addresses the following topics:

1.4.1. Monitoring of mental load levels using HRV features

Today's mobile healthcare applications offer a variety of opportunities to capture physiological signals, process the data in real-time and provide just-in-time feedback. To achieve a day-by-day quantification of mental load for a long-term health monitoring, first different load levels which occur during everyday have to be discriminated. This thesis investigates the applicability of HRV features obtained from a mobile ECG system. In contrast to state of the art which generally discriminate between baseline (rest) and high cognitive load, the aim of this thesis is to show that HRV features obtained from a mobile ECG logger allows to discriminate different levels of mental workload induced by a continuous performance task.

1.4.2. Assessment of mental load levels in everyday life using individual calibration measures

Existing studies on mental workload mostly focus on laboratory settings where the sympathetic and parasympathetic responses to stress or high cognitive load are collected from several subjects within a standardized, controlled experiment. The collected data are then used to build generalized models to predict stress or mental workload scores. However, such models are usually not feasible in real life settings since controlled experiments often do not reflect the real life setting. Additionally, high variations in HRV responses may exist between subjects because of factors such as age, health status or physical activity level. As a consequence, two open questions are: (1) How to approach different levels of mental load occurring during real life settings? (2) How to incorporate individual differences when building models for cognitive load estimation? In order to target individual differences, this thesis considers a calibration procedure where each subject's individual HRV responses to different levels of mental workload are measured in a laboratory setting. This thesis then investigates whether the data obtained in the calibration session are appropriate to discriminate low, medium and high mental workload levels occurred during a daily life office-work scenario.

1.4.3. Development of a wearable user interface for measuring reaction time

Conducting cognitive assessment tests throughout normal daily life offers new opportunities to early detect changes in cognitive efficiency. Such tests would allow identification of early symptoms of cognitive impairment, monitor the progress of disease processes related to cognitive efficiency and reduce the risk of cognitive overload. The goal of this thesis is to transfer a well-defined cognitive test into daily life in order to obtain a continuous measurement of cognitive performance. For the cognitive test, RT tests are considered in this thesis, since RT tests offer high sensitivity for detecting variation in cognitive functioning and they can be repeated virtually an unlimited amount of times. A shortcoming of traditional computerized RT tests is that they require the full attention of a test person, which prohibits the measurement of cognitive efficiency during daily routine tasks. In order to overcome this limitation, this thesis aims to design a wearable reaction time test which can be operated throughout everyday life. The following requirements associated with the design of the wearable device are addressed: (1) wearability of the device (2) choosing a stimulus type that is convenient for daily life (3) choosing a predefined response gesture (4) ensuring a continuous operation for long-term recordings.

1.4.4. Feasibility evaluation of the wearable reaction time test

This thesis aims to evaluate the feasibility of the wearable RT device with two empirical studies. The first study investigates to what extent changes in duration and variability of user's reaction time can be measured with the wearable RT interface in comparison to traditional desktop-based reaction time tests. The aim of this study is to determine whether the wearable RT test is sensitive to measure changes in reaction times occurred due to altered cognition. The second experiment aims to conduct a long-term case study in a real working environment. The case study targets the following goals: (1) to continuously collect reaction times in a real-world working setting without interrupting the daily routine of the subject (2) to investigate the observed variations in length and variability of reaction times regarding typical work-related workload factors (3) to investigate the correlations between RT features and subjective ratings on mood and perceived workload.

1.4.5. Effect of daily activities on reaction times

Most studies which employ RT tests are examined in carefully controlled laboratories. During these studies the independent variable is changed while other factors are kept constant. However, in daily life the activities performed by the subjects can also affect the reaction times. This thesis investigates how common daily activities affect the reaction times of young and elderly subjects. The main research aims of the thesis are: (1) to examine the effect of everyday life activities on reaction times, (2) to determine if it is feasible to measure variations in reaction times occurred due to high cognitive load while the subjects are engaged in common daily life tasks, (3) to investigate the differences between young and elderly age group regarding their reaction times.

1.5. Thesis outline

This thesis includes six scientific publications (Chapter 3 to 8) addressing the aims summarized in the previous section. Figure 1.1 depicts the thesis's aims and the corresponding chapters addressing them. The arrows in the figure show the relationship between the chapters. Table 1.1 shows the assignment of each publication to its respective chapter.

Chapter 2 presents the summary of contributions while providing an insight into the relevant outcomes. The limitations and new research directions for future work are highlighted. Chapter 3 presents a laboratory study which investigates the use of heart rate variability to discriminate three levels of mental load. Chapter 4 extends the previous study by incorporating individual calibration measures. Each subject's heart rate response obtained during the laboratory experiment is used as calibration measure. This information is then used to differentiate different mental load levels occurring in an office work scenario. Chapter 5 presents a new sensing modality, the so-called wearable reaction time test which allows monitoring one's cognitive efficiency during daily activities. Furthermore, the first results are presented in which the reaction times obtained by the wearable system are compared to those collected with a well accepted desktop-based tests. Chapter 6 extends the previous work and reports results of a detailed statistical analysis. In addition to the comparison between desktop-based and wearable reaction time tests, this chapter presents a comparison between two implementations of the wearable reaction time test. Chapter 7 presents a case study conducted in a real working environment and investigates the link between different workload factors and reaction times. Finally, combining the findings achieved in Chapter 6 and 7, Chapter 8 presents detailed results from a study on examining the effects of daily activities on reaction times. Besides, this chapter highlights the differences between young and elderly age groups by comparing how their reaction times change when they are subjected to high cognitive load while performing different daily life activities.

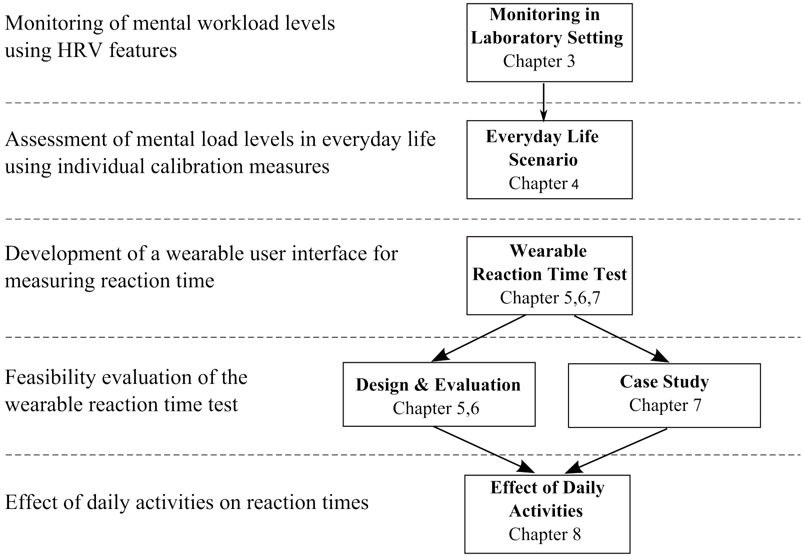
Monitoring of mental workload levels using HRV features
Monitoring in Laboratory Setting
Chapter 3
Assessment of mental load levels in everyday life using individual calibration measures
Everyday Life Scenario
Chapter 4
Development of a wearable user interface for measuring reaction time
Wearable Reaction Time Test
Chapter 5,6,7
Feasibility evaluation of the wearable reaction time test
Design & Evaluation
Chapter 5,6
Case Study
Chapter 7
Effect of daily activities on reaction times
Effect of Daily Activities
Chapter 8

Figure 1.1. Visualization of the thesis outline showing each chapter according to the aims presented in Section 1.4.

Table 1.1. Publications and corresponding chapters included in this thesis.

Chapter	Publication
3	Monitoring of Mental Workload Levels Burcu Cinaz, Roberto La Marca, Bert Arnrich and Gerhard Tröster Proceedings of IADIS eHealth Conference, 189-193, July 2010.
4	Monitoring of Mental Workload Levels during an Everyday Life Office-Work Scenario Burcu Cinaz, Bert Arnrich, Roberto La Marca and Gerhard Tröster Personal and Ubiquitous Computing, 17(2), pp 229-239, Springer, February 2013.
5	A Wearable User Interface for Measuring Reaction Time Burcu Cinaz, Christian Vogt, Bert Arnrich and Gerhard Tröster International Joint Conference on Ambient Intelligence, pp 41-50, Springer, 2011.
6	Implementation and Evaluation of Wearable Reaction Time Tests Burcu Cinaz, Christian Vogt, Bert Arnrich and Gerhard Tröster Pervasive and Mobile Computing, 8(6), pp 813 - 821, Elsevier, December 2012.
7	A Case Study on Monitoring Reaction Times with a Wearable User Interface during Daily Life Burcu Cinaz, Bert Arnrich, Roberto La Marca, Gerhard Tröster International Journal of Computers in Healthcare, Inderscience, 1(4), pp 283-303, 2012.
8	Effects of Daily Activities on Reaction Times: Comparison between Young and Elderly Subjects Burcu Cinaz, Bert Arnrich, Nathan Theill, Vera Schumacher, Mike Martin and Gerhard Tröster Pervasive and Mobile Computing, 2013 (submitted)

1.6. Additional publications

The following publications have been written in addition to those presented in this thesis:

- Burcu Cinaz, Roberto La Marca, Bert Arnrich, Gerhard Tröster, “Towards continuous monitoring of mental workload”, 5th International Workshop on Ubiquitous Health and Wellness. 2010.
- Burcu Cinaz, Bert Arnrich, Gerhard Tröster, “Monitoring of Cognitive Functioning by Measuring Reaction Times with Wearable Devices”, 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), Workshop on Pervasive Care for People with Dementia and their Carers, 2011.
- Mirco Rossi, Burcu Cinaz and Gerhard Tröster, “Ready-To-Live: Wearable Computing Meets Fashion”, Adjunct Proceedings of the 13th International Conference on Ubiquitous Computing (Ubicomp 2011), 2011.

Bibliography

[1] http://www.daimler.com.

[2] http://www.emercedesbenz.com/.

[3] J. Aasman, G. Mulder, and L. Mulder. Operator effort and the measurement of heart-rate variability. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(2):161–170, 1987.

[4] R. A. Abrams and D. A. Balota. Mental chronometry: Beyond reaction time. *Psychological Science*, 2(3):153–157, May 1991.

[5] R. H. Baayen. Analyzing reaction times. *International Journal of Psychological Research*, 3:12–28, 2010.

[6] K. E. Beede and S. J. Kass. Engrossed in conversation: The impact of cell phones on simulated driving performance. *Accident Analysis &amp; Prevention*, 38(2):415 – 421, 2006. ISSN 0001-4575.

[7] A. Benoit, L. Bonnaud, A. Caplier, P. Ngo, L. Lawson, D. G. Trevisan, V. Levacic, C. Mancas, and G. Chanel. Multimodal focus attention and stress detection and feedback in an augmented driver simulator. *Personal Ubiquitous Comput.*, 13(1):33–41, 2009. ISSN 1617-4909.

[8] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 78(Supplement 1):B231–B244, May 2007.

[9] E. R. Braverman, A. L. Chen, T. J. Chen, J. D. Schoolfield, A. Notaro, D. Braverman, M. Kerner, S. H. Blum, V. Arcuri, M. Varshavskiy, U. Damle, B. W. Downs, R. L. Waite, M. Oscar-Berman, J. Giordano, and K. Blum. Test of variables of attention (TOVA) as a predictor of early attention complaints, an antecedent to dementia. *Neuropsychiatr Dis Treat*, 6(1):681–690, 2010.

[10] G. E. Cooper and R. P. Harper. The use of pilot rating in the evaluation of aircraft handling qualities. Technical Report 567, AGARD, London, Apr. 1969.

[11] F. C. Donders. On the speed of mental processes. *Acta Psychol (Amst)*, 30:412–431, 1969.

[12] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *Intelligent Transportation Systems, IEEE Transactions on*, 12(2):596–614, 2011.

[13] J. Engström, E. Johansson, and J. Östlund. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8 (2):97 – 120, 2005. ISSN 1369-8478.

[14] European Foundation for the Improvement of Living and Working Conditions. Work-related stress. http://www.eurofound.europa.eu/.

[15] E. Gorus, R. De Raedt, M. Lambert, J. C. Lemper, and T. Mets. Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. *J Geriatr Psychiatry Neurol*, 21(3):204–218, Sep 2008.

[16] E. Gorus, R. De Raedt, and T. Mets. Diversity, dispersion and inconsistency of reaction time measures: Effects of age and task complexity. *Aging Clin Exp Res*, 18(5):407–417, Oct 2006.

[17] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[18] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris. Out of the lab and into the fray: Towards modeling emotion in everyday life. In *Pervasive Computing*, vol. 6030/2010, pp. 156–173, 2010.

[19] A. Henelius, K. Hirvonen, A. Holm, J. Korpela, and K. Muller. Mental workload classification using heart rate metrics. *Conf Proc IEEE Eng Med Biol Soc*, 1:1836–1839, 2009.

[20] D. F. Hultsch, S. W. MacDonald, and R. A. Dixon. Variability in reaction time performance of younger and older adults. *J Gerontol B Psychol Sci Soc Sci*, 57(2):P101–115, Mar 2002.

[21] A. Ivorra, C. Daniels, and B. Rubinsky. Minimally obtrusive wearable device for continuous interactive cognitive and neurological assessment. *Physiol Meas*, 29(5):543–554, May 2008.

[22] A. R. Jensen. *Clocking the mind: Mental chronometry and individual differences.* Elsevier, 2006.

[23] D. Kim, Y. Seo, , and L. Salahuddin. Decreased long term variations of heart rate variability in subjects with higher self reporting stress scores. In *Pervasive Healthcare*, 2008.

[24] D. Kim, Y. Seo, J. Cho, and C. H. Cho. Detection of subjects with higher self-reporting stress scores using heart rate variability patterns during the day. *Conf Proc IEEE Eng Med Biol Soc*, 2008: 682–685, 2008.

[25] J. Klingner, R. Kumar, and P. Hanrahan. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of ETRA*, pp. 69–72, 2008.

[26] R. J. Kosinski. A literature review on reaction time, August 2009.

[27] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple Task Performance*, pp. 279–328, 1991.

[28] R. Langner, M. B. Steinborn, A. Chatterjee, W. Sturm, and K. Willmes. Mental fatigue and temporal preparation in simple reaction-time performance. *Acta Psychol (Amst)*, 133(1):64–72, Jan 2010.

[29] H. R. Lieberman, F. M. Kramer, S. J. Montain, and P. Niro. Field assessment and enhancement of cognitive performance: Development of an ambulatory vigilance monitor. *Aviation, Space, and Environmental Medicine*, 78(5, Suppl.):B268–75, May 2007.

[30] R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford University Press, 1986.

[31] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz. 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, Mar 1996.

[32] M. Morris and F. Guilak. Mobile heart health: Project highlight. *IEEE Pervasive Computing*, 8(2):57–61, 2009. ISSN 1536-1268.

[33] L. Mulder. Cardiovascular reactivity and mental workload. *International Journal of Psychophysiology*, 7(2):321–322, 1989.

[34] F. Paas and J. Merriënboer. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6:351–371, 1994. ISSN 1040-726X.

[35] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1):63–71, 2003.

[36] O. Palinko, A. Kun, A. Shyrokov, and P. Heeman. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pp. 141–144. ACM, 2010.

[37] C. Papadelis, C. Kourtidou-Papadeli, P. Bamidis, and M. Albani. Effects of imagery training on cognitive performance and use of physiological measures as an assessment tool of mental effort. *Brain and Cognition*, 64(1):74 – 85, 2007. ISSN 0278-2626.

[38] C. J. Patten, A. Kircher, J. Östlund, L. Nilsson, and O. Svenson. Driver experience and cognitive workload in different traffic environments. *Accident Analysis &amp; Prevention*, 38(5):887 – 894, 2006. ISSN 0001-4575.

[39] G. B. Reid and T. E. Nygren. The Subjective Workload Assessment Technique: A scaling procedure for measuring mental workload. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 8, pp. 185–218. Elsevier, 1988.

[40] B. Reimer and B. Mehler. The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics*, 54(10):932–942, 2011.

[41] A. Riener. Simulating on-the-road behavior using a driving simulator. In *Advances in Computer-Human Interactions, 2010. ACHI'10. Third International Conference on*, pp. 25–31. IEEE, 2010.

[42] A. Riener, A. Ferscha, and M. Aly. 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*, pp. 99–106, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-571-0.

[43] S. Rubio, E. Díaz, J. Martín, and J. M. Puente. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004. ISSN 1464-0597.

[44] D. Salvucci. Predicting the effects of in-car interface use on driver performance: An integrated model approach. *International Journal of Human-Computer Studies*, 55(1):85–107, 2001.

[45] D. Salvucci. Modeling driver behavior in a cognitive architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(2):362–380, 2006.

[46] D. D. Salvucci and K. L. Macuga. Predicting the effects of cellular-phone dialing on driver performance. *Cognitive Systems Research*, 3(1):95 – 102, 2002. ISSN 1389-0417. Computational Cognitive Modeling.

[47] U. Scholz, R. L. Marca, U. M. Nater, I. Aberle, U. Ehlert, R. Hornung, M. Martin, and M. Kliegel. Go no-go performance under psychosocial stress: Beneficial effects of implementation intentions. *Neurobiology of Learning and Memory*, 91(1):89 – 92, 2009. ISSN 1074-7427.

[48] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine: Personal Healh Systems*, 2010.

[49] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 extended*

abstracts on Human factors in computing systems, pp. 2651–2656. ACM, 2007.

[50] C. Soga, S. Miyake, and C. Wada. Recovery patterns in the physiological responses of the autonomic nervous system induced by mental workload. In *SICE, 2007 Annual Conference*, pp. 1366–1371, Sept. 2007.

[51] S. Sternberg. Memory-scanning: mental processes revealed by reaction-time experiments. *Am. Sci.*, 57(4):421–457, 1969.

[52] J. Taelman, S. Vandeput, E. Vlemincx, A. Spaepen, and S. Huffel. Instantaneous changes in heart rate regulation due to mental load in simulated office work. *European Journal of Applied Physiology*, 111:1497–1505, 2011. ISSN 1439-6319.

[53] A. Tattersall and G. Hockey. Level of operator control and changes in heart rate variability during simulated flight maintenance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(4):682–698, 1995.

[54] G. van Daalen, T. Willemsen, K. Sanders, and M. van Veldhoven. 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, 2009.

[55] J. van den Berg and G. Neely. Performance on a simple reaction time task while sleep deprived. *Percept Mot Skills*, 102(2):589–599, Apr 2006.

[56] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky. Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands. *Neuropsychologia*, 47(12):2389 – 2396, 2009. ISSN 0028-3932.

[57] T. W. Victor, J. L. Harbluk, and J. A. Engström. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):167 – 190, 2005. ISSN 1369-8478.

[58] J. Westerink, E. L. Broek, M. H. Schut, J. Herk, and K. Tuinenbreijer. Computing emotion awareness through galvanic skin response and facial electromyography. In J. Westerink, M. Ouwerkerk, T. Overbeek, W. Pasveer, and B. Ruyter, ed., *Probing Experience*, vol. 8 of *Philips Research*, pp. 149–162. Springer Netherlands, 2008. ISBN 978-1-4020-6592-7.

[59] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[60] F. H. Wilhelm and P. Grossman. Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biol Psychol*, Feb 2010.

[61] G. F. Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *Int. J. Aviat. Psychol*, 12:3–18, 2002.

[62] K. Younsi, J.-M. Girard, J.-C. Popieul, P. Loslever, and P. Simon. Improving safety through online driver workload assessment. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pp. 1107–1112, Oct. 2008.

2

Thesis Summary

2.1. Summary of Contributions

The key research contributions of this thesis which advance the state-of-the-art in monitoring cognitive load and cognitive performance are presented below. The summary is structured according to the thesis aims introduced in Section 1.4 and depicted in Figure 1.1. Detailed descriptions on methods and results can be found in the corresponding publication chapters referenced in this summary chapter (see Table 1.1).

2.1.1. Monitoring of mental load levels using HRV features

This work investigated how heart rate variability features originated from a mobile ECG system could be used to discriminate three levels of mental load. In order to investigate this relationship, first an experiment was designed to induce three levels of mental load on subjects. Afterwards, the changes in heart rate variability for each load level were analyzed. Regarding the first objective (see Section 1.4.1, page 12) the following contributions were achieved.

Experiment design

In order to simulate different mental load levels, an experiment in a laboratory setting was designed. Seven healthy male subjects (age between 25 and 34 years) were involved in the experiment. Three sessions with low, medium and high workload were defined. Each of the mental load sessions had the same structure composed of a baseline (10 min), workload (20 min) and recovery (15 min) period. Subjects performed each session on separate days. In order to counterbalance learning effects, each workload session was randomly assigned to the subjects. Three difficulty levels of a continuous performance task, the so-called Dual N-Back task [1, 7] (see Section 3.2.1, page 62), were employed to induce low, medium and high workload. During all three sessions, the baseline and the recovery periods were the same: the subjects watched a relaxing documentary film in order to calm down. The workload phases differed in the amount of mental workload induced by the N-Back task. Directly after the workload period, the subject was asked to assess his perceived workload by filling out the the NASA-TLX [5]. The recording of the ECG data was accomplished with a Zephyr BioHarness chest belt (see Figure 4.1, page 73). In addition to the ECG data, the chest belt pro-

Table 2.1. Comparison of HRV Features during low, medium and high workload periods. F and p values from ANOVA test are given (from Section 3.3, page 65).

HRV Features	Low Workload	Medium Workload	High Workload	$F; p$
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*	1.9 ± 0.2	2.5 ± 0.3	4.6 ± 1.0	$4.59; 0.02$

Mean ± standard error
$*p < 0.05$

vided RR intervals by measuring the duration between two consecutive R waves of the ECG.

HRV analysis

The following HRV features (see Section 3.2.2, page 63) in the time domain were investigated: 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 (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).

The analysis of HRV in the frequency domain investigated two frequency bands: low frequency (LF: 0.04-0.15 Hz) and high frequency (HF: 0.15-0.4 Hz). Next, the ratio of LF/HF was calculated by using the normalized values of LF and HF.

All features were calculated on the overall workload periods (20 min) for each session. The features obtained for the three workload periods were then analyzed by using the ANOVA test.

Results

- The analysis of the perceived workload scores obtained from the NASA-TLX showed that each subject perceived the induced load levels as intended from the experiment design ($p < 0.01$)(see Section 3.3, Figure 3.1, page 64).

- The investigated HRV features could be classified into two distinct groups with respect to their response: with increasing workload, the features RMSSD, pNN50 and HF showed a statistically significant decrease while LF and LF/HF ratio showed a statistically significant increase with increased workload ($p < 0.05$) (see Table 2.1). The remaining features showed a consistent but non-significant increase or decrease, what might be explained by the limited number of subjects.

- These findings provide evidence that an increase in workload is associated with a decrease in parasympathetic nervous activity and with a concomitant increase in sympathetic activity. The results are in line with other studies which showed that in comparison to a resting state, high workload or stress is associated with a decrease in parasympathetic nervous activity and with an increase in sympathetic activity [9–11, 15, 18].

2.1.2. Assessment of mental load levels in everyday life using individual calibration measures

Each individual's physiological response to high mental load can be different depending on several factors such as health status, activity level, age or medications. This work addresses the challenge of individual differences by incorporating individual calibration measures. The main research question of this work is how different mental load levels occurring during a daily life scenario can be discriminated by incorporating individual calibration measures. Regarding the second objective (see Section 1.4.2, page 12) the following contributions were done.

Calibration sessions

The data collected in the laboratory experiment described in the previous section (see Section 2.1.1, page 28) were used as individual calibration measurements. Seven healthy male participants (age between 25 and 34 years) participated in the overall experiment. Each individual's

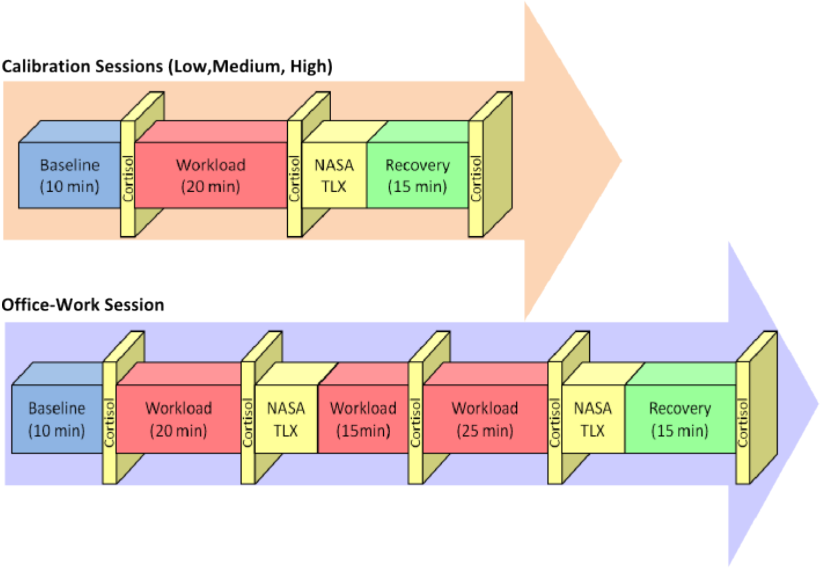

Figure 2.1. Experiment procedure for calibration and office-work sessions. A total of three calibration sessions were conducted which differed in the level of induced workload: low, medium, and high. The office-work condition consisted of 1 h of normal office working activities. The subjective rating of perceived workload was assessed with the NASA-TLX, whereas an objective measurement was assessed by collecting salivary cortisol at particular points in time (from Section 4.2.2, page 74).

HRV responses to three different levels of mental load were measured in a controlled laboratory setting. The first three sessions were designed to induce low, medium and high levels of mental workload in order to establish an individual calibration. This was enabled by applying three variants of Dual N-Back task (see Section 4.2.2, page 75).

Daily life office-work session

The laboratory experiment was extended with a real life working scenario. After obtaining each subject's baseline and workload heart rate features in the calibration sessions, an additional session was conducted in which the same seven subjects were monitored during one hour of normal office-work (see Section 4.2.2, page 76). The daily office tasks were freely chosen by the subjects and contained working activities such as programming, reading or writing research papers. Both calibration and office-work session included baseline and recovery stages (see Figure 2.1 for the experimental procedure). In addition to the heart rate data, subjective ratings of the perceived workload was collected with the NASA-TLX [5].

Data analysis

First, time and frequency features from the RR interval data were extracted (see Section 4.3.1, page 77). 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 (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).

As the frequency domain feature, the normalized value of the LF/HF ratio was calculated. For each subject, so called "relative features" were computed by dividing the HRV features obtained during the workload stages by the corresponding mean value of the baseline feature.

Calibration procedure

The goal of incorporating the individual calibration measures was to investigate whether the data collected during the calibration sessions were appropriate to discriminate low, medium and high mental load levels experienced during one hour of office-work. First, in order to assign the subjective workload score of the office-work into one of three classes (low, medium, and high), individual boundaries according to the subjective workload scores collected during the calibration conditions were defined (see Section 4.4.1, page 80 and Figure 4.7, page 82).

Afterwards, two approaches were developed that use the individual calibration data. For a given 2-min RR segment the objectives were (1) to predict the corresponding subjective workload score by using relevant HRV features and, (2) to identify the mental workload class

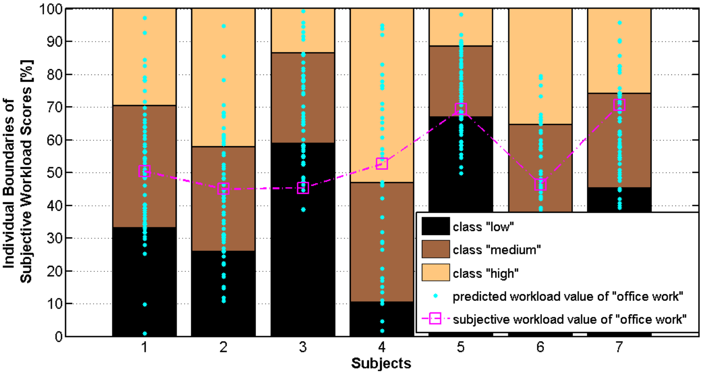

Figure 2.2. Predicted workload scores of the office-work session based on a linear regression model. The regression problem was transformed into a classification problem using the majority rule, i.e. the predicted class was computed as the class to which the majority of predicted values falls into (from Section 4.4.6, page 86).

(low, medium, or high) to which a new observation belongs. For the first objective, a multiple regression analysis was employed to model the relationship between HRV features and the subjective ratings of NASA-TLX. For each subject, the HRV features were used as the predictor variables while the NASA-TLX scores served as response variables. The NASA-TLX scores of the office-work session were then predicted based on this model (see Section 4.4.6, page 85 and Figure 2.2). For the second objective, the linear discriminant analysis (LDA), k-nearest neighbor algorithm (k-NN), and support vector machine (SVM) were used as classification methods.

For the multiple regression and all three classification methods, the entire calibration data were used as training set and the office-work data were used as test set.

Results

- According to the subjective NASA-TLX ratings collected in the calibration sessions, it was shown that all participants perceived the induced load levels as intended from the experiment design. As expected, the N-Back performance of the subjects collected

Table 2.2. Multiple linear regression office-work session: actual workload class obtained from NASA-TLX rating vs. estimated class. Note that the regression problem was transformed into a classification problem using the majority rule: the predicted class was computed as the class to which the majority of predicted values falls into (from Section 4.4.6, page 88).

Subjects	Actual class	Predicted class
1	Medium	Medium
2	Medium	Medium
3	**Low**	**Medium**
4	High	High
5	Medium	Medium
6	Medium	Medium
7	Medium	Medium

False identified classes are indicated in bold.

Table 2.3. Classification results for each subject (from Section 4.4.7, page 89).

Subjects	True Class	Predicted Class		
		LDA	k-NN	SVM
1	M	M	M	M
2	M	M	M	**L**
3	L	**M**	**M**	L
4	H	H	**L**	**M**
5	M	M	M	M
6	M	M	M	M
7	M	M	M	M

False identified classes are indicated in bold.
L (low), M (medium), H (high)

in the calibration sessions decreased with increasing workload (see Section 4.4.3, page 83). Compared to the calibration conditions, subjective NASA-TLX ratings of the office-work session were ranked either between low and medium or between medium and high (see Section 4.4.1, page 80 and see Figure 4.5, page 80).

- The resulting predictions of the workload class (low, medium, high) were correct for six out of the seven subjects when applying multiple linear regression (see Section 4.4.6, page 85 and see Table 2.2). In only one subject (subject 3), there was a confusion between low and medium workload.
- The comparison of the three classification models showed that the best results were obtained with linear discriminant analysis (LDA) which resulted in 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 (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 (see Section 4.4.7, page 86 and see Table 2.3).
- The overall results showed that the perceived mental workload level of an individual occurred during an office-work scenario could be discriminated by modeling the relationship between individual HRV features and the subjective ratings gathered in a calibration session. This provides evidence that incorporating calibration measures for everyday life scenarios as proposed by Morris et al. [13] is feasible.

2.1.3. Development of a wearable user interface for measuring reaction time

Reaction time is one of the most sensitive parameters for detecting variation in cognitive functional ability. There are several desktop-based reaction time tests in which users have to respond to visual stimuli by using keyboard, mouse or special buttons. The main shortcoming of existing computerized RT tests is that they require the full attention of the subject, i.e. the subject has to interrupt his daily routine for several minutes in order to perform the task on the computer. This restriction prohibits the measurement of reaction times during daily activities. This work investigated the design of a wearable reaction time test which

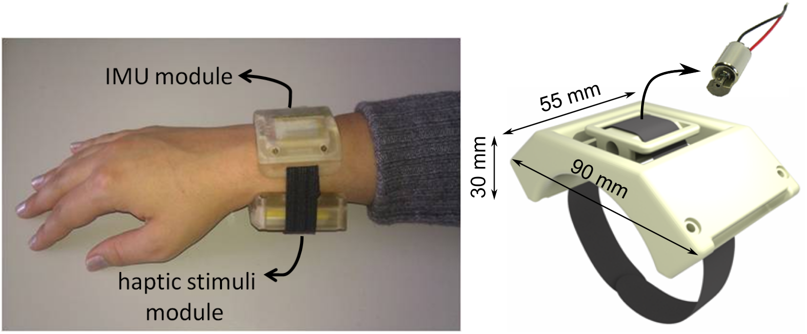

Figure 2.3. The left picture shows the wearable implementation of a simple reaction time test. The haptic stimuli are generated as vibrations on the lower side of the wrist. The IMU is placed on the opposite side to recognize the user's hand gesture responses. The right picture depicts a bracelet-like housing of the stimuli module. For the setups of the wearable reaction time test, see Section 5.3.1, page 100 and Section 6.3.1, page 119 and Section 7.3.1, page 143.

can be operated continuously throughout everyday life. Regarding the third objective (see Section 1.4.3, page 13) the following contributions were done.

System Design

The following requirements were considered for the design of a wearable reaction time test; (1) delivering stimulus during everyday life tasks, (2) choosing a predefined response gesture, (3) ensuring a continuous operation for long-term monitoring.

Since visual and auditory channel is usually occupied during daily life activities, the traditional visual stimulus used in desktop-based tests was replaced with haptic stimulus. The keyboard response was replaced with a forearm rotation allowing the user to perform the test in a natural way without pressing any extra push button. The haptic stimulus was applied on the wrist since it is known that the perception is increased particularly in the near of anatomical anchor points such as the wrist.

The final design of the wearable RT test (see Section 5.3.1, page 99 and Section 6.3.1, page 118) consisted of two main modules: the stimuli

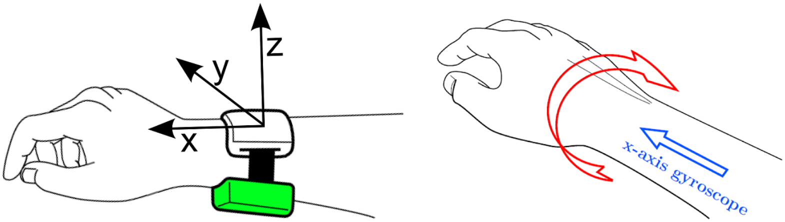

Figure 2.4. Illustration of the three-dimensional gyroscope axes of the IMU module and the gesture response defined as an outward rotation of the forearm along the x-axis of the gyroscope (from Section 6.3, page 121).

module to generate haptic stimuli and the inertial measurement unit (IMU) module for detecting forearm rotations (see Figure 2.3). In order to generate vibro-tactile stimuli, a coreless mini DC vibration motor with a diameter of 6 mm and a resonant frequency of 200 Hz was used. The motor was placed in a separate plastic enclosure with dimensions of 90 x 55 x 30 mm to maximize the sense of the vibration (see Figure 2.3). The stimuli module had its own battery supply to ensure a continuous operation during at least one day. The IMU module consisted of the so-called ETH Orientation Sensor (ETHOS) [4], in which a three-axis accelerometer and a three-axis gyroscope are integrated. The IMU module was used to recognize the subject's gesture response. The response gesture was defined as an outward rotation of the forearm (see Figure 2.4). The accelerometer and gyroscope were sampled with a frequency of 128 Hz. The firmware version of the ETHOS was modified to control the vibro-tactile component.

Validation of the device

Before applying the device in real experiments, a preliminary study (see Section 5.3.1, page 99 and Section 6.3.1, page 118) was conducted to investigate whether the predefined response was detectable at different arm positions. Three subjects performed a simple reaction time test during three different conditions. In the first condition, the subject was sitting on a chair while the arms were heading towards the floor. In the second condition, the arms were placed on the table. In the third condition, the subject was walking with a moderate speed (4 km/h) on a treadmill. In each condition, the haptic stimuli were generated randomly. As soon as the subject perceived the target vibration stimulus he had to rotate his forearm outwards. Based on a visual inspection of the recorded data, it was clearly visible that the wrist-turn axis of the gyroscope (x-axis) was the most sensitive axis for detecting the rotation of the forearm (see Figure 2.4). The reaction time was defined as the time difference between the haptic stimulus and the occurrence of the forearm rotation.

2.1.4. Feasibility evaluation of the wearable reaction time test

This work investigated the feasibility of the wearable reaction time test with two empirical studies. The first study investigated to what extent the changes in the duration and variability of user's reaction time could be measured with the wearable device in comparison to traditional desktop-based reaction time tests. In the second study a long-term monitoring in a real working environment was conducted. Regarding the fourth objective (see Section 1.4.4, page 14) the following contributions were done.

Evaluation in laboratory - Study Design

The main goal was to evaluate the wearable user interface by comparing the reaction times obtained by the wearable reaction time tests with those collected by a desktop-based reaction time test. In order to compare changes in the duration and variability of reaction times between the two measurement techniques, additional cognitive load was applied to the subjects in both setups. The observed changes in reaction times occurred due to the cognitive load were then compared between desktop-based and wearable reaction time tests.

As the desktop-based reaction time test, a free version of the go/no-go TOVA test was used (see Section 6.3.3, page 121). Traditional TOVA test consists of one target and one non-target stimuli type. The participant must respond only when the target stimulus appears and must not respond if the non-target stimulus appears (see Figure 2.5).

In this experiment, two different wearable implementations were evaluated. In the first design, the target haptic stimulus was placed on the lower side of the dominant wrist whereas the non-target haptic stimulus was placed on the upper side (one-handed design). In the second two-handed design, target and non-target modules were placed

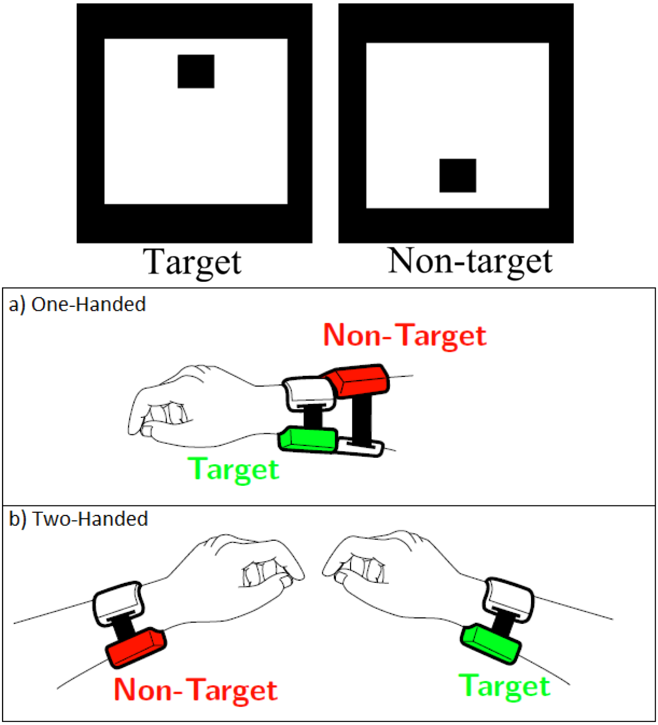

Figure 2.5. The above picture shows the target and non-target stimulus types of the TOVA test (from Section 6.3.3, page 122). In the picture below, two implementations of the wearable TOVA reaction time test are depicted. In the first one-handed design, target and non-target stimulus modules are both placed on the same wrist. In the second two-handed design, the target and non-target stimulus modules are placed on separate wrists (from Section 6.3.1, page 121).

separately on both wrists, i.e. the target stimulus module was placed on the lower side of the dominant wrist while the non-target stimulus was placed on the lower side of the non-dominant wrist (see Figure 2.5).

Twenty subjects (12 male, 8 female, average age 24.3 years) were recruited for the experiment. Each session lasted approximately 70 min. Participants were randomly assigned to one of the two experimental groups, which differs in sensor placement, i.e. 10 subjects used the one-handed design and the remaining 10 subjects used the two-handed de-

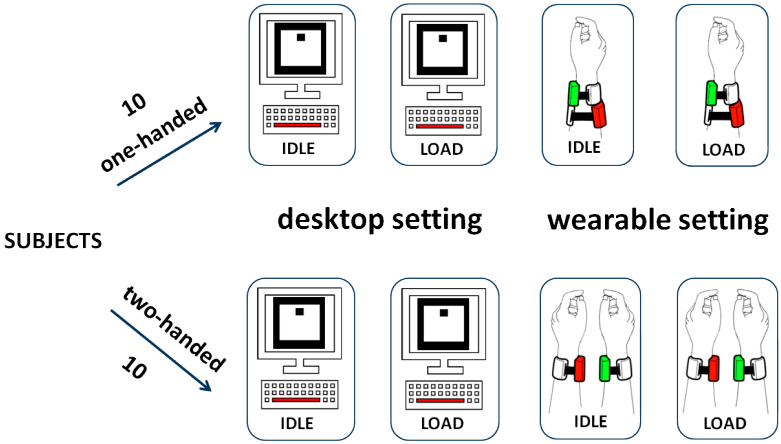

Figure 2.6. Experimental procedure: subjects were randomly assigned in one of the two experimental groups i.e. one-handed vs. two-handed groups. Each subject performed idle and load conditions for desktop-based and wearable reaction time test separately (from Section 6.3.4, page 123).

sign. In the first part of the experiment, participants performed the desktop-based TOVA test both in idle and cognitive load conditions. During the idle condition, the subjects had to respond to each visual target stimulus by using the keyboard. In the cognitive load condition, the subjects had to solve two tasks simultaneously: in addition to the TOVA test, the so-called Audio 2-Back task, a variant of the N-Back task [1, 7] was presented to the subject (Section 6.3.4, page 122). In the Audio 2-Back task every 3 sec a letter was pronounced to the subject via an audio message. The subject had to respond by saying "match" if the currently pronounced letter was the same as the one that was pronounced two positions back. In the second part of the experiment, subjects performed the same idle and cognitive load conditions with the wearable RT test. In the idle condition, the subjects had to perform only the wearable RT test. In the cognitive load condition, the wearable RT test and the Audio 2-Back task were performed simultaneously. Each condition lasted about 10 min and and a total of 640 reaction times were collected from each subject during the whole experiment (see Section 6.3.4, page 122 and for the experimental procedure see Figure 2.6).

Evaluation in laboratory - Data analysis

The hand response of the subject to a stimulus was detected by a threshold approach (see Section 6.3.1, Figure 6.2, page 120). The reaction time was then computed as the time difference between the onset of the haptic stimulus and the onset of the forearm rotation.

For the statistical analysis, mean reaction time and reaction time variability measured as the coefficient of variation (CV) (standard deviation divided by the mean reaction time) were used as evaluation metrics. Both within subjects and between subjects comparisons were done using ANOVA test (see Section 6.3.6, page 124). Within subject analyses addressed the questions (1) whether the reaction time performance and variability differ between desktop and wearable settings, and (2) whether the induced cognitive load differs from the idle state by means of reaction time performance and variability. Between subject comparison examined whether wearable reaction times differ between different sensor placements (one-handed vs. two-handed).

Evaluation in laboratory - Results

- The results (see Section 6.4, page 125) showed that during load condition, mean reaction time and variability were significantly increased ($p < 0.001$) for both settings (desktop vs. wearable) and both groups (one-handed vs. two-handed) (see Table 6.1, page 128). These first results provided evidence that the wearable reaction time test was suitable to measure changes in reaction times occurred due to the cognitive load. A similar cognitive load effect has also been demonstrated in a study of Brunken et al. [2]. The authors showed that the reaction times were significantly faster in a single-task condition (a visual reaction time test alone) when compared to a cognitive demanding dual-task condition (multimedia learning task in addition to the reaction time test).

- Compared to desktop test, the subjects demonstrated significant faster reaction times for both one-handed ($p < 0.05$) and two-handed ($p < 0.01$) wearable settings under the idle condition (see Table 6.1, page 128). This might be explained by the fact that the transduction of a visual stimulus takes generally longer than the perception of a haptic stimulus as known from literature [8]. However, in case of cognitive load, this effect tended to diminish and a significant difference between both settings was not observed (see Section 6.4, see Figure 6.6, page 126).

- The comparison between one-handed and two-handed wearable designs showed that the only difference between both groups was a significant faster mean reaction time for the two-handed design under idle condition ($p < 0.05$) (see Table 6.2, page 129). A reasonable explanation is that differentiating between target and non-target on the same hand is more difficult than differentiating of target and non-target on two hands.
- In conclusion, the results showed that both implementations of the wearable reaction time tests are suitable to measure factors that influence length and variability of reaction times.

Feasibility study in daily life - Study Design

This case study extended the monitoring of reaction times from a controlled laboratory setting to an unrestricted work environment (see Section 7.4, page 145). During the study, reaction time data and subjective ratings on mood were collected throughout 15 working days of a graphic designer. The first three working days were characterized by normal job demands as baseline measurement. During the remaining 12 days, the designer was confronted with four workload factors: stress, sleep deprivation, night shift, and moderate alcohol consumption (see Figure 2.7). Each factor was studied on three consecutive days (for a detailed description of each factor see Section 7.4.1, page 147).

Four aims were specified for this case study: (1) to continuously collect reaction times in a real-world working setting without interrupting the daily routine of a graphic designer, (2) to investigate the observed variations in length and variability of reaction times regarding the four workload factors, (3) to correlate the observed reaction time features with perceived workload ratings, and (4) to investigate the correlations between reaction time features and subjective ratings on mood.

Feasibility study in daily life - Data collection

The reaction time module was placed on the dominant wrist of the subject. A simple reaction time test was implemented which generated haptic stimuli at random intervals varying between 60 sec and 90 sec. Within 30 min periods a total of 20 reaction time measurements were collected.

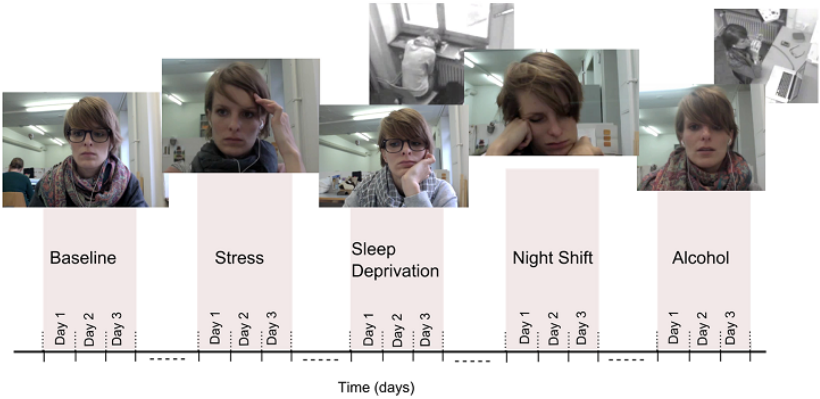

Figure 2.7. Case study with a graphic designer throughout 15 working days. The first three working days were characterized by normal job demands as baseline measurement. During the remaining 12 days, the designer was confronted with four workload factors: stress, sleep deprivation, night shift, and moderate alcohol consumption (from Section 7.4, page 146).

In order to examine the relationship between reaction time data and self-reported ratings on mood and workload, an experience sampling method (ESM) (see Section 7.4.7, page 149) was used. The subject was randomly probed eight times per day to complete a set of questionnaires. The minimum and maximum time interval between successive questionnaires was set to 30 and 120 min, respectively. The subject was notified to fill out the questionnaires by a particular haptic stimuli pattern, i.e. 10 sequential haptic stimuli were generated at 1 sec intervals (see Figure 7.4, page 149).

In addition, the subject was asked to assess her perceived workload regarding the last working activities using the NASA-TLX rating [5]. To recognize subject's hand response to a stimulus, the maximum peak occurring within a window of 2 sec after the stimulus was detected (see Figure 2.8). The relation between reaction times and self-rated mood and workload items was examined by creating data segments around each experience sampling events (see Section 7.5.1, page 151 and see Figure 7.6, page 152). The reaction time features were calculated for each segment and then correlated with the mood and perceived workload ratings.

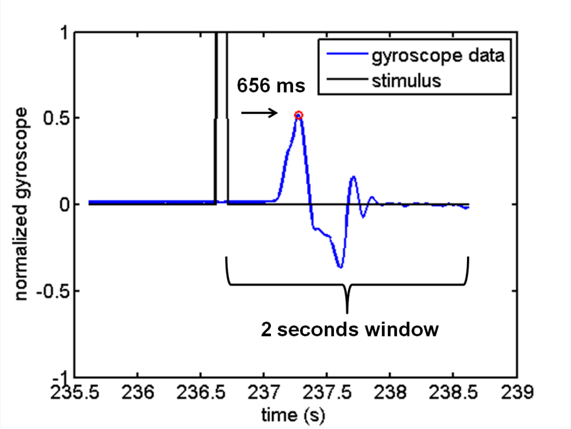

Figure 2.8. After each stimulus event the maximum peak within a fixed time window is detected. Reaction time is calculated as the time difference between the haptic stimulus and the occurrence of the peak response (from Section 7.5, page 151).

Feasibility study in daily life - Results

- The results (see Section 7.6, page 152) showed that the subject showed a significant increased variability in reaction times during the stress condition compared to the baseline ($p < 0.05$). Sleep deprivation did not show a significant difference to baseline. Compared to baseline condition, a significant increased variability of reaction times ($p < 0.05$) and a significant higher number of omission errors ($p < 0.05$) were observed for the night shift and alcohol consumption conditions (see Section 7.6.1, Figure 7.7, page 154 and Figure 7.8, page 156).

- The correlation analysis of reaction time measures with perceived workload (see Section 7.6.2, page 157) showed that mean reaction time was significantly positive correlated with mental and temporal demand ($p < 0.01$). Total workload showed a significant positive correlation with reaction time variability ($p < 0.01$) and number of omission errors ($p < 0.05$).

- The correlation analysis of reaction time measures with subjective ratings on mood (see Section 7.6.3, page 159) showed that mean reaction time was significantly positive correlated with the high arousal items concentrated ($p < 0.01$), alert ($p < 0.01$), active ($p < 0.01$) and stressed ($p < 0.01$). The reaction time variability was significantly positive correlated with alert ($p < 0.05$), active ($p < 0.05$), stressed ($p < 0.01$) and negative correlated with the low arousal mood items bored ($p < 0.01$) and tired ($p < 0.01$). In the valence-arousal plane (see Figure 7.10, page 159), it could be observed that a high level of arousal yielded to increased reaction times and higher variability whereas a low arousal yielded to lower variability of reaction times.

- In conclusion, this case study showed that a wearable watch-like reaction time test tool could be operated in a long-term measurement and enabled the interpretation of reaction times during work life activities. The results revealed that there was a correlation between workload factors and reaction time features. Besides, the ratings of the subject about her mood while performing working activities showed a correlation with the reaction time measures.

2.1.5. Effect of daily activities on reaction times

This work investigated how common daily activities affect the reaction times of young and elderly subjects. By knowing the effect of a particular daily activity on the reaction times, the variation in reaction times occurred due to the other factors can be extracted in a reasonable way. This can also enable a context-aware interpretation of the reaction times, i.e. the variations in reaction times can be analyzed within each activity class. Regarding the fifth objective (see Section 1.4.5, page 14) the following contributions were done.

Study design

The aim of the experiment was to investigate the effect of everyday activities on reaction times of young and elderly subjects. The experiment involved four common daily life activities: sitting, walking, reading and writing. In order to measure changes in reaction times as a result of an alteration in cognitive functionality, an additional cognitive load factor was added to each activity. Therefore, each activity was performed once under an “idle” condition in which subjects just performed the activity, and once under “load” condition in which subjects were confronted with additional cognitive load while performing the activity.

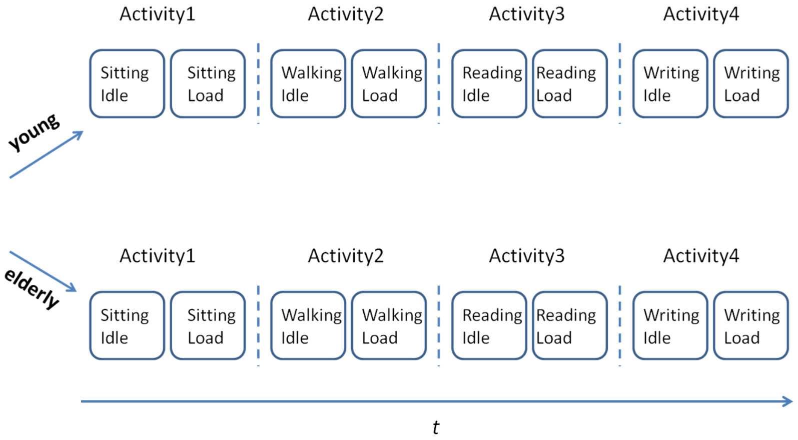

Figure 2.9. Experiment design: the four common daily life activities sitting, walking, reading and writing were investigated in both age groups. Each activity was performed once under an idle condition and once under a load condition. Each subject started with the sitting activity while the remaining daily activities were randomly assigned (from Section 8.3.4, page 175).

Having idle and load conditions enabled to compare the changes in reaction times between both young and elderly age groups. Additionally, it was investigated how accurate the idle and load conditions could be discriminated from each other (see Section 8.3.1, page 173).

The experiment addressed the following questions; (1) Do daily life activities affect the reaction time measures? (2) Do reaction times differ between young and elderly subjects for the chosen activities? (3) Do reaction times differ between idle and cognitive load conditions whilst performing daily life activities? (4) How accurate can idle and load conditions be discriminated from each other?

The data from 12 healthy elderly subjects with an average age of 70.17 years and 14 healthy young subjects with an average age of 25.79 were investigated. During the cognitive load condition, the subjects had to solve the Audio-2 Back task [1, 7] in addition to the main activity. The Audio 2-Back task was configured in a way that every 5 sec a letter was presented to the subject via an audio message. The subject was asked to respond only if the currently pronounced letter was the same as the one that was pronounced two positions back. The subject responded by saying “match” whenever a sound match was detected by the subject. Each subject started with the baseline activity which was defined as sitting on a chair at rest and just performing the wearable reaction time test. The remaining daily activities (walking, reading and writing) were randomly assigned to each subject. Each activity started with the idle condition and was followed by the cognitive load condition (see Section 8.3.4, page 175 and for the experimental procedure see Figure 2.9).

Data collection and analysis

The wearable reaction time module was placed on the non-dominant wrist of the subject. Simple reaction time test was implemented which generated haptic stimuli at random intervals varying between 5 sec and 25 sec. The subject was asked to respond as fast as possible to each haptic stimulus by performing the dedicated hand gesture, i.e. rotating his forearm outwards (see Section 8.3.2, Figure 8.1, page 173). Within each 10 min block an average of 40 reaction time measurements were collected. Directly after each block, each subject completed the NASA-TLX questionnaire [5] to indicate his perceived workload.

For the statistical analysis, mean reaction time and reaction time variability measured as the coefficient of variation (CV) (standard deviation divided by the mean reaction time) were used as evaluation metrics. In order to analyze the activity and cognitive load effects for both young and elderly subjects, a mixed 2x2x4 ANOVA was performed (see Section 8.3.7, page 178).

In order to investigate the discriminative power when distinguishing between idle and cognitive load conditions, logistic regression analysis was performed (see Section 8.3.8, page 178).

Results

- Mean reaction time and reaction time variability were significantly increased during cognitive load condition for both young and elderly subjects ($p < 0.001$) (see Section 8.4.1, page 179 and see Section 8.4.2, page 181). This confirmed that the cognitive load could be assessed during the daily activities under investigation.
- Mean reaction time and reaction time variability were significantly affected by the type of the performed activity ($p < 0.001$).

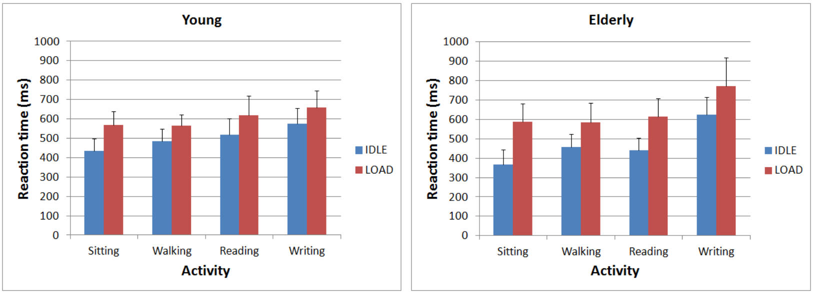

Figure 2.10. Mean reaction times of young and elderly subjects across four activities (from Section 8.4.1, page 179).

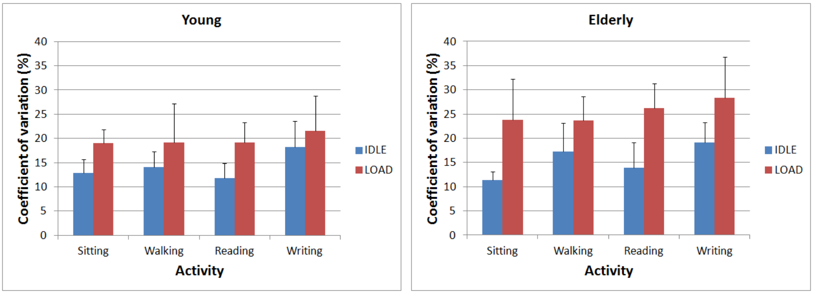

Figure 2.11. Reaction time variability of young and elderly subjects across four activities (from Section 8.4.2, page 181).

For both age groups, the highest mean reaction time and variability were observed for the writing activity (see Figure 2.10 and Figure 2.11).

- The test of between-subjects effect showed a significant main effect of age ($p < 0.01$) for the reaction time variability, reflecting that elderly subjects showed an overall higher reaction time variability compared to the younger participants (see Section 8.4.2, page 181). This result is consistent with other studies which have shown increasing variability with increasing age [3, 6].
- There was not a significant difference between young and elderly regarding to overall mean reaction times (see Section 8.4.1, page 179) although other studies showed that older adults were slower

than younger ones [3]. On the contrary, older participants showed a slightly faster mean reaction time during the idle condition (see Figure 2.10). One possible explanation could be the underlying motivational processes, e.g. it was observed that the elderly subjects were more motivated then the younger subjects during the whole experiment. Elderly perceived the accomplishment of the experiment as a sense of achievement whereas the main motivation for the younger subjects was the compensation money. The positive affective state of the elderly might have induced an increased arousal level which is a known factor that enhances the reaction time performance [8, 17]. However, the underlying motivational processes need to be investigated further.

- The main difference between young and elderly was observed in the interaction effect between condition and age group for mean reaction time ($p < 0.05$) and variability ($p < 0.01$). The increase in mean reaction time and reaction time variability from idle to cognitive load condition was higher for the older participants compared to the younger ones. In other words, although older participants showed slightly faster reaction times during idle condition, when they were confronted with high cognitive load their reaction time performance and variability were deteriorated to a larger extent compared to younger subjects.

- The classification of idle and load conditions confirmed the effect of activity on reaction times (see Section 8.4.3, page 183). In a leave-one-subject-out cross validation, the classification accuracy increased from 75% to 80.77% after adding activity class features to the classification model as predictors (see Figure 8.6, page 182). When considering a context-aware wearable reaction time test which is able to detect when the hand is active or not, a classification accuracy of 87.5% was achieved (see Section 8.4.3, page 183).

- The perceived workload scores obtained from the NASA-TLX questionnaire (see Section 8.4.4, page 184) showed that the elderly subjects reported in average lower perceived workload than younger subjects. This might be explained by the fact that elderly often tend to be reluctant to admit their cognitive difficulties because of a fear of losing their independence [14].

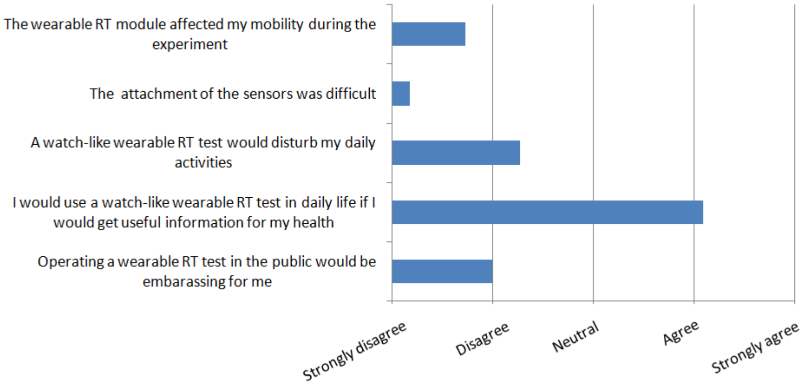

Figure 2.12. Usability questionnaire filled by the elderly subjects.

- The relationship between NASA-TLX workload scores and reaction time measures was modeled by applying a multiple linear regression analysis. Mean reaction time and reaction time variability served as predictors whereas the perceived workload score was used as the dependent variable. The results showed that the mean reaction time ($\beta = 0.085, p < 0.001$) and reaction time variability($\beta = 0.72, p < 0.05$) were significant predictors. The overall model fit was $R^2 = 0.282$ (see Section 8.4.4).

- Finally, the link between N-Back scores obtained from the Audio 2-Back task and reaction times was analyzed. Elderly subjects showed significantly lower performance than younger subjects in overall scores ($p < 0.01$). The lowest N-Back scores were observed during reading and writing activities since these activities were cognitively more demanding than sitting and walking activities.

- One of the main findings was that the increase in the length and variability of reaction times occurred due to the high cognitive load was higher for the elderly subjects compared to the younger age group. Besides, the type of daily activity affected the reaction times significantly. It was also shown how a simple activity recognition system integrated into the wearable reaction time test device would improve the detection of changes in reaction time performance and variability during common daily life activities.

- It should be noted that the elderly participants in this experiment were recruited from a dedicated university for elderly people. Cognitive activities such as education or similar intellectual engagement in older ages may postpone average age-related changes in cognition [16]. Therefore investigating elderly group with cognitive impairments could yield more deterioration in reaction times.

- Lastly, the usability questionnaire filled by elderly showed that operating the wearable reaction time test was not disturbing. Besides, most of the elderly subjects would accept using a wearable reaction time test in daily life when the device looks very similar to a watch (see Figure 2.12). Therefore, minimizing the size of the wearable reaction time module would meet the comfort requirements.

2.2. Conclusion

This thesis has investigated wearable sensing techniques towards assessing cognitive load and cognitive performance in daily life scenarios. In the first part of the thesis, cognitive load was assessed by means of HRV features obtained from a mobile ECG logger. In the second part of the thesis, a wearable watch-like reaction time test was introduced which enables a continuous monitoring of cognitive performance during daily life. As summarized in Section 2.1, the following conclusions can be drawn from the achieved contributions:

- Variants of a continuous performance task (Dual N-Back task) which differ in their degree of difficulty allow inducing different levels of mental workload. Data from a mobile ECG logger allow discriminating different levels of mental workload induced by N-Back tasks. The investigated HRV features can be classified into two distinct groups with respect to their response: with increasing workload the features RMSSD, pNN50 and HF showed a decrease in their values while LF and LF/HF ratio showed an increase with increased workload.

- Mental workload levels in an everyday life office-work scenario can be discriminated with data from a mobile ECG logger by incorporating individual calibration measures. Employing a calibration procedure allows modeling the relationship between relevant HRV features and the subjective ratings of perceived mental workload during an office-work scenario. The best results were obtained with LDA followed by k-NN and SVM.

- A wearable watch-like reaction time test was designed and implemented in a way that enables a continuous operation during daily life. In comparison to state-of-the-art reaction time tests, the visual stimulus was replaced by a haptic stimulus and instead of a keyboard response the detection of a forearm rotation was implemented.

- The wearable reaction time test was evaluated within two empirical studies. The first experiment was conducted to investigate to what extent the changes in length and variability of user's reaction time could be measured with the wearable interface in comparison to a well-accepted, traditional desktop-based test. The results showed that the variations in reaction times occurred due to an additional cognitive task were similar for both desktop and wearable settings which provides evidence that wearable reaction time tests are suitable to measure factors that influence length and variability of reaction times.

The second study investigated the feasibility of employing the wearable reaction time test in an unrestricted work environment. In a case study, reaction times of a graphic designer were collected continuously throughout 15 working days. The graphic designer was confronted with four workload factors that are common when a designer has to deliver creative outcomes: stress, sleep deprivation, night shift and moderate alcohol consumption. During each working day, an experience sampling method was employed to gather subjective data on mood and perceived workload. High level of arousal resulted in slowing of reaction times and higher variability whereas low arousal resulted in lower variability of reaction times. The results provide evidence that the wearable reaction time test is suited to perform long-term recordings and interpretations of reaction times during everyday work life.

- The effect of common daily activities on reaction times were investigated and compared between young and elderly subjects. The results showed that mean reaction times and reaction time variability were significantly affected by the type of activity. When comparing young and elderly subjects, it could be shown that the effect of activity on mean reaction times was significantly different between the young and elderly group. When elderly subjects were confronted with additional cognitive load, they showed slower reaction times and higher variability in comparison to younger subjects. The results of logistic regression analysis used for discriminating between idle and cognitive load conditions, confirmed the effect of activity on reaction times. A context-aware wearable reaction time test achieved the highest discrimination accuracy of 87.5%.

2.3. Limitations and Relevance

This work has investigated mobile and wearable techniques towards assessing cognitive load and cognitive performance in daily life. Cognitive load was assessed by means of HRV features obtained from a mobile ECG logger and by employing a wearable reaction time test. In the conducted laboratory experiments, variants of the Dual N-Back task were successfully employed to induce different levels of cognitive load in young and elderly test subjects.

It could be shown that HRV features obtained from a mobile ECG logger allows discriminating different levels of mental workload induced by the N-Back tasks in healthy subjects. The investigated HRV features showed a consistent response to increased workload. In an office-work scenario, it could be shown that the relationship between relevant HRV features and the subjective ratings of perceived mental workload could be modeled by employing a calibration procedure. Since the subjects under investigation were healthy and of young age, it remains open whether the results can be generalized to a broader population, e.g. regarding subject's age. The investigated office work scenario was restricted to working in front of a computer. Thus, it remains an open issue to what extent the proposed calibration and modeling approach has to be adapted to other scenarios such as surgeons and nurses in health care working environments [11, 12].

A wearable watch-like reaction time test was designed and implemented in a way that enables a continuous operation during daily life. In a laboratory evaluation, it could be shown that both variants of the wearable reaction time tests are suitable to measure factors that influence length and variability of reaction times. In a case study it could be shown that a long-term measurement and interpretation of reaction times during normal work life is feasible. In particular, it could be shown that subjective assessment of work demand and arousal level is significantly correlated with the obtained reaction times measures. It remains an open question whether the observed correlations are valid for other persons and different working environments.

In an experimental setting it could be shown how four common daily activities affect reaction time measures in young and elderly subjects. Incorporating the activity information into a context-aware wearable reaction time test resulted in an accuracy of 87.5% when discriminating between idle and cognitive load. It was concluded that a wearable reaction time test combined with a simple hand activity recognition system is feasible to detect changes in reaction time performance and variability during the investigated common daily life activities. Since in daily life many other activities occur, it remains an open issue whether the simple hand activity recognition is sufficient to achieve high accuracy when detecting cognitive load.

In the presented work, healthy subjects were investigated and the induced variations in cognitive functioning were induced by applying additional cognitive load. It remains an open issue to what extent the wearable reaction time test is suited for detecting variations in cognitive functioning in patients with cognitive impairments or other disabilities such as stroke.

2.4. Outlook

In this thesis mobile and wearable approaches were investigated in order to enable cognitive load and cognitive performance monitoring in daily life. Further research should address the following challenges:

- **Generalisability of the calibration procedure:** The relationship between relevant HRV features and the subjective ratings of perceived mental workload could be modeled by employing a calibration procedure. Future work should extend the investigated office-work scenario with a broader range of working activities. This will lead to new insights into the generalisability of the proposed calibration and modeling approach in other working scenarios.
- **Compatibility with daily life:** It was shown that the developed wearable reaction time test can be operated during common daily working activities. In order to meet comfort and user acceptance issues, both modules of the wearable reaction time test should be integrated into a normal watch. Depending on user's preferences, alternative response gestures may facilitate daily life employment.
- **Context-aware wearable reaction time test:** It was shown that a wearable reaction time test combined with a simple hand activity recognition system allows detecting cognitive load during four common daily activities. In future work the context recognition should be an integral part of the wearable RT test in order to generate stimuli events only in cases when the hand is not active for a certain amount of time. More advanced, the wearable RT test could be combined with additional sensor networks in order to control the generation of stimuli events depending on the actual context of the user. Such an advanced context-aware reaction time test would facilitate long-term assessments in daily life.

Bibliography

[1] Brain Workshop - a Dual N-Back game. http://brainworkshop.sourceforge.net/.

[2] R. Brunken, J. L. Plass, and D. Leutner. Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1):53–61, 2003.

[3] E. Gorus, R. De Raedt, and T. Mets. Diversity, dispersion and inconsistency of reaction time measures: Effects of age and task complexity. *Aging Clin Exp Res*, 18(5):407–417, Oct 2006.

[4] H. Harms, O. Amft, R. Winkler, J. Schumm, M. Kusserow, and G. Tröster. ETHOS: Miniature orientation sensor for wearable human motion analysis. In *Sensors, 2010 IEEE*, pp. 1037 –1042, nov. 2010.

[5] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[6] D. F. Hultsch, S. W. MacDonald, and R. A. Dixon. Variability in reaction time performance of younger and older adults. *J Gerontol B Psychol Sci Soc Sci*, 57(2):P101–115, Mar 2002.

[7] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[8] A. R. Jensen. *Clocking the mind: Mental chronometry and individual differences.* Elsevier, 2006.

[9] D. Kim, Y. Seo, , and L. Salahuddin. Decreased long term variations of heart rate variability in subjects with higher self reporting stress scores. In *Pervasive Healthcare*, 2008.

[10] D. Kim, Y. Seo, J. Cho, and C. H. Cho. Detection of subjects with higher self-reporting stress scores using heart rate variability patterns during the day. *Conf Proc IEEE Eng Med Biol Soc*, 2008: 682–685, 2008.

[11] C. Langelotz, M. Scharfenberg, O. Haase, and W. Schwenk. Stress and heart rate variability in surgeons during a 24-hour shift. *Arch Surg*, 143:751–755, Aug 2008.

[12] R. R. Looser, P. Metzenthin, S. Helfricht, B. M. Kudielka, A. Loerbroks, J. F. Thayer, and J. E. Fischer. Cortisol is significantly correlated with cardiovascular responses during high levels of stress in critical care personnel. *Psychosom Med*, 72(3):281–289, Apr 2010.

[13] M. Morris and F. Guilak. Mobile heart health: Project highlight. *IEEE Pervasive Computing*, 8(2):57–61, 2009. ISSN 1536-1268.

[14] M. Pavel, H. Jimison, T. Hayes, J. Kaye, E. Dishman, K. Wild, and D. Williams. Continuous, unobtrusive monitoring for the assessment of cognitive function. *Handbook of cognitive aging: Interdisciplinary perspectives. Thousand Oaks, CA: Sage Publications*, pp. 524–543, 2008.

[15] A. Riener, A. Ferscha, and M. Aly. 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*, pp. 99–106, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-571-0.

[16] V. Schumacher and M. Martin. Comparing age effects in normally and extremely highly educated and intellectually engaged 65 - 80 year-olds: potential protection from deficit through educational and intellectual activities across the lifespan. *Curr Aging Sci*, 2(3):200–204, Dec 2009.

[17] S. VaezMousavi, R. J. Barry, and A. R. Clarke. Individual differences in task-related activation and performance. *Physiology & Behavior*, 98(3):326 – 330, 2009. ISSN 0031-9384.

[18] S. Vandeput, J. Taelman, A. Spaepen, and S. Van Huffel. Heart rate variability as a tool to distinguish periods of physical and mental stress in a laboratory environment. In *Proceedings of the 6th international workshop on biosignal interpretation (BSI), New Haven, CT*, pp. 187–190, 2009.

3

Mental Workload Monitoring

Burcu Cinaz, Roberto La Marca, Bert Arnrich and Gerhard Tröster

Full publication title: Monitoring of Mental Workload Levels

Proceedings of IADIS eHealth Conference, 189-193, July 2010.

Abstract

Mobile 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 day-by-day quantification of mental load, first different load levels which occur during a workday have to be discriminated. This work goes one step towards this goal: we present our experiment design and preliminary results in discriminating different levels of mental workload based on heart rate features obtained from a mobile ECG system. Based on the subjective ratings of the participants under study, we show that all participants perceived the induced load levels as intended from the experiment design. The heart rate variability (HRV) features under investigation could be classified into two distinct groups. Features in the first group, representing markers associated with parasympathetic nervous activity, show a decrease in their values with increased workload. Features in the second group, representing markers associated with sympathetic nervous activity, show an increase of their values with increased workload. These results provide evidence that a mobile ECG system is suited to discriminate different levels of mental workload. This would enable the development of mobile applications to monitor mental workload levels and to prevent long-term damage by giving early warning signs in case of prolonged high workload.

3.1. Introduction

Recently, the European Foundation for the Improvement of Living and Working Conditions called the attention on the increasing level of mental disorders due to work-related stress. The workplace has changed due to globalization, use of new information and communication technology, resulting in an increased mental workload. Work-related stress was found to be the second most common work-related health problem across the EU15 [4]. Work-related stress occurs when there is a mismatch between job load and the capabilities, resources or needs of the worker. If the worker is not able to recover, long-term damage may result in the development of mental disorders [14]. Mobile healthcare offers new opportunities to prevent long-term damage by continuously monitoring mental workload and providing just-in-time feedback increasing workers awareness for improving self-management of mental workload.

Mobile monitoring of work-related stress or mental workload is still in an exploratory stage. One example is the exploratory research project "Mobile Heart Health", which aims to detect early signs of stress triggered by physiological or contextual changes and provide just-in-time mobile coaching [11]. Most of the existing studies often try to discriminate a state of mental load from a resting condition. In [2, 10, 12] two stress factors relevant at the workplace were under investigation: high cognitive load under time pressure and social-evaluative threat. In all three studies mild cognitive load was discriminated from a constant high stress level but different stress intensities were not investigated. [13] used a mental arithmetic task to induce mental workload and investigated the recovery patterns of physiological responses as indicators of stress. [8] studied heart rate variability (HRV) features of subjects under chronic stress. Subjects were divided into a high and a low stress group based on their self-reporting stress scores. [6] investigated the ability of short-term HRV metrics to discriminate between low and high level of mental workload.

In an office workplace scenario however, a worker is confronted with different levels of mental load during an office day. In order to achieve a day-by-day quantification of the mental load, first the different load levels have to be discriminated, and in a second step, the overall load can be estimated by accumulating these levels accordingly. This work goes one step towards this goal: we present our experimental results in discriminating different levels of mental workload. For an "everyday life application", a minimal sensor setup is desired for comfort reasons. This work therefore focuses on a single sensor modality: a mobile ECG system to measure heart rate (HR). The analysis of the HRV was chosen, because it represents a sensitive stress and mental load measure. Increased stress leads to an activation of the sympathetic nervous system and withdrawal of the parasympathetic nervous system [15]. In this work, we investigate HRV features in the time as well as in the frequency domain.

3.2. Methods

3.2.1. Experiments

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. The experiment was designed to investigate different levels of mental workload. Three sessions with low, medium and high workload were chosen. Each session consisted of a baseline (10 minutes), workload (20 minutes) and recovery (15 minutes) period. Subjects performed each session on separate days in the afternoon, while the different sessions were randomly assigned for each subject in order to avoid sequence effects and, therefore, to counterbalance learning effects. To induce different levels of mental workload, we used the N-Back Test [7]. Three variants of this task were used to induce low, medium and high workload which are likely to be present during an office work day:

1. **Position 1 Back (Low Workload; very easy task with visual stimuli):** A square appears every 4.5 seconds in one of eight different positions on a regular grid on the screen. The subject has to respond by using the keyboard if the position of the currently shown square is the same as the one that was presented just before. 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 seconds 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 just before. The result of the calculation has then to be entered on the keyboard. This task reflects medium cognitive load since the 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 seconds 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 2 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. The result of the calculation has then to be entered on the keyboard. This task represents a high cognitive load since the subject has to memorize the position and the value of two numbers and has to perform a math task in the given time.

Directly after the workload period, the subject was asked to assess his perceived workload. For this subjective rating we employed the NASA Task Load Index (TLX) [5]. First, the subject has to rate 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 is systematically asked which of the items represents the more important contributor to the workload. Based on these comparisons, the total workload is computed as a weighted average of the ratings. In addition, the individual performance for each workload task is recorded. The physiological responses were measured with the Zephyr BioHarness chest belt [1]. The monitoring belt consists of three smart fabric sensors to acquire cardiac activity, breathing rate and skin temperature. The ECG data was sampled with 250Hz. In this work, we focus on the analysis of heart rate variability features in the time and frequency domain.

3.2.2. Data analysis

We investigated the subjective ratings of the total workload obtained with the NASA Task Load Index by comparing the individual ratings for each workload period. Based on this data we performed an ANOVA test to investigate whether the perceived ratings differed significantly between the workload periods.

For the analysis of the heart rate data, we first removed RR intervals which differ more than 20% from their predecessors in order to remove artifacts. Next, we calculated a set of time and frequency HRV features following the guidelines of the European Task Force [9]: mean heartbeat intervals (Mean RR), standard deviation of RR intervals (SDNN), root

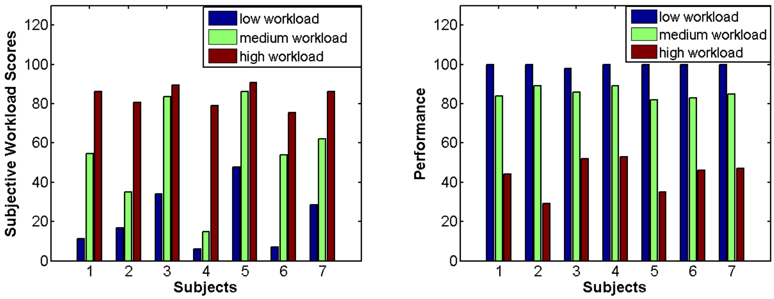

Figure 3.1. Perceived workload obtained from the NASA task load index and performance scores for each task.

mean square of successive differences (RMSSD), and the percentage of intervals that vary more than 50ms from the previous interval (pNN50). In addition, the HRV index (bin width 1/128 sec.), and the triangular interpolation of the R peak interval histogram (TINN) were extracted as geometric parameters. All features were calculated on the overall workload periods (20min) for each session. The analysis 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 [3]. 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 [9]. 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 component. We obtained all HRV features for each phase of the experiment (baseline, workload and recovery). We compared these features obtained for the three workload periods by using the ANOVA test. As significance level, $p < 0.05$ was considered.

3.3. Results

As shown in Figure 3.1 (left), all subjects perceived the induced load levels by the three variants of the N-Back as intended from the experiment design (ANOVA, $p < 0.01$). In Figure 3.1 (right) it is shown that also the individual performance reflects the three different workload levels. The mean values including standard errors of all HRV features

Table 3.1. Comparison of Mean HRV Features ± standard error during low, medium and high workload periods.

HRV Features	Low Workload	Medium Workload	High Workload	**F; p**
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 [n.u.]*	64.3 ± 2.9	70.1 ± 2.7	77.8 ± 4.6	$3.66; 0.04$
HF [n.u.]*	35.6 ± 2.9	29.8 ± 2.7	22.2 ± 4.6	$3.66; 0.04$
LF/HF*	1.9 ± 0.2	2.5 ± 0.3	4.6 ± 1.0	$4.59; 0.02$

Mean ± standard error

$*p < 0.05$

are listed in Table 3.1. 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, pNN50 and HF ($p < 0.05$) while Mean RR, SDNN, HRV Index and TINN show a consistent but non-significant decrease. In contrast, features in the second group show an increase of their values with increased workload. A statistically significant increase can be observed for the features LF and LF/HF ratio ($p < 0.05$).

3.4. Conclusion

We have presented an experiment design to induce three different levels of mental workload and to discriminate the workload levels based on heart rate features obtained from a mobile ECG system. According to the subjective ratings and the performance of the participants, 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. The investigated HRV features 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 of their values. The features RMSSD, pNN50 and HF showed a statistically significant decrease while LF and 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. Therefore, an increase in workload seems to be associated with a decrease in parasympathetic nervous activity and probably a concomitant increase in sympathetic activity. In conclusion, our experimental results show that a mobile heart rate sensor is suited to discriminate different levels of mental workload induced by cognitive tasks. In future work we are going to employ the mobile heart rate sensor in monitoring mental load during real office tasks.

Acknowledgment: This work is partly funded by the EU research project ProFiTex (http://www.project-profitex.eu/), grant agreement no.: CT-TP 228855-2. All views presented here reflect the author's opinion and not that of the EU commission. The authors thank Philip Dieringer for his support in conducting the experimental sessions.

Bibliography

[1] Zephyr. http://www.zephyr-technology.com/. URL http://www.zephyr-technology.com/.

[2] B. Arnrich, C. Setz, R. La Marca, G. Tröster, and U. Ehlert. What does your chair know about your stress level? *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):207 –214, march 2010. ISSN 1089-7771.

[3] G. D. Clifford. *Signal Processing Methods For Heart Rate Variability Analysis*. PhD thesis, St Cross College, 2002.

[4] European Foundation for the Improvement of Living and Working Conditions. Work-related stress. http://www.eurofound.europa.eu/.

[5] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[6] A. Henelius, K. Hirvonen, A. Holm, J. Korpela, and K. Muller. Mental workload classification using heart rate metrics. *Conf Proc IEEE Eng Med Biol Soc*, 1:1836–1839, 2009.

[7] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[8] D. Kim, Y. Seo, , and L. Salahuddin. Decreased long term variations of heart rate variability in subjects with higher self reporting stress scores. In *Pervasive Healthcare*, 2008.

[9] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz. 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, Mar 1996.

[10] R. L. Marca, P. Waldvogel, H. Thorn, M. Tripod, P. H. Wirtz, J. C. Pruessner, and U. Ehlert. Association between Cold Face Test-induced vagal inhibition and cortisol response to acute stress. *Psychophysiology*, Jul 2010.

[11] M. Morris and F. Guilak. Mobile heart health: Project highlight. *IEEE Pervasive Computing*, 8(2):57–61, 2009. ISSN 1536-1268.

[12] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine: Personal Healh Systems*, 2010.

[13] C. Soga, S. Miyake, and C. Wada. Recovery patterns in the physiological responses of the autonomic nervous system induced by mental workload. In *SICE, 2007 Annual Conference*, pp. 1366–1371, Sept. 2007.

[14] G. van Daalen, T. Willemsen, K. Sanders, and M. van Veldhoven. 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, 2009.

[15] J. A. Veltman and A. W. Gaillard. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5): 656–669, May 1998.

4

Individual Calibrations for Daily Life Monitoring

Burcu Cinaz, Bert Arnrich, Roberto La Marca and Gerhard Tröster

Full publication title: Monitoring of Mental Workload Levels during an Everyday Life Office-Work Scenario

Personal and Ubiquitous Computing, 17(2), pp 229-239, Springer, February 2013

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 classi-

fication of the mental workload level during office-work for five out of the seven subjects.

4.1. 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 [9]. Work-related stress occurs when there is a mismatch between job load and the capabilities of the worker [24]. 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 [3] and [22], 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 [23] 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. [13, 14] 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. [11] investigated the ability of short-term HRV metrics to discriminate between low and high level of mental workload.

In our previous work [5], 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 [16, 20, 25, 26]. In this work, we investigate HRV features in the time as well as in the frequency domain.

4.1.1. 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. [19] 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.

4.2. Data collection

4.2.1. Mobile ECG measurement

The physiological responses were measured with the Zephyr BioHarness chest belt as depicted in Fig. 4.1. The monitoring belt consists of three

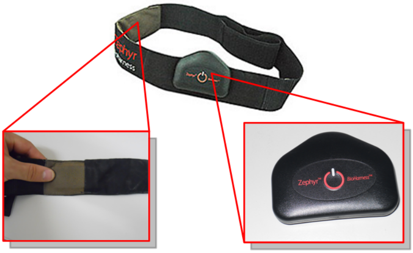

Figure 4.1. Zephyr BioHarness monitoring system.

smart fabric sensors to acquire cardiac activity, breathing rate and skin temperature [1]. 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.

4.2.2. 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 [15, 21].

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 ques-

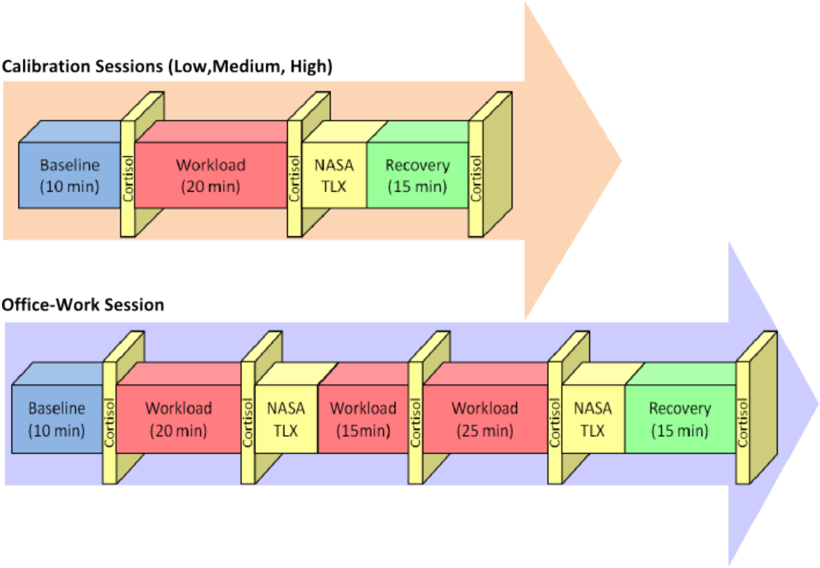

Figure 4.2. Experiment procedure for calibration and office-work sessions. A total of three calibration sessions were conducted which differed in the level of induced workload: low, medium, and high. The office-work condition consisted of 1 h of normal office working activities. The subjective rating of perceived workload was assessed with the NASA-TLX, whereas an objective measurement was assessed by collecting salivary cortisol at particular points in time.

tionnaires and cortisol collection). The experimental procedure can be seen in Fig. 4.2.

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) [10]. 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 [18]. 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. 4.2). 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 [8].

Calibration conditions: investigation of mental workload levels

Since individual's response to stress can vary to a huge extend, Morris et al. [19] 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, 12] 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 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. 4.3. 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.

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.

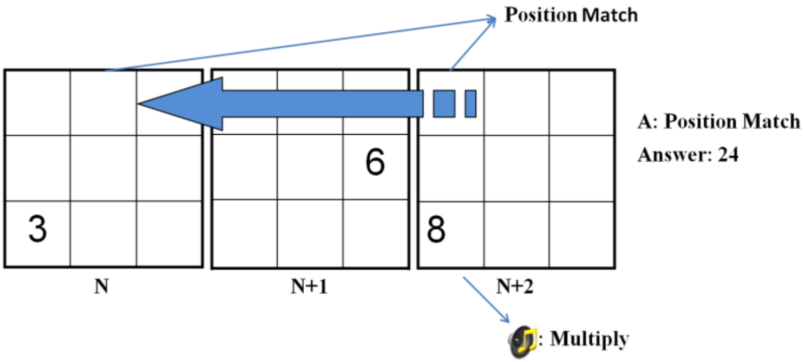

Figure 4.3. Dual Arithmetic 2 Back Task was used to induce high mental workload on subjects. An integer number between 0 and 9 appears every 4.5 s in one of eight different positions on a regular grid. In each step, a math operator (add, subtract, multiply, or divide) is presented via an audio message. 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.

4.3. 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 4.4 illustrates the complete data processing chain comprising the steps of preprocessing, feature extraction, and application of methods.

4.3.1. 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

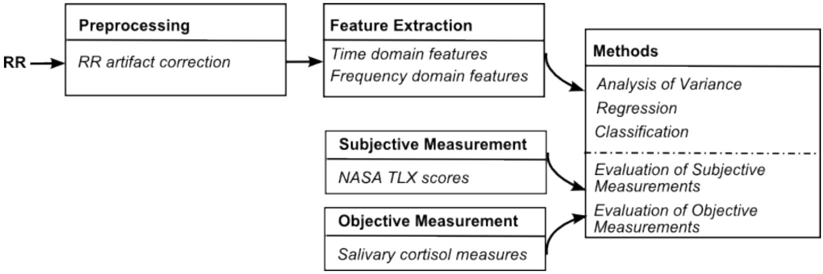

Figure 4.4. Block diagram showing the preprocessing, feature extraction, subjective and objective measurements, and mental workload evaluation steps.

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 (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 [6]. 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.

4.3.2. 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 ob-

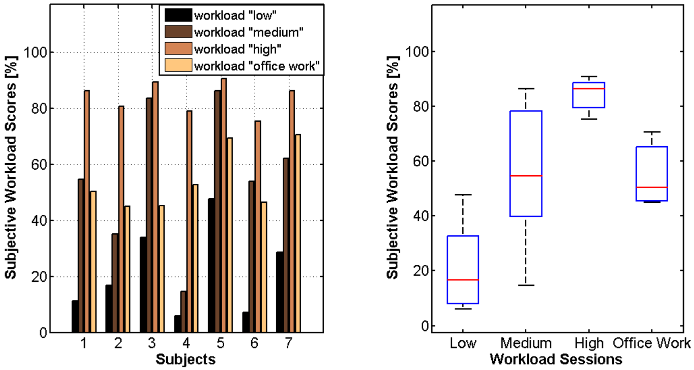

Figure 4.5. Subjective workload scores obtained from the NASA Task Load Index for each session and each subject (left). Comparison of the workload sessions for all subjects using box plots (right)

tained using MATLAB Arsenal toolbox [27] that encapsulates various classification algorithms and machine learning packages such as WEKA or libSVM [4]. 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.

4.4. 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.

4.4.1. Subjective measurement of mental workload

Figure 4.5 shows subjective workload scores for each subject. It can be seen that all subjects perceived the induced load levels by the three vari-

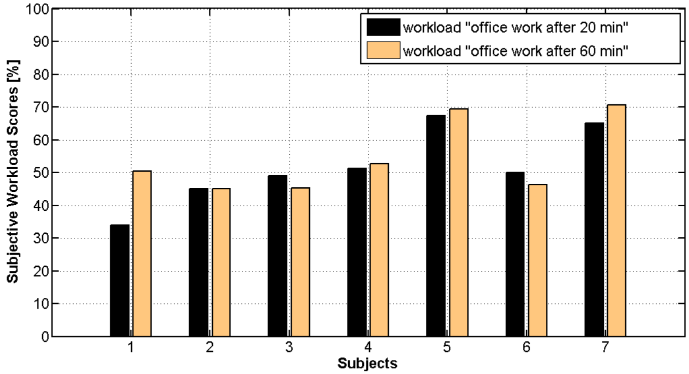

Figure 4.6. Comparison of the NASA results from two particular points in time (after 20 min and at the end of the working session)

ants 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. 4.5 (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. 4.6.

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,

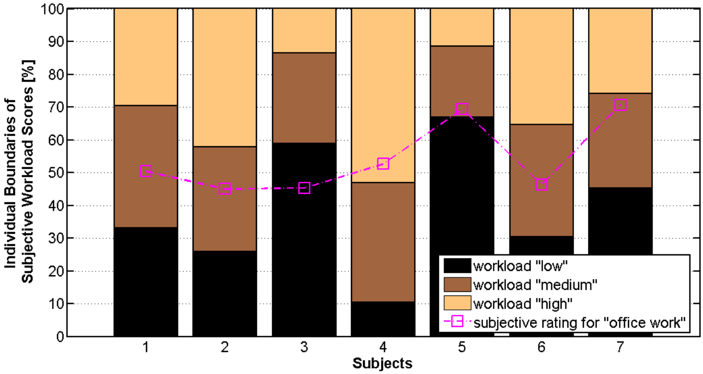

Figure 4.7. Individual boundaries for low-, medium-, and high-workload classes and subjective rating for the office-work session

$$
\begin{aligned}
low &< (low_c + medium_c)/2 \\
(low_c + medium_c)/2 \leq medium &\leq (medium_c + high_c)/2 \\
high &> (medium_c + high_c)/2
\end{aligned}
$$

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. 4.7.

4.4.2. 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 4.8 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

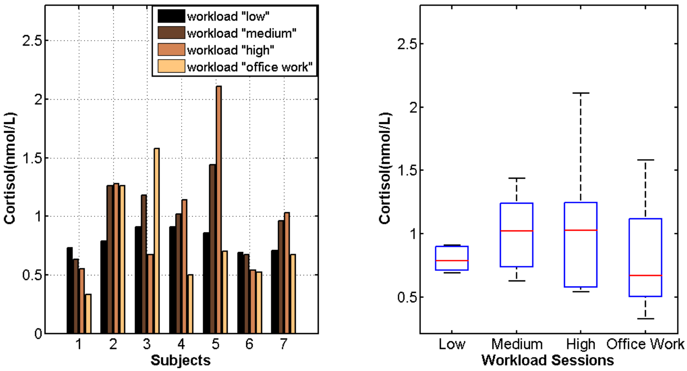

Figure 4.8. Normalized salivary cortisol levels of each subject for different workload periods (left). Comparison of the workload sessions for all subjects using boxplots (right)

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 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 [7]. In our case, the stressor was a continuous performance task that was controlled and not characterized by social-evaluative threat. By adding socialevaluative threat such as judging the subject about his performance by others during the experiment might increase cortisol levels.

4.4.3. Performance results

In each calibration session, the individual task performance was recorded. In Fig. 4.9, 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, $p < 0.001$).

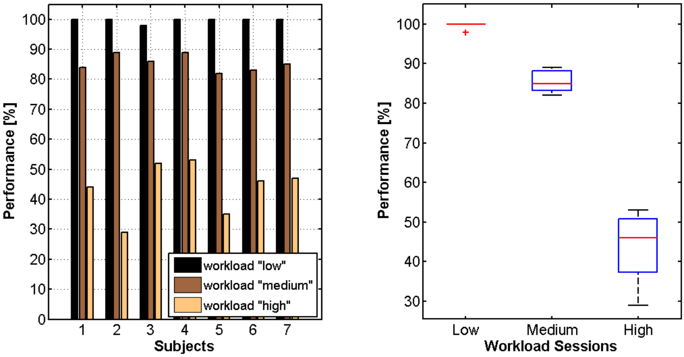

Figure 4.9. Performance scores of each subject for each N-Back session (left). Comparison of the workload sessions of the calibration condition for all subjects using boxplots (right)

4.4.4. 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 4.1. 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 ($p < 0.05$).

4.4.5. 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

Table 4.1. 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*	1.9 ± 0.2	2.5 ± 0.3	4.6 ± 1.0	$4.59; 0.02$

Mean ± standard error

$*p < 0.05$

SDNN, and TINN with SDNN ($r > 0.9$).

4.4.6. 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 4.2. 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 4.10 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.

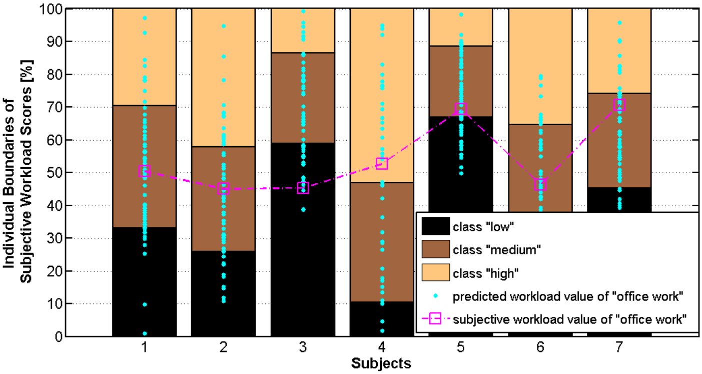

Figure 4.10. Predicted workload scores of the office-work session based on linear regression model

Table 4.3 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.

4.4.7. Classification

Table 4.4 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 subjects), whereas k-NN and SVM worked successfully for 5 subjects.

4.5. 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,

Table 4.2. 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	$24.966 * **$	$3.947*$	$11.499 * **$	$8.238 * **$

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

Table 4.3. Workload score, actual workload class, and estimated class with corresponding accuracy.

Subjects	NASA-office (%)	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.

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 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

Table 4.4. 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)
k-NN	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 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.

Bibliography

[1] Bioharness validation testing.

[2] Brain Workshop - a Dual N-Back game. http://brainworkshop.sourceforge.net/.

[3] B. Arnrich, C. Setz, R. La Marca, G. Tröster, and U. Ehlert. What does your chair know about your stress level? *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):207 –214, march 2010. ISSN 1089-7771.

[4] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, May 2011. ISSN 2157-6904.

[5] B. Cinaz, R. La Marca, B. Arnrich, and G. Tröster. Monitoring of mental workload levels. In *Proceedings of IADIS eHealth Conference*, 2010.

[6] G. D. Clifford. *Signal Processing Methods For Heart Rate Variability Analysis*. PhD thesis, St Cross College, 2002.

[7] S. S. Dickerson and M. E. Kemeny. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychol Bull*, 130:355–391, May 2004.

[8] R. Dressendörfer, C. Kirschbaum, W. Rohde, F. Stahl, and C. Strasburger. Synthesis of a cortisol-biotin conjugate and evaluation as a tracer in an immunoassay for salivary cortisol measurement. *The Journal of Steroid Biochemistry and Molecular Biology*, 43(7):683 – 692, 1992. ISSN 0960-0760.

[9] European Foundation for the Improvement of Living and Working Conditions. Work-related stress. http://www.eurofound.europa.eu/.

[10] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[11] A. Henelius, K. Hirvonen, A. Holm, J. Korpela, and K. Muller. Mental workload classification using heart rate metrics. *Conf Proc IEEE Eng Med Biol Soc*, 1:1836–1839, 2009.

[12] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[13] D. Kim, Y. Seo, , and L. Salahuddin. Decreased long term variations of heart rate variability in subjects with higher self reporting stress scores. In *Pervasive Healthcare*, 2008.

[14] D. Kim, Y. Seo, J. Cho, and C. H. Cho. Detection of subjects with higher self-reporting stress scores using heart rate variability patterns during the day. *Conf Proc IEEE Eng Med Biol Soc*, 2008: 682–685, 2008.

[15] C. Kirschbaum, B. M. Kudielka, J. Gaab, N. C. Schommer, and D. H. Hellhammer. Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. *Psychosom Med*, 61:154–162, 1999.

[16] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple Task Performance*, pp. 279–328, 1991.

[17] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz. 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, Mar 1996.

[18] R. L. Marca, P. Waldvogel, H. Thorn, M. Tripod, P. H. Wirtz, J. C. Pruessner, and U. Ehlert. Association between Cold Face Test-induced vagal inhibition and cortisol response to acute stress. *Psychophysiology*, Jul 2010.

[19] M. Morris and F. Guilak. Mobile heart health: Project highlight. *IEEE Pervasive Computing*, 8(2):57–61, 2009. ISSN 1536-1268.

[20] A. Riener, A. Ferscha, and M. Aly. 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*, pp.

99–106, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-571-0.

[21] N. Sato, S. Miyake, J. Akatsu, and M. Kumashiro. Power spectral analysis of heart rate variability in healthy young women during the normal menstrual cycle. *Psychosom Med*, 57:331–335, 1995.

[22] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine: Personal Healh Systems*, 2010.

[23] C. Soga, S. Miyake, and C. Wada. Recovery patterns in the physiological responses of the autonomic nervous system induced by mental workload. In *SICE, 2007 Annual Conference*, pp. 1366–1371, Sept. 2007.

[24] G. van Daalen, T. Willemsen, K. Sanders, and M. van Veldhoven. 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, 2009.

[25] G. F. Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *Int. J. Aviat. Psychol*, 12:3–18, 2002.

[26] G. F. Wilson and F. T. Eggemeier. Psychophysiological assessment of workload in multitask environments. *Multiple-Task Performance*, pp. 329–360, 1991.

[27] R. Yan. Matlabarsenal: A matlab package for classification algorithms. Technical report, School of Computer Science, Carnegie Mellon University, 2006.

5

Wearable Reaction Time Test

Burcu Cinaz, Christian Vogt, Bert Arnrich and Gerhard Tröster

Full publication title: A Wearable User Interface for Measuring Reaction Time

International Joint Conference on Ambient Intelligence, 41-50, Springer, 2011.

Abstract

Reaction time (RT) tests are known as simple and sensitive tests for detecting variation in cognitive efficiency. RT tests measure the elapsed time between a stimulus and the individual's response to it. A drawback of existing RT tests is that they require the full attention of a test person which prohibits the measurement of cognitive efficiency during daily routine tasks. In this contribution we present the design and the evaluation of a wearable RT test user interface which can be operated throughout everyday life. We designed a wearable watch-like device which combines the generation of a haptic stimulus and the recognition of subject's hand movement response. In order to show to what extent the wearable RT test is convenient to measure reaction times, we designed an experiment in which we measured the reaction times of ten subjects from two different setups. In the first half of the experiment, the participants performed a desktop-based RT test whereas in the second half of the experiment they performed the wearable RT test. In order to measure changes in the duration and variability of reaction times we induced additional cognitive load in both setups. We show that individual changes of reaction times occurred due to the cognitive load manipulation are similar for both desktop-based and wearable RT test. Additionally we investigate the subjective ratings of perceived workload. We conclude that the presented wearable RT test allows to measure changes in reaction times occurred due to additional cognitive load and hence would allow the assessment of changes in cognitive efficiency throughout everyday life.

5.1. Introduction and motivation

Reaction time (RT) tests measure how rapidly information can be processed and a response to it can be activated [14]. In other words, RT tests measure the elapsed time between a stimulus and the individual's response to it. According to Jensen [10], RT tests are well suited for cognitive assessment tests since in comparison to conventional psychometric techniques, RT tests offer a high sensitivity for detecting variation in cognitive efficiency and they can be virtually unlimited repeated. Several desktop-based RT tests have been developed in which users have to respond to visual stimuli by using keyboard, mouse or special buttons. An extensive review about computer-based cognitive tests can be found in [20]. There are several examples on applying RT tests to assess cognitive functioning: early detection of cognitive decline such as dementia or Alzheimer's disease in elderly; determining the ability to manage complex activities such as driving, piloting or search and rescue; identifying of children with intellectual disabilities such as Attention Deficit Hyperactive Disorder (ADHD).

The main drawback of existing desktop-based RT tests is the requirement of the full attention of the subject, i.e. the subject has to interrupt his daily routine for several minutes in order to perform the task on the computer. This restriction prohibits the measurement of cognitive efficiency during daily routine tasks, e.g. to determine the ability to manage complex activities such as piloting. Our goal is to develop reaction time tests which can be operated throughout everyday life by means of wearable devices. An important step in the development is to ensure that wearable reaction time tests are suitable to measure changes in reaction times similar to desktop-based approaches.

In this paper, we present the design and the evaluation of a wearable RT test user interface. We designed a wearable watch-like device which combines the generation of haptic stimuli and the recognition of subject's responses. Haptic stimulus is generated by a vibration motor. The subject's responses to a stimulus are performed by a hand movement which is recognized with an inertial measurement unit (IMU). For the evaluation of the wearable interface, we conducted an experiment to investigate to what extent we can measure the user's reaction time with our interface compared to desktop-based tests.

5.2. Related work

Three kinds of RT tests are commonly employed in literature [11]: simple, recognition and choice RT tests. Simple RT tests consist of one stimulus and one response. For instance the subject has to press a button as soon as the letter "X" appears at a predefined position or as soon as a light or sound appears. In recognition RT tests, the subject has to respond to a stimulus (target) and ignore other stimuli (non-target). This is sometimes called as "go/no-go" RT task. Recognition of a particular sound or symbol belongs to this category. Lastly, choice RT tests include multiple stimuli and multiple responses. The subject has to respond to each stimulus with a corresponding response, e.g. by pressing a certain key whenever a corresponding letter appears on the screen. A detailed series of recommendations on how to conduct experiments using reaction times and how to analyze the collected data can be found in [10, 14, 19].

Increasing age and age-related diseases like cognitive impairment are important factors which influence length and variability of reaction times [11]. It has been known that with increasing age, reaction times become more variable and longer. Gorus et al. showed that persons with cognitive deterioration demonstrated more intra-individual performance variability and more slowing in their reaction times than cognitively healthy elderly [5]. Braverman et al. showed in a clinical setting that the test of variables of attention (TOVA) is an accurate predictor of early attention complaints and memory impairments [2]. The effect of stress was investigated in an experiment which examines the cognitive performance under psychosocial stress [17]. The results showed that participants under stress were slower in their reaction times. Another application area of RT tests is the Attention Deficit Hyperactive Disorder (ADHD) patients. Children with ADHD have often difficulties in focusing on tasks and one of the most consistent findings is increased moment-to-moment variability in reaction time [18].

Most of the studies have in common that RT tests are operated with a computerized test which requires the full attention of the subject. Since the user has to interrupt his current activity to perform the test, most of these techniques are not feasible to be used during normal life activities. There exist only a few studies which measure one's cognitive performance continuously during everyday activities. Lieberman et al. implemented visual stimuli (3 LEDs), auditory stimuli (a miniature speaker) and two push buttons on a wrist-worn device to assess vigilance [13]. Ivorra et al. implemented a haptic stimulus to interrogate the central nervous system in a minimally obtrusive way [8]. As the response, the detection of a wrist movement is defined. By doing so, they showed that a simple RT test can be continuously administered throughout the course of normal life activities. However, a comparison of the wearable implementation with desktop-based RT tests is missing.

5.3. Materials and methods

5.3.1. Design of the wearable reaction time test

The wearable user interface to measure reaction times consists of two main modules: the stimuli module to generate haptic stimuli and the inertial measurement unit (IMU) module for detecting wrist movements. According to the literature, the wrist is a recommended stimulus site for wearable tactile displays [3, 12, 15, 16]. Therefore we designed a wrist-mounted tactile display in order to deliver the stimulus information to the user. For generating vibro-tactile stimuli, we used a coreless mini DC vibration motor with a diameter of 6mm and a resonant frequency around 200 Hz (manufactured by Precision Microdrivers Ltd.). In order to maximize the vibration amplitude and to ensure a proper sense of the vibration, we placed the motor in a separate plastic enclosure resulted in WxLxH dimensions of 90x55x30 mm which can be attached to the wrist of the user by using a strap. The stimuli module has its own battery supply. The vibration motor needs a continuous current of 83 mA and a start current of 150 mA. In a conservative calculation (continuous current of 150 mA, single stimulus duration 500 ms, 160 stimuli in 12 minutes), a total of 400 mAh would be required to perform a continuous reaction time test over 24 hours. In order to guarantee a continuous operation during at least one day, we have selected a lithium ion battery with 650 mAh. In addition, we have integrated an audio driver (MAX4410 by Maxim Inc.) in order to allow the generation of auditory stimuli through headphones. The IMU module consists of the ETH Orientation Sensor (ETHOS) which was developed in our laboratory [6]. The ETHOS includes a 3D accelerometer and gyroscope which allows to recognize the subject's gesture response.

The accelerometer and gyroscope were sampled with a frequency of 128 Hz. The detailed description of the ETHOS hardware platform can be found in [6]. We modified the firmware version of the ETHOS to control the vibro-tactile component. An implementation of a go/no-go task which is comprised of two wearable user interfaces to induce target and non-target stimuli can be seen in Fig. 5.1.

In order to automatically recognize a predefined hand gesture response to the haptic stimulus, we performed a preliminary experiment. Similar to the wearable RT test presented in [8], we have defined the response gesture as a fast rotation of the wrist. Three subjects performed a RT test on the wearable device during three different conditions. In the first condition, the subject was sitting on a chair while the arms

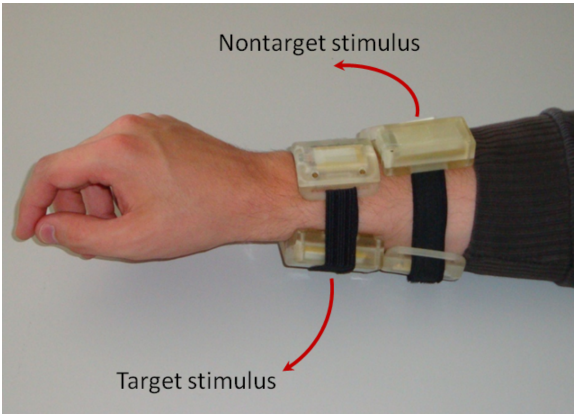

Figure 5.1. Wearable implementation of a go/no-go RT test. The left module generates the target stimuli as vibrations on the lower side of the wrist, whereas the right module generates non-target stimuli on the upper side of the wrist. For both modules, the IMU is placed on the opposite side of the vibration motor to recognize the user's hand gesture response.

were heading towards the floor. In the second condition, the arms were placed on the table. In the third condition, the subject was walking with a moderate speed (4km/h) on a treadmill. In each condition we recorded 3D acceleration and gyroscope data. Based on a visual inspection of the recorded data, we manually labeled each wrist response. For all conditions, it was clearly visible that the wrist-turn axis of the gyroscope (x-axis) was the most sensitive axis for detecting the fast rotation of the hand. In order to define a threshold for automatically detecting this hand gesture, we computed the correctly identified responses for different thresholds. With a threshold of 0.5 rad/sec, we could correctly detect the occurrence of this hand gesture response in all conditions. The raw gyroscope data and the occurrence of haptic stimuli are exemplary shown in Fig. 5.2. According to the simple threshold approach mentioned above, we compute the time point when the user was assumed to have reacted.

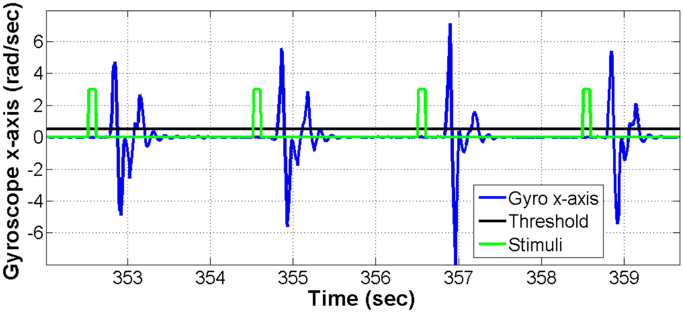

Figure 5.2. X-component (wrist-turn axis) of the gyroscope data while reacting to four stimuli with a fast rotation of the hand. Based on the threshold approach the time point when a subject has reacted was computed.

5.3.2. Experiment: Comparison of Wearable and Desktop RT Tests

Ten healthy subjects (6 male, 4 female, average age 26.15 years) participated in our experiment. All participants were paid 30 Swiss Francs for participating in one session of approximately 70 minutes. The goal of this experiment was to evaluate our wearable user interface by comparing the reaction times obtained by the wearable reaction time test with a desktop-based approach.

Experimental setup

For the desktop-based reaction time test we used a free version of the TOVA test which is implemented with the psychology experiment building language (PEBL) [1]. The implementation of the test is based on the description in [4]. A white square appears briefly on the screen, with a black square within it. Participant must respond only to targets (the black square on top) and ignore the non-targets (the black square on the bottom). Each stimulus is presented for 100 ms at 2000 ms intervals. For the wearable RT test, we placed two RT modules on the dominant wrist of the user as shown in Fig. 5.1. The left module generates the haptic target stimuli on the lower side of the wrist (volar side), whereas the right module generates haptic non-target stimuli on the upper side of the wrist (dorsal side). Similar to the desktop-based

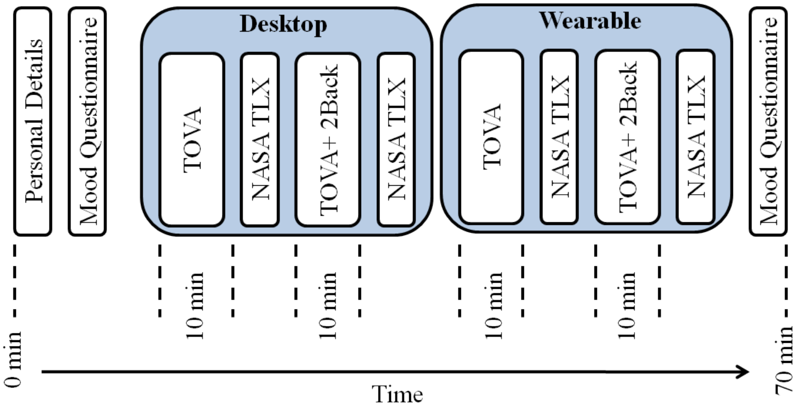

Figure 5.3. Experimental procedure including two conditions (baseline as TOVA single task and cognitive load as TOVA + 2Back dual-task) for each setup (desktop based RT and wearable RT).

RT test, each stimulus is generated for 100 ms at 2000 ms intervals. For recognizing the user's wrist turn response, we only use the data collected with the IMU placed opposite to the left module.

Experimental procedure

Each setup (wearable and desktop-based) consists of two experimental conditions: (i) single-task in which the subject has to respond to the target stimulus, and (ii) dual-task in which the subject has to solve a cognitive task in parallel to the single-task. Each condition lasts 10 minutes and contains 320 stimuli (160 targets and 160 non-targets). This leads to a total of 640 reaction times for each subject (160 targets x 2 setup x 2 conditions). As cognitive task we employed a variant of the N-Back test, the so-called "Audio 2-Back" [9] as explained in the following. The four phases used for each subject are:

- *Desktop-based RT (single-task):* The subject has to respond to each target stimulus by pressing the space bar on the keyboard and ignore non-target stimuli types. This is the typical variant of the test of variables of attention (TOVA).
- *Desktop-based RT with N-Back (dual-task):* In this condition a second task is added to the traditional desktop-based TOVA test.

The subject has to solve an Audio 2-Back task which is presented to the user simultaneously with the TOVA test. Thereby a letter is presented to the subject via an audio message and the subject has to respond if the currently pronounced letter is the same as the one that was pronounced 2 positions back. The response to the Audio 2-Back was done by saying "match" whenever a sound match occurs. The investigator controls if the subject answers correctly and gives feedback continuously to the user about correct and false answers to keep him concentrated on both of the tasks.

- *Wearable RT (single-task):* The subject has to respond to each target stimulus generated on the wrist by performing a wrist movement and ignore the non-target stimuli types.
- *Wearable RT with N-Back (dual-task):* The subject has to respond to target stimuli with hand movements, and solve Audio 2-Back task simultaneously.

In the following, we denote the single task of each setup as "baseline" and dual task as "cognitive load" condition. Directly after each condition for both settings, each subject was asked to indicate his perceived workload by completing the multidimensional assessment tool NASA Task Load Index (TLX) [7]. The rating consists of the following six scales: mental demand, physical demand, temporal demand, own performance, effort and frustration. Based on the ratings, the total workload was computed as a weighted average. The experimental procedure can be seen in Fig. 5.3.

5.4. Results

5.4.1. Reaction times

For the analysis, the mean reaction time and the standard deviation are considered as evaluation metrics. In Table 5.1 and Fig. 5.4 the means and standard deviations of the reaction times for all subjects in each condition are presented. First, it can be observed that for both desktop and wearable RT test, the mean reaction time is always increased during the cognitive load condition compared to the baseline condition. Mean reaction times of the desktop-based RT test are significantly correlated with the wearable RT test for the baseline condition ($r = 0.8336$, $p < 0.01$) as well as for the cognitive load condition

Table 5.1. Comparison of mean reaction times including standard deviation for the four experimental conditions.

Subjects	Desktop-based reaction times (ms)		Wearable reaction times (ms)	
	Baseline	Cognitive load	Baseline	Cognitive load
1	455 (102)	560 (217)	408 (103)	574 (208)
2	375 (93)	400 (173)	351 (89)	423 (173)
3	339 (68)	455 (208)	314 (80)	429 (134)
4	317 (62)	475 (192)	257 (44)	429 (182)
5	336 (94)	401 (123)	263 (83)	333 (115)
6	398 (70)	439 (135)	348 (57)	424 (133)
7	350 (59)	513 (229)	347 (142)	498 (187)
8	338 (91)	417 (132)	303 (111)	373 (185)
9	311 (42)	367 (129)	321 (49)	471 (170)
10	334 (59)	432 (181)	302 (70)	418 (147)

mean (standard deviation)

($r = 0.7070$, $p < 0.05$). Second, it can be observed that the increase in mean reaction times from baseline to cognitive load conditions is similar within subjects for both desktop and wearable setting. The relative difference (mean RT during cognitive load minus mean RT during baseline condition) between desktop and wearable setting are significantly correlated ($r = 0.7095$, $p < 0.05$). Consistently, the variability of reaction times was always higher in the cognitive load condition compared to the baseline condition for both desktop and wearable setting. No significant correlations were observed for the standard deviation of reaction times. Besides, it can be observed that for most subjects the mean reaction time in the wearable setting is lower compared to the desktop-based approach during baseline (exception is subject 9). In the cognitive load condition the mean wearable reaction times are again lower for most subjects (exceptions are subjects 1, 2 and 9). This might be explained by the fact that the transduction of a visual stimulus takes generally longer than the perception of a haptic stimulus as known from literature [10].

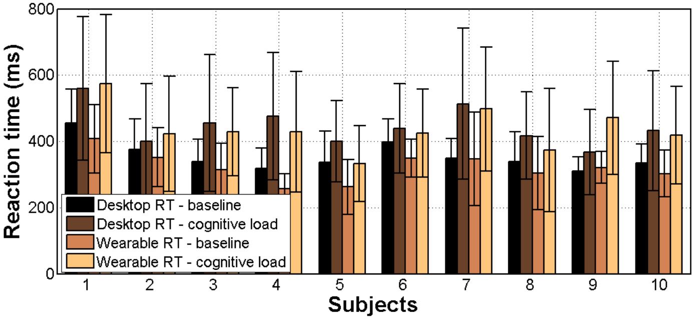

Figure 5.4. Mean reaction times for each subject collected from two conditions in each setting. Error bars indicate standard deviation.

5.4.2. Subjective ratings

Fig. 5.5 shows the subjective NASA task load index for each subject. As intended from the experiment design, it can be observed that for both desktop and wearable RT test, the subjective ratings of the cognitive load condition are higher than the respective baseline condition. A comparison between both tests shows that 50% of the subjects perceived higher total workload during baseline condition when using the wearable device. This is due to the fact that 90% of all subjects rated the "physical demand" item of NASA-TLX with higher values for the wearable setting since additional physical demand was required for responding with the wrist movement. However, the comparison between both cognitive load conditions shows that 70% of the subjects perceived lower workload when using the wearable device. These results indicate that operating the wearable device results in lower perceived mental load when the user is engaged in a primary task which requires a certain amount of information processing.

5.5. Conclusion and outlook

In this paper, we presented our experimental design and initial results in measuring reaction times of a person using a wearable RT test. In order to show to what extent a wearable interface is convenient to measure reaction times, we designed an experiment in which we measured response times of ten subjects from two different setups. In the first half

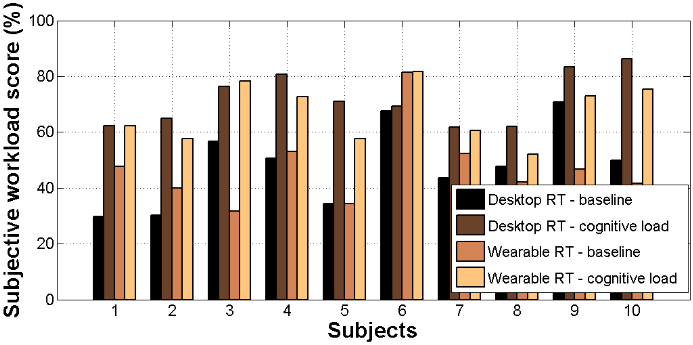

Figure 5.5. Subjective workload scores of each subject obtained from the NASA task load index for each condition.

of the experiment, the participants performed a desktop-based RT test whereas in the second half of the experiment they performed the wearable RT test. In order to measure changes in the duration and variability of reaction times we induced additional cognitive load in both setups. Besides the recording of reaction time data, subjective ratings of perceived workload were collected with the NASA-TLX. In a comparison of the obtained wearable reaction times with desktop-based reaction times, we showed that individual changes of reaction times due to the cognitive load are similar for both desktop-based and wearable RT test. According to the subjective ratings of the participants, we could show that all participants perceived the induced cognitive load as intended from the experiment design. Furthermore, subjective ratings showed that operating the wearable RT test interface under cognitive load results in lower perceived mental load compared to desktop-based reaction time test. These results suggest that the wearable RT test is more appropriate when the user is engaged in a second task which requires a certain amount of information processing. Based on the achieved results, we conclude that wrist-mounted reaction time tests seem feasible to measure factors which influence length and variability of reaction times and would allow the measurement of variation in cognitive efficiency throughout everyday life where the individuals are engaged in multiple tasks.

In our future work, we will perform statistical comparisons of different wearable RT setups, e.g. generation of the target/non-target stimulus on the dominant/non-dominant hand. In addition we are going to conduct long-term measurements of reaction times throughout daily life as cognitive performance indicator. We are planning to measure reaction times in real time from employers which have to perform complex intellectual tasks like flight operators. The obtained reaction times would allow us to identify loss of cognitive efficiency and to reduce the risk of cognitive overload.

Acknowledgments: This research has been partly funded by the European FP7 project ProFiTex, grant agreement number CT-TP 228855-2.

Bibliography

[1] Pebl: Psychological test battery. http://pebl.sourceforge.net/.

[2] E. R. Braverman, A. L. Chen, T. J. Chen, J. D. Schoolfield, A. Notaro, D. Braverman, M. Kerner, S. H. Blum, V. Arcuri, M. Varshavskiy, U. Damle, B. W. Downs, R. L. Waite, M. Oscar-Berman, J. Giordano, and K. Blum. Test of variables of attention (TOVA) as a predictor of early attention complaints, an antecedent to dementia. *Neuropsychiatr Dis Treat*, 6(1):681–690, 2010.

[3] H.-Y. Chen, J. Santos, M. Graves, K. Kim, and H. Z. Tan. Tactor localization at the wrist. In *Proceedings of the 6th international conference on Haptics: Perception, Devices and Scenarios*, EuroHaptics '08, pp. 209–218, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-69056-6.

[4] G. B. Forbes. Clinical utility of the Test of Variables of Attention (TOVA) in the diagnosis of attention-deficit/hyperactivity disorder. *J Clin Psychol*, 54(4):461–476, Jun 1998.

[5] E. Gorus, R. De Raedt, M. Lambert, J. C. Lemper, and T. Mets. Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. *J Geriatr Psychiatry Neurol*, 21(3):204–218, Sep 2008.

[6] H. Harms, O. Amft, R. Winkler, J. Schumm, M. Kusserow, and G. Tröster. ETHOS: Miniature orientation sensor for wearable human motion analysis. In *Sensors, 2010 IEEE*, pp. 1037 –1042, nov. 2010.

[7] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[8] A. Ivorra, C. Daniels, and B. Rubinsky. Minimally obtrusive wearable device for continuous interactive cognitive and neurological assessment. *Physiol Meas*, 29(5):543–554, May 2008.

[9] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[10] A. R. Jensen. *Clocking the mind: Mental chronometry and individual differences.* Elsevier, 2006.

[11] R. J. Kosinski. A literature review on reaction time, August 2009.

[12] S. C. Lee and T. Starner. BuzzWear: alert perception in wearable tactile displays on the wrist. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pp. 433–442, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9.

[13] H. R. Lieberman, F. M. Kramer, S. J. Montain, and P. Niro. Field assessment and enhancement of cognitive performance: Development of an ambulatory vigilance monitor. *Aviation, Space, and Environmental Medicine*, 78(5, Suppl.):B268–75, May 2007.

[14] R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford University Press, 1986.

[15] M. Matscheko, A. Ferscha, A. Riener, and M. Lehner. Tactor placement in wrist worn wearables. In *14th annual IEEE International Symposium on Wearable Computers (ISWC'10)*, pp. 1 –8, 2010.

[16] I. Oakley, Y. Kim, J. Lee, and J. Ryu. Determining the feasibility of forearm mounted vibrotactile displays. In *Proceedings of the Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, HAPTICS '06, pp. 27–34, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 1-4244-0226-3.

[17] U. Scholz, R. L. Marca, U. M. Nater, I. Aberle, U. Ehlert, R. Hornung, M. Martin, and M. Kliegel. Go no-go performance under psychosocial stress: Beneficial effects of implementation intentions. *Neurobiology of Learning and Memory*, 91(1):89 – 92, 2009. ISSN 1074-7427.

[18] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky. Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands. *Neuropsychologia*, 47(12):2389 – 2396, 2009. ISSN 0028-3932.

[19] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[20] K. Wild, D. Howieson, F. Webbe, A. Seelye, and J. Kaye. Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's and Dementia*, 4(6):428 – 437, 2008. ISSN 1552-5260.

6

Evaluation of the Wearable Reaction Time Tests

Burcu Cinaz, Christian Vogt, Bert Arnrich and Gerhard Tröster

Full publication title: Implementation and Evaluation of Wearable Reaction Time Tests

Pervasive and Mobile Computing, 8(6), pp 813 - 821, Elsevier, December 2012

Abstract

Conducting cognitive assessment tests throughout normal daily life offers new opportunities to early detect changes in cognitive efficiency. Such tests would allow identification of early symptoms of cognitive impairment, monitor the progress of disease processes related to cognitive efficiency and reduce the risk of cognitive overload. Reaction time tests are known as simple and sensitive tests for detecting variation in cognitive efficiency. A drawback of existing reaction time tests is that they require the full attention of a test person, which prohibits the measurement of cognitive efficiency during daily routine tasks. In this contribution we present the design, implementation and empirical evaluation of two wearable reaction time tests that can be operated throughout everyday life. We designed and implemented wearable watch-like devices, which combine the generation of haptic stimuli and the recognition of hand gestures as the subject's response. For the evaluation of the wearable interface, we conducted a user study with 20 subjects to investigate to what extent we can measure changes in length and variability of user's reaction time with the wearable interfaces in comparison to well accepted, traditional desktop-based tests. Based on the achieved statistical results, we conclude that the presented wearable reaction time tests are suitable to measure factors that influence length and variability of reaction times.

6.1. Introduction

Conducting cognitive assessment tests throughout normal daily life offers new opportunities to early detect changes in cognitive efficiency. On the one hand, such tests would allow identification of early symptoms of cognitive impairment in risk-groups like the elderly or to monitor the progress of disease processes related to cognitive efficiency like Alzheimer's disease. On the other hand, a continuous assessment of cognitive efficiency would allow reducing the risk of cognitive overload for employers who have to perform complex intellectual tasks like flight assistants. At present, cognitive assessment is usually done in hospital environments by clinical assessments such as Mental State Examination (MSE), neuropsychological tests or mental state questionnaires

[3, 7, 20, 23]. Promising alternatives are computerized assessments of cognitive functioning. In comparison to traditional neuropsychological assessment methods, computerized tests offer benefits such as cost reduction and time savings. An extensive review about computer-based cognitive tests for detecting cognitive decline can be found in [26].

A simple and sensitive computerized cognitive assessment test is the reaction time (RT) test, which is defined as a measure of how rapidly information can be processed and a response to it can be activated [18]. In other words, it is the elapsed time between a stimulus and the individual's response to it. According to Jensen [13], RT tests are well suited for cognitive assessment since they offer a high sensitivity for detecting variation in cognitive efficiency and in comparison to conventional psychometric tests they can be repeated virtually an unlimited amount of times. In recent years several desktop-based RT tests have been developed. Commonly, visual stimuli are generated at random time points and the subject has to respond by using a keyboard, mouse or special buttons. There are several examples of applying RT tests to assess cognitive functioning: early detection of cognitive decline such as mild cognitive impairment or Alzheimer's disease; determining the ability to manage complex activities such as driving or piloting, and identifying children with intellectual disabilities such as Attention Deficit Hyperactive Disorder (ADHD).

The main drawback of existing desktop-based RT tests is their incompatibility with daily life since the subject has to interrupt his daily routine for several minutes in order to perform the RT test. As a consequence the measurement of cognitive efficiency during daily routine tasks is not feasible so far. In order to overcome this limitation our goal is to develop wearable reaction time tests that can be operated throughout everyday life. An important step in the development is to validate whether wearable reaction time tests measure changes in reaction times similar to desktop-based approaches.

In this contribution we present the design, implementation and empirical evaluation of two wearable reaction time test variants. We have designed and implemented wearable watch-like devices, which generates haptic stimuli and recognizes the user's hand gesture as a response. Haptic stimuli are generated from time to time by a vibration motor. The subject has to respond to a stimulus by performing a forearm rotation, which is recognized with an inertial measurement unit (IMU). For the evaluation of the wearable reaction time interfaces, we conducted an experimental comparison with traditional desktop-based tests. We have compared duration and variation of reaction times between the wearable interfaces and desktop-based RT tests.

In the following we first provide an overview about related work. Next, we describe the implementation of wearable reaction time tests and the experimental evaluation procedure. In the result section we present the outcomes of the performed evaluation. Finally, we conclude our work, discuss the results and provide an outlook on future work.

6.2. Related work

There exist mainly three kinds of reaction time tests [15]: simple, recognition and choice reaction. In simple reaction time tests the subject has to respond to one stimulus with a dedicated response. For instance the subject has to press a certain button as soon as a particular symbol appears at the screen. In recognition reaction time tests, which are sometimes called "go/no-go" tasks, the subject has to respond to target stimuli and ignore non-target stimulus types. Lastly, in choice reaction time tests the subject has to respond to multiple stimuli with a corresponding response e.g. pressing a dedicated key whenever a corresponding letter appears on the screen. A detailed series of recommendations on how to conduct experiments using reaction times and how to analyze the obtained data can be found in [13, 18, 25].

Increasing age and age-related diseases like cognitive impairment are known to influence length and variability of reaction times [15]. For instance, Gorus et al. showed that reaction times and performance variability are potential markers for the early detection of Alzheimer's disease. Persons with cognitive deterioration demonstrated more intra-individual performance variability and more slowing in their reaction times than cognitively healthy elderly. Thus, the authors suggest that intra-individual performance variability and RT are predictors for mild cognitive impairment (MCI) and Alzheimer's disease (AD) [9]. Braverman et al. showed that the test of variables of attention (TOVA) is an accurate predictor of early attention complaints and memory impairments in a clinical setting [3]. The effect of stress had been studied by investigating the cognitive performance under psychosocial stress [22]. Subjects were randomly assigned to the Trier Social Stress Test (TSST) versus a rest condition. After the stress test, a go/no-go task was performed by each participant. Participants in the stress condition were slower in their reaction time than in the rest condition. Another application area of reaction time tests is the investigation of Attention

Deficit Hyperactive Disorder (ADHD) patients. Children with ADHD have in general difficulty in focusing on tasks. In [24] the authors performed a study to examine the RT variability in ADHD using go/no-go tasks with differing levels of cognitive demand. A total of 140 children (57 with ADHD) completed both a simple go/no-go task and a more complex go/no-go task with increased working memory load. The resulting findings showed increased variability in both tasks for ADHD children.

Most of the studies have in common that the employed RT tests are operated with a computerized test, which requires the full attention of the subject for several minutes. Hence, most of these techniques are not feasible to be used without interrupting normal life activities. There exist only a few studies which investigate the feasibility of measuring reaction times continuously during everyday activities. Lieberman et al. implemented a wrist-worn device to assess vigilance. The device consists of visual stimuli (3 LEDs), auditory stimuli (a miniature speaker) and two push buttons [17]. Ivorra et al. [11] implemented a haptic stimulus to interrogate the central nervous system in a minimally obtrusive way. As the response the detection of a wrist movement is defined. In a first feasibility study they showed that a simple RT test can be continuously administered throughout the course of normal life activities. However, an evaluation of the wearable implementation in comparison with state of the art desktop-based RT tests is missing and the obtained reaction times were not further analyzed. In our recent work [6], we followed the approach of Ivorra et al. and designed a wearable reaction time device which combines the generation of haptic stimuli and the recognition of forearm rotation as subject's response. In our preliminary results we presented the mean and standard deviation of reaction times obtained from ten subjects. In this contribution we present a user study with 20 subjects in order to evaluate two wearable reaction time tests. In particular, we investigate to what extent we can measure changes in length and variability of user's reaction time in comparison to well accepted, traditional desktop-based tests. We analyze within subjects as well as between subjects effects.

6.3. Materials and methods

6.3.1. Design of the wearable reaction time test

As outlined in the related work section above, most of the desktop-based reaction time tests consist of visual stimuli and keyboard responses. In order to design a wearable reaction time test, we follow the approach of Ivorra et al. [11]: we replaced the visual stimulus by a haptic stimulus and instead of a keyboard response we employ a forearm rotation. Haptic stimulus was chosen since the visual and auditory channel is often already occupied for everyday life tasks. We apply the haptic stimuli on the wrist since from the literature it is known that the performance of perceiving vibro-tactile stimulus is considerable enhanced when the stimulation fall near natural anatomical anchor points such as wrist and elbow [5]. In addition, the placement on the wrist was recommended by several studies about wearable tactile displays [4, 16, 19, 21]. The forearm rotation as response to the haptic stimulus was selected to allow performing the test in a natural way without the need of pressing any extra button.

The wearable user interface consists of two main modules: the stimuli module to generate haptic stimuli and the inertial measurement unit (IMU) module for detecting forearm rotations. For generating haptic stimuli, we used a coreless mini DC vibration motor with a diameter of 6 mm and a resonant frequency around 200 Hz (manufactured by Precision Microdrivers Ltd.). In order to maximize the vibration amplitude and to ensure a proper sense of the vibration, we placed the motor in a separate enclosure with a dimension of 90x55x30 mm. The module is attached to the wrist of the user by using a strap. The complete housing was constructed using the 3D CAD software Autodesk Inventor and fabricated with the rapid prototyping equipment available at the department of mechanical and process engineering at ETH Zurich. The stimuli module has its own battery supply in order to guarantee a continuous operation during at least one day. In a conservative calculation (current of 150 mA, single stimulus duration 500 ms, 160 stimuli in 12 min), a total of 400 mA h would be required to perform a continuous reaction time test over 24 h. Hence, we have selected a lithium ion battery with 650 mA h. In a feasibility study we have investigated the turn-on-delay of 10 Vibration motors. Each of the motors showed a turn-on time derivation of at most ±5 ms for repeatedly applied stimuli. All tested motors were within 20 ms uncertainty with respect to their turn-on delay.

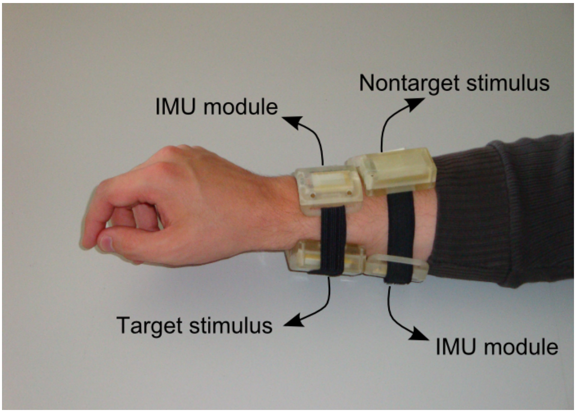

Figure 6.1. Wearable implementation of a go/no-go RT test. The left module generates the haptic target stimuli as a vibration on the lower side of the wrist, whereas the right module generates non-target haptic stimuli on the upper side of the wrist. For both modules, the IMU is placed on the opposite side of the vibration motor to control the generation of haptic stimuli and to recognize the user's gesture responses.

The IMU module consists of the so-called ETH Orientation Sensor (ETHOS), which was developed in our laboratory [10]. The ETHOS includes a three-axis accelerometer and a gyroscope, which is used to recognize the subject's gesture response. The accelerometer and gyroscope were sampled with a frequency of 128 Hz. The detailed description of the ETHOS hardware platform can be found in [10]. We modified the firmware version of the ETHOS to control the vibro-tactile component. An implementation of a wearable go/no-go task can be seen in Fig. 6.1.

In order to develop a procedure for automatically recognizing a predefined hand gesture response, we performed a preliminary experiment. As a response gesture we have defined an outward rotation of the forearm (see right side of Fig. 6.3) similar to the wearable RT test presented in [11]. Three subjects performed a simple RT test during three different conditions. In the first condition, the subject was sitting on a chair while the arms were heading towards the floor. In the second condition, the arms were placed on the table. In the third condition, the subject was walking with a moderate speed (4 km/h) on a treadmill. In each condition we randomly applied haptic stimuli. As soon as the subject perceived the target vibration stimulus he had to

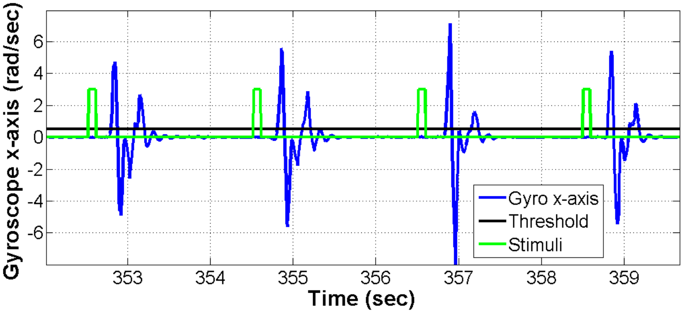

Figure 6.2. X-component (wrist-turn axis) of the gyroscope data while reacting to four stimuli with a fast rotation of the hand. Based on the threshold approach the time point when a subject has reacted was computed.

rotate his forearm outwards. 3D acceleration and gyroscope data were sampled with a frequency of 128 Hz. Based on a visual inspection of the recorded data, it was clearly visible that the wrist-turn axis of the gyroscope (x-axis) was the most sensitive axis for detecting the rotation of the forearm. Each of the axes of the IMU module and the hand gesture are illustrated in Fig. 6.3. In order to define a threshold for automatically detecting the forearm rotation, we computed the correctly identified responses for different thresholds. With a threshold of 0.5 rad/s, we could correctly detect the onset of the forearm rotation in all conditions for all subjects. The raw gyroscope data and the time points when haptic stimuli were applied are exemplary shown in Fig. 6.2. The reaction time is defined as the time difference between the onset of the haptic stimulus and the onset of the forearm rotation.

We have implemented the wearable go/no-go task in two variants. In the first design, we placed the target stimulus on the lower side and the non-target stimulus on the upper side of the dominant wrist. In the second two-handed design, the target stimulus module is placed on the lower side of the dominant wrist while the non-target stimulus is placed on the lower side of the non-dominant wrist. The two sensor placements are depicted in Fig. 6.3. In the following, we denote the first setup as "one-handed" and the second setup as "two-handed".

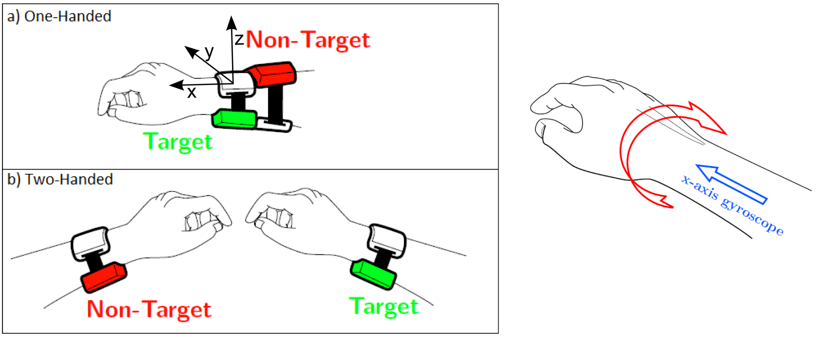

Figure 6.3. (Left) Two implementations of wearable go/no-go reaction time tests. In the first one-handed design, target and non-target stimulus modules are both placed on the dominant wrist. In the second two-handed design, the target stimulus module is placed on the dominant wrist while the non-target stimulus is placed on the non-dominant wrist. (Right) Illustration of the outward rotation of the forearm including x-axis of the gyroscope.

6.3.2. Participants

Twenty subjects (12 male, 8 female, average age 24.3 years) were recruited for the experiment. All participants were compensated with 30 Swiss Francs for participating in one session of approximately 70 min. Before starting the experiment, participants were briefly informed about the experimental procedure and they were offered a detailed report of their reaction time results after completing the analysis of their reaction times. Participants were randomly assigned to one of the two experimental groups, which differ in sensor placement (one-handed vs. two-handed). Each group consists of 4 female and 6 male participants.

6.3.3. Experimental overview

This experiment addresses two main goals: (i) to evaluate our wearable user interface by comparing the reaction times obtained by the wearable reaction time tests with those collected by a desktop-based reaction time test and (ii) to compare reaction times collected by the one-handed with those collected by two-handed wearable setup. As a desktop-based reaction time test we used a free version of the go/no-go TOVA test [2]. The test procedure follows the description provided in

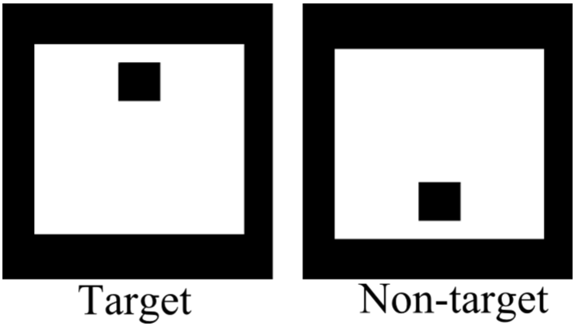

Figure 6.4. Target and non-target stimulus types of the TOVA test.

[8]. A white box appears for 100 ms on the screen. In the white box, a black square is shown on the top or on the bottom. The participant must respond only when the black square appears on the top (target stimulus) and must not respond if the black square appears on the bottom (non-target stimulus). Fig. 6.4 depicts the two target and non-target stimuli types of the TOVA test. Each stimulus is presented at 2000 ms intervals. Similar to the desktop-based RT test, in the wearable variant each haptic stimulus is generated for 100 ms at 2000 ms intervals. The IMU, used for recognizing the user's forearm rotation, is placed on the opposite to the stimuli module.

6.3.4. Experimental procedure

For both desktop-based and wearable setting two experimental conditions are investigated: (i) idle condition in which the subject is only performing the reaction time test, and (ii) cognitive load condition in which the subject is performing a continuous performance task in parallel to the reaction time test. Each condition lasts about 10 min and contains 320 stimuli (160 targets and 160 non-targets). In total 640 reaction times (160 targets x 2 setups x 2 conditions) were collected from each subject. For inducing cognitive load we have employed the so-called "Audio 2-Back" test, which is a variant of N-Back tests [1, 12]. In the following the two setups and the two conditions are described in more detail.

1. *Desktop-based RT (idle):* The subject has to respond as fast as possible to each visual target stimulus by pressing the space bar

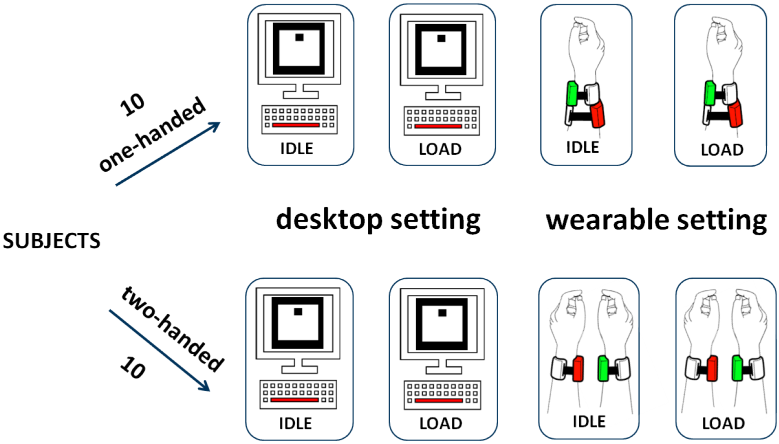

Figure 6.5. Experimental procedure: Subjects are randomly assigned in one of the two experimental groups i.e. one-handed and two-handed. Each subject performed idle and load conditions for desktop-based and wearable reaction time test separately.

on the keyboard and ignore non-target stimuli. This is the typical variant of the test of variables of attention (TOVA).

2. *Desktop-based RT with N-Back (cognitive load):* In this condition a second task is added to the traditional desktop-based TOVA test. The subject has to solve an Audio 2-Back task, which is presented to the user in addition to the TOVA test. In the Audio 2-Back task every 3 s a letter is presented to the subject via an audio message. The subject must only respond if the currently pronounced letter is the same as the one that was pronounced two positions back. Since the subject has to memorize the two prior letters and has to perform a comparison with the current letter in time, this task induces additional cognitive load. In order to respond the subject has to say “match” whenever a sound match occurs. In order to keep the subject engaged in the second task, the experiment leader continuously controls the performance of the subject and provides direct feedback to the user about correct and false answers.

3. *Wearable RT (idle):* The subject has to respond as fast as possible to each haptic target stimulus by performing a forearm rotation

and ignore the non-target stimuli.

4. *Wearable RT with N-Back (cognitive load):* Like in (3) the subject has to respond as fast as possible to each haptic target stimulus by performing a forearm rotation. Like in (2) the subject has to solve the Audio 2-Back task simultaneously.

The experimental procedure can be seen in Fig. 6.5. Each experimental group performed the same procedure. The only difference was the sensor placement during the wearable reaction time tests.

6.3.5. Measures

For the analysis, mean reaction time, intra-individual variability measured as the coefficient of variation (CV) [14, 24] and accuracy (number of commission errors defined as the number of responses to non-targets) are used as evaluation metrics. For each individual, the coefficient of variation is computed as the standard deviation divided by the mean reaction time.

6.3.6. Statistical analysis

We performed separate statistical analyses to analyze within subjects and between subjects effects. For within subjects comparison, we performed two-way repeated measures ANOVA for each experimental group (one-handed vs. two-handed) using the three measures as dependent variables, while setting (desktop vs. wearable) and condition (idle vs. load) serve as independent variables. Data were tested for normal distribution using Kolmogorov-Smirnov test. Within subject comparisons address the following research questions: Do the mean reaction time, variability and accuracy differ between desktop and wearable settings? Do mean reaction time, variability and accuracy differ between idle and load conditions?

In order to analyze the effects in detail, we performed a pairwise comparison (with Bonferroni correction) using repeated measures one-way ANOVA. In this manner, we compared pairs of independent variables with each other. In order to analyze the setting effect, we did a pairwise comparison of desktop-idle vs. wearable-idle and desktop-load vs. wearable-load. In order to analyze the condition effect, we did a pairwise comparison of desktop-idle vs. desktop-load and wearable-idle vs. wearable load. For one-way repeated measures ANOVA, data were tested for homogeneity of variance using Mauchly's sphericity test and if necessary Greenhouse-Geisser correction was performed.

In order to compare different sensor placements for the wearable setting, we performed between subjects comparison. For each measure, we performed a mixed 2x2 ANOVA with group (one-handed vs. two-handed) as between-subject variable and condition (idle vs. load) as within-subject variable. Between subjects comparison addresses the following question: Do wearable reaction times differ between different sensor placements (one-handed vs. two-handed)? Data were tested for normal distribution and homogeneity of variance using the Kolmogorov-Smirnov and Levene's test. In order to investigate the differences between each condition of each group we performed an independent samples t-test. Greenhouse-Geisser correction was performed if necessary in the repeated measures data.

6.4. Results

We first present the boxplots of mean reaction time, coefficient of variation, and commission error for both groups and both settings in Fig. 6.6. First, it can be observed that in each group the mean reaction time of the subjects is always increased during the load condition compared to the idle condition. Second, it is visible that the mean reaction times of the wearable setting are faster than those of the desktop-based setting during the idle condition. Third, we can observe that the variability is increased during the load condition for both settings and groups. The number of commission errors does not show a clear visible difference between settings and groups. In the following we provide detailed statistical analysis for within subjects comparison and for between subject comparison.

6.4.1. Within subjects comparison

Mean reaction times

Two-way repeated measures ANOVA of mean reaction time showed a significant main effect for condition for one-handed as well as two-handed group ($p < 0.001$). Subjects were thus always slower during the cognitive load condition compared to the idle condition. There was not a significant main effect of the setting for the one-handed group. Mean reaction times of the subjects were not significantly affected by whether subjects performed desktop or wearable reaction time tests.

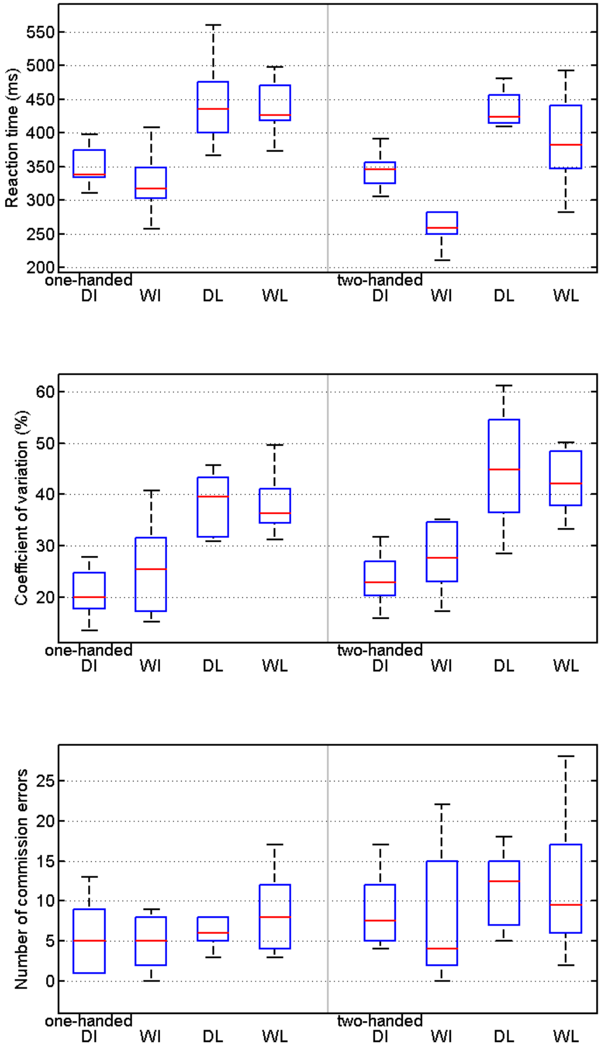

Figure 6.6. Boxplots of mean reaction time (above), coefficient of variation (middle), and commission error (below) for both groups and both settings. (DI = desktop-idle, WI = wearable-idle, DL = desktop-load, WL = wearable-load)

On the contrary, the setting showed a significant main effect for the two-handed group ($p < 0.01$).

As shown in Table 6.1, pairwise comparison revealed that mean reaction times of the desktop-based reaction time test were significantly slower than the wearable reaction time test during the idle condition (desktop-idle vs. wearable-idle) for both one-handed ($p < 0.05$) and two-handed group ($p < 0.01$). However for both of the groups a significant difference between both settings was not observed during the load condition (desktop-load vs. wearable-load).

Variability

Two-way repeated measures ANOVA of CV indicated a significant main effect for condition (idle vs. load) for both of groups showing that subjects demonstrated higher variability under load condition compared to idle condition ($p < 0.001$). There was not a significant main effect for setting for both groups.

As shown in Table 6.1, pairwise comparison confirms the significant increase of variability under cognitive load for both groups and settings.

Accuracy

Two-way repeated measures ANOVA of commission errors showed a significant main effect for condition only for two-handed group ($p < 0.01$). However, pairwise comparison between idle and load conditions for each setting (desktop-idle vs. desktop-load and wearable-idle vs. wearable-load) revealed that there was not a significant increase in commission errors during load condition compared to idle condition.

6.4.2. Between subjects comparison

Mean reaction times

Mixed 2x2 ANOVA showed a significant condition ($p < 0.001$) and group effect ($p < 0.05$) on mean reaction times. As shown in Table 6.2, post-hoc group by group comparison revealed that during the idle condition subjects in the two-handed group were significantly faster than the subjects in the one-handed group ($p < 0.05$).

Table 6.1. The results of one-way repeated measures ANOVA and Bonferroni corrected pairwise comparison on each reaction time task.

		Desktop setting		Wearable setting		Comparisons (p-values)			
		IDLE(1)	LOAD(2)	IDLE(3)	LOAD(4)	Setting effect		Condition effect	
						1-3	2-4	1-2	3-4
One handed	Mean RT	355.2(43.5)	445.9 (57.8)	321.4 (44.8)	437.1 (66.1)	*	NS	***	***
	CV	20.7 (4.7)	38.3 (5.8)	25.7 (8.5)	37.5 (5.6)	NS	NS	***	**
	Errors	5.7 (4.5)	8.4 (7.2)	6.0 (5.5)	8.4 (5.4)	NS	NS	NS	NS
Two handed	Mean RT	343.5 (25.1)	434.9 (50.8)	274.0 (46.8)	385.7 (64.4)	**	NS	***	***
	CV	23.5 (5.2)	45.3 (10.6)	27.6 (6.1)	42.8 (5.8)	NS	NS	***	***
	Errors	8.8 (4.8)	11.6 (4.6)	8.0 (8.3)	11.8 (8.0)	NS	NS	NS	NS

Mean (standard deviation); NS: not significant.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

Table 6.2. Between group pairwise comparison (independent samples t-test).

	Group:one handed		Group:two handed		Comparisons (p-values)	
	IDLE(1)	LOAD(2)	IDLE(3)	LOAD(4)	1-3	2-4
Mean RT	321.4 (44.8)	437.1 (66.1)	274.0 (46.8)	385.7 (64.4)	*	NS
CV	25.7 (8.5)	37.5 (5.6)	27.6 (6.1)	42.8 (5.8)	NS	NS
Errors	6.0 (5.5)	8.4 (5.4)	8.0 (8.3)	11.8 (8.0)	NS	NS

Mean (standard deviation); NS: not significant.

* $p < 0.05$.

Variability

Mixed 2x2 ANOVA indicated a significant effect for condition ($p < 0.001$) but not a significant effect for group on variability. Subjects overall demonstrated higher CV during load condition compared to the idle condition. There was not a significant difference between groups.

Accuracy

Mixed 2x2 ANOVA showed a significant effect for condition on commission error ($p < 0.05$). Subjects overall showed more errors under load condition compared to idle condition. There was not a significant main effect for group.

6.5. Conclusion and discussion

In this contribution we have presented the implementation and empirical evaluation of two wearable reaction time test designs. We have implemented wearable watch-like devices, which combine the generation of haptic stimuli with the recognition of hand gestures as subject's response. For the evaluation of the wearable interfaces, we have conducted an experiment to investigate to what extent we can measure changes in length and variability of user's reaction time with the wearable interfaces in comparison to a well accepted, traditional desktop-based test.

In total we have investigated three research questions. First, we have analyzed whether mean reaction times, variability and accuracy differ between the induced idle and load conditions. Since we were interested in whether our wearable user interface is sensitive to measure changes in reaction times occurred due to altered cognition, we induced cognitive load on subjects to alter their cognitive state. The cognitive load is one of the factors that influences the duration and variability of reaction times. As mentioned in related work, altered cognition could also be mild cognitive impairment, dementia or Alzheimer's disease. Since in our experiment healthy subjects participated, an additional cognitive task was used to simulate an alteration in subject's cognitive efficiency. The results show that during load condition mean reaction time and variability are significantly increased for both settings (desktop vs. wearable) and both groups (one-handed vs. two-handed). This first result shows us that our wearable user interface is suitable to measure changes in reaction times occurred due to one of the factors affecting human reaction time. Second, we have investigated whether there are differences between desktop and wearable setting. We observed significant faster reaction times for both wearable settings under the idle condition. A possible explanation is the faster perception of haptic stimuli in comparison to visual stimuli. Third, we have investigated differences between the two wearable reaction time variants. The only difference between both experimental groups was a significant faster mean reaction time for the two-handed design under idle condition. A possible explanation is that differentiating between target and non-target on the same hand is more difficult than differentiation of target and non-target on two hands. This result shows us that both variants were appropriate to measure changes in reaction times but the one-handed variant was more complex. Thus, depending on the experiment design, one could use the one-handed design to perform more complex reaction time test such as go/no-go task, which needs more concentration to differentiate between two similar stimuli. Two-handed design is less complex because the discrimination of the two stimuli is easier. Based on the achieved results, we finally conclude that both presented implementations of wearable reaction time tests are suitable to measure factors that influence length and variability of reaction times.

6.6. Future work

In our future work, we will investigate long-term measurements of reaction times throughout daily life as cognitive performance indicators. We will collect reaction time data from several subjects during daily life activities. Since an interrogation every two seconds would not be acceptable during normal life activities, the time interval between two stimuli has to be increased. Up to now we collected reaction times from one subject during 8 h with random intervals varying between 60 and 90 s which results in 384 reaction times. Based on the collected reaction times the minimum number of reaction times will be defined by analyzing the correlations between reaction times and self-reported items.

Since different activities can also affect the reaction times, we will investigate the variations in reaction times occurred due to different daily life activities. In addition to reaction times, we will collect self-assessments such as concentration level, alertness, tiredness, etc. throughout the day and we will investigate the relation between reaction times and self-reported items on cognitive efficiency.

In a second study, we want to employ our wearable reaction time test on the elderly with cognitive impairment. The main goal is to use the reaction time test continuously as a secondary task while performing activities of daily living. Cognitive impairment or attentional deficit is reflected in the performance of daily activities among elderly people. A decrement on reaction time performance could provide us a measure of attentional resources needed to perform that particular daily activity. Another possible employment could be for patients after a stroke. Reaction times can be collected while the patients perform different motor skills during a rehabilitation program. An increase in reaction time performance (more attentional resources are allocated for the cognitive performance task) could indicate that the patient needs to allocate fewer attentional resources to maintain the motor skills and that the rehabilitation has made progress.

In addition, we are planning to extend the implementation of wearable reaction times by adding context-awareness in order to provide a context-aware generation of stimuli events.

Bibliography

[1] Brain Workshop - a Dual N-Back game. http://brainworkshop.sourceforge.net/.

[2] Pebl: Psychological test battery. http://pebl.sourceforge.net/.

[3] E. R. Braverman, A. L. Chen, T. J. Chen, J. D. Schoolfield, A. Notaro, D. Braverman, M. Kerner, S. H. Blum, V. Arcuri, M. Varshavskiy, U. Damle, B. W. Downs, R. L. Waite, M. Oscar-Berman, J. Giordano, and K. Blum. Test of variables of attention (TOVA) as a predictor of early attention complaints, an antecedent to dementia. *Neuropsychiatr Dis Treat*, 6(1):681–690, 2010.

[4] H.-Y. Chen, J. Santos, M. Graves, K. Kim, and H. Z. Tan. Tactor localization at the wrist. In *Proceedings of the 6th international conference on Haptics: Perception, Devices and Scenarios*, EuroHaptics '08, pp. 209–218, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-69056-6.

[5] R. W. Cholewiak and A. A. Collins. Vibrotactile localization on the arm: effects of place, space, and age. *Percept Psychophys*, 65 (7):1058–1077, Oct 2003.

[6] B. Cinaz, C. Vogt, B. Arnrich, and G. Tröster. A wearable user interface for measuring reaction time. In *Ambient Intelligence*, vol. 7040 of *Lecture Notes in Computer Science*, pp. 41–50. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-25166-5.

[7] M. F. Folstein, S. E. Folstein, and P. R. McHugh. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12(3):189–198, Nov 1975.

[8] G. B. Forbes. Clinical utility of the Test of Variables of Attention (TOVA) in the diagnosis of attention-deficit/hyperactivity disorder. *J Clin Psychol*, 54(4):461–476, Jun 1998.

[9] E. Gorus, R. De Raedt, M. Lambert, J. C. Lemper, and T. Mets. Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. *J Geriatr Psychiatry Neurol*, 21(3):204–218, Sep 2008.

[10] H. Harms, O. Amft, R. Winkler, J. Schumm, M. Kusserow, and G. Tröster. ETHOS: Miniature orientation sensor for wearable human motion analysis. In *Sensors, 2010 IEEE*, pp. 1037 –1042, nov. 2010.

[11] A. Ivorra, C. Daniels, and B. Rubinsky. Minimally obtrusive wearable device for continuous interactive cognitive and neurological assessment. *Physiol Meas*, 29(5):543–554, May 2008.

[12] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[13] A. R. Jensen. *Clocking the mind: Mental chronometry and individual differences.* Elsevier, 2006.

[14] S. Kaiser, A. Roth, M. Rentrop, H.-C. Friederich, S. Bender, and M. Weisbrod. Intra-individual reaction time variability in schizophrenia, depression and borderline personality disorder. *Brain and Cognition*, 66(1):73 – 82, 2008. ISSN 0278-2626.

[15] R. J. Kosinski. A literature review on reaction time, August 2009.

[16] S. C. Lee and T. Starner. BuzzWear: alert perception in wearable tactile displays on the wrist. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pp. 433–442, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9.

[17] H. R. Lieberman, F. M. Kramer, S. J. Montain, and P. Niro. Field assessment and enhancement of cognitive performance: Development of an ambulatory vigilance monitor. *Aviation, Space, and Environmental Medicine*, 78(5, Suppl.):B268–75, May 2007.

[18] R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford University Press, 1986.

[19] M. Matscheko, A. Ferscha, A. Riener, and M. Lehner. Tactor placement in wrist worn wearables. In *14th annual IEEE International Symposium on Wearable Computers (ISWC'10)*, pp. 1 –8, 2010.

[20] R. Mistur, L. Mosconi, S. D. Santi, M. Guzman, Y. Li, W. Tsui, and M. J. de Leon. Current Challenges for the Early Detection of Alzheimer's Disease: Brain Imaging and CSF Studies. *J Clin Neurol*, 5(4):153–166, Dec 2009.

[21] I. Oakley, Y. Kim, J. Lee, and J. Ryu. Determining the feasibility of forearm mounted vibrotactile displays. In *Proceedings of the Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, HAPTICS '06, pp. 27–34, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 1-4244-0226-3.

[22] U. Scholz, R. L. Marca, U. M. Nater, I. Aberle, U. Ehlert, R. Hornung, M. Martin, and M. Kliegel. Go no-go performance under psychosocial stress: Beneficial effects of implementation intentions. *Neurobiology of Learning and Memory*, 91(1):89 – 92, 2009. ISSN 1074-7427.

[23] P. T. Trzepacz and R. W. Baker. *The psychiatric mental status examination*. Oxford University Press, New York :, 1993. ISBN 0195062515.

[24] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky. Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands. *Neuropsychologia*, 47(12):2389 – 2396, 2009. ISSN 0028-3932.

[25] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[26] K. Wild, D. Howieson, F. Webbe, A. Seelye, and J. Kaye. Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's and Dementia*, 4(6):428 – 437, 2008. ISSN 1552-5260.

7

Collecting Reaction Times in Daily Life

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

Full publication title: A Case Study on Monitoring Reaction Times with a Wearable User Interface during Daily Life

International Journal of Computers in Healthcare, Inderscience, 1(4),pp 283-303, 2012

Abstract

Reaction time tests are known as sensitive tests for measuring cognitive efficiency, cognitive decline, early attention complaints, and memory impairments. A main drawback of existing computer-based reaction time tests is their incompatibility with daily life. As a consequence, it is not feasible so far to assess variations in cognitive efficiency which are caused by influencing factors occurring in daily life. In order to overcome this limitation, in recent work, we have designed and implemented a wearable watch-like reaction time test tool. In this contribution, we present a case study when employing the wearable reaction time test in the work environment of a graphic designer. We show that our tool allows to continuously collect reaction times throughout 15 working days. Besides collecting reaction times, we employed an experience sampling method to gather subjective data on mood and perceived workload. The correlation analysis showed that mean reaction time is significantly positive correlated with mental and temporal demand. High level of arousal results in slowing of reaction times and higher variability whereas a low arousal results in lower variability of reaction times.

7.1. Introduction

Reaction time tests measure how rapidly stimuli information can be processed and a response to it can be activated [18]. In other words, reaction time tests measure the elapsed time between the occurrence of a stimulus and the individual's response to it. Since the middle of the 19th century, reaction time has been extensively investigated by experimental psychologists. In the last decades, several computer-based reaction time tests have been developed in which a subject has to respond to visual stimuli by using keyboard, mouse or special buttons. According to Jensen [13], reaction time tests are well suited for cognitive assessment tests since reaction time tests offer a high sensitivity for detecting variation in cognitive efficiency and they can be repeated virtually an unlimited amount of times. An extensive review about computer-based cognitive tests can be found in Wild et al. [27]. There are several examples on how reaction time tests are applied to assess variations in cognitive functioning, such as detection of cognitive decline or determination of the ability to manage complex activities like driving, piloting, search and rescue. For example, Kosinski [14] reported that cognitive impairment influences length and variability of reaction times. In Gorus et al. [8], it was shown that persons with cognitive deterioration demonstrated more intra-individual performance variability and more slowing in their reaction time than cognitively healthy elderly. Further, Braverman et al. [3] showed that a reaction time test is an accurate predictor of early attention complaints and memory impairments.

The main drawback of existing computer-based reaction time tests is their incompatibility with daily life since subjects have to interrupt their daily routine for several minutes and provide their full attention in order to perform the reaction time test. This restriction prohibits the measurement of cognitive efficiency during daily routine tasks. As a consequence, it is not feasible so far to assess variations in cognitive efficiency which are caused by influencing factors occurring in daily life like variations in cognitive load during work or activities of daily life. In order to overcome this limitation, in previous work, we have designed and implemented a wearable watch-like reaction time test tool which allows to collect reaction times throughout daily life. Instead of a visual stimulus a haptic stimulus is employed from time to time and instead of a keyboard response the user is able to react with a simple hand movement. Haptic stimulus was chosen since the visual and auditory channel is often already occupied for everyday life tasks. User's hand movement response is automatically recognized with an integrated inertial measurement unit (IMU) and thus the reaction time to the haptic stimulus can be measured.

In order to show to what extent the wearable reaction time test is an accurate new tool to measure reaction times, in recent work we have conducted an experiment in which we analyzed the reaction times of 20 subjects in an idle condition and under cognitive load [6]. During the idle condition, subjects had to perform the wearable reaction time test followed by operating a state of the art desktop-based reaction time test. In order to influence length and variability of reaction times, we applied cognitive load to all subjects in a separate load condition. Here, the subjects had to solve mental tasks in addition to performing the reaction time test. Like in the first idle condition, subjects had to perform the wearable reaction time test followed by operating a state of the art desktop-based reaction time test. The results showed that the mean reaction times of the subjects were always increased during the load condition compared to the idle condition in wearable and desktop settings. Second, it could be observed that the variability of reaction times was increased during the load condition in both wearable and desktop settings. Based on the achieved results, we could show that the wearable reaction time test is an accurate tool to measure variations in length and variability of reaction times.

In this contribution, we extend our work by transferring our approach from controlled laboratory settings to an unrestricted daily life environment. In a case study, we present our methods and findings when employing the wearable reaction time test tool in the daily work life of a graphic designer. We continuously collected reaction time data and self-experience ratings on mood and workload throughout 15 working days. The first three working days were characterized by normal job demands, which we denote as “baseline” measurement. During the remaining 12 days, the designer was confronted with four workload factors that are common when a designer has to deliver a creative outcome in a limited time: stress, sleep deprivation, night shift, and moderate alcohol consumption.

In this case study, we target four main topics:

1. continuously monitoring of reaction times in a real world working setting without interrupting the daily routine of a graphic designer
2. investigating the observed variations in length and variability of reaction times regarding the four workload factors
3. investigating the correlations between reaction time features and perceived workload
4. investigating the correlations between reaction time features and subjective ratings on mood.

In the following, we first present related work. Next, we describe our wearable reaction time test implementation. Afterwards, we explain the real life employment followed by data analysis methods. Finally, we present and discuss the achieved results and provide a conclusion and an outlook.

7.2. Related work

There exist mainly three kinds of reaction time tests: simple, recognition and choice reaction time tests [14]. Simple reaction time tests consist of one stimulus and one response. In desktop-based simple reaction time tests, a dedicated stimulus is temporary shown at random time intervals on the screen. The subject has to concentrate on the appearance of the stimulus and every time the subject recognizes the stimulus he has to press a dedicated key as fast as possible. Differently, in recognition reaction time tests, the subject has to differentiate between a dedicated stimulus (target) and other stimuli (non-target). This is sometimes called “go/no-go” reaction time task. In desktop-based go/no-go tests the target stimulus and the non-target stimuli are temporary shown in a random sequence on the screen. The subject has to be attentive to the appearance of the target stimulus and ignore the non-target stimuli. As soon as the subject recognizes the target, he has to respond as fast as possible by pressing a dedicated key. Lastly, choice reaction time tests include multiple stimuli and multiple responses. The subject has to respond to each stimulus with a corresponding response, e.g. by pressing a certain key whenever a corresponding letter appears on the screen. A detailed series of recommendations on how to conduct laboratory experiments using reaction time tests and how to analyze the collected data can be found in [13, 18, 26].

Since reaction time has been extensively investigated for many years, there exists a broad variety of clinical studies which have been summarized in a literature review by Kosinski [14]. For example, it is known that age-related diseases like cognitive impairment are important factors which influence length and variability of reaction time. Gorus et al. [8] showed that persons with cognitive deterioration demonstrated more intra-individual performance variability and more slowing in their reaction time than cognitively healthy elderly. Braverman et al. [3] showed in a clinical setting that a go/no-go reaction time test is an accurate predictor of early attention complaints and memory impairments. In [24], it was shown that children diagnosed with attention deficit hyperactive disorder (ADHD) are characterized by an increased variability in reaction time. In [22], the effect of psychosocial stress on reaction time was investigated in a controlled experimental setting. The authors were able to show that participants under stress revealed slower reaction time.

Most of the studies have in common that desktop-based reaction time tests were employed in laboratory settings which require the full attention of the subject for several minutes. Hence, it is not feasible so far to measure reaction times during everyday-life in order to assess variations in cognitive efficiency which are caused by influencing factors occurring in daily life like variations in cognitive load during work or activities of daily life. Only a few studies exist which investigated the feasibility of measuring reaction times during everyday activities. Lieberman et al. [17] implemented a wrist-worn reaction time device to assess vigilance. Visual and auditory stimuli were generated with three LEDs and a miniature speaker. The user had to react by pressing one of the two push buttons on the device. Ivorra et al. [11] implemented a haptic stimulus to interrogate the central nervous system in a minimally obtrusive way. As the response the detection of a wrist movement is defined. In a first feasibility study, they showed that a simple reaction time test can be continuously administered throughout the course of normal life activities. However, an evaluation of the wearable implementation in comparison with state of the art desktop-based reaction time tests is missing and the obtained reaction times were not further analyzed. In our recent work [5], we followed the approach of Ivorra et al. [11] and designed a wearable reaction time device which combines the generation of haptic stimuli and the recognition of forearm rotation as subject's response.

7.3. Wearable reaction time test

7.3.1. System design

In our previous work, we have designed and implemented a watch-like wearable user interface to measure reaction time [5]. The main design goal was to enable a mobile measurement of reaction time throughout daily life. In order to achieve this goal, we replaced the state of the art visual stimuli by haptic stimuli which can be recognized without paying continuous attention to it. Second, we replaced the state of the art push buttons by detecting a dedicated rotation of the forearm as response to the stimuli. As a result the wearable reaction time test consists of two main modules: a stimulus module generating haptic stimuli and an IMU module detecting forearm rotations. We applied the haptic stimuli on the wrist since according to the literature the wrist is a recommended stimulus site for wearable tactile displays [4, 16, 19, 20]. For generating vibro-tactile stimuli, we used a coreless mini DC vibration motor. In order to maximize the vibration amplitude and to ensure a proper sense of the vibration, we placed the motor in a separate plastic enclosure. In order to guarantee a continuous operation during at least one day, the stimuli module has its own battery supply. In addition, we

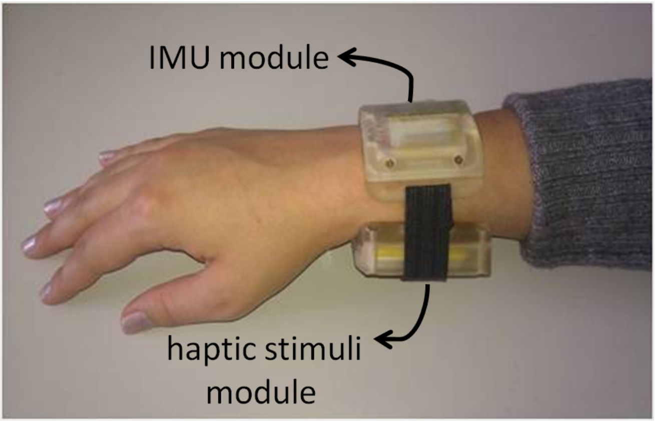

Figure 7.1. Wearable implementation of a simple reaction time test. The haptic stimuli are generated as vibrations on the lower side of the wrist. The IMU is placed on the opposite side to recognize the user's hand gesture responses.

integrated an audio driver in order to allow the generation of auditory stimuli through headphones. The IMU module consists of the so-called ETHOS which includes a three-axis accelerometer and gyroscope to recognize the subject's gesture response [9]. The accelerometer and gyroscope is sampled with a frequency of 128 Hz. An implementation of a wearable simple reaction time test can be seen in Figure 7.1. The detailed description of the hardware platform can be found in Cinaz et al. [5] and Harms et al. [9].

We modified the firmware version of the ETHOS to control the haptic stimuli module. Similar to the wearable reaction time test presented in Ivorra et al. [11], we have defined the response gesture as a fast rotation of the forearm.

In our recent work, we have implemented a wearable go/no-go task [6]. In order to evaluate the wearable interface, we have conducted an experiment to compare the reaction times obtained by the wearable interface with those collected by a desktop-based reaction time test. For the desktop-based reaction time test, we used a free version of the go/no-go TOVA test [2, 7]. As exemplary shown in Figure 7.2, a white square appears for 100 ms on the screen, with a black square within it. Participant must respond only to targets (the black square on top)

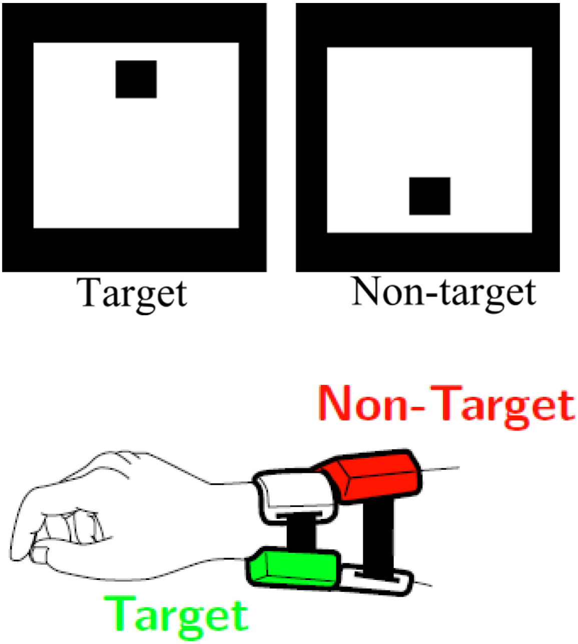

Figure 7.2. (a) Target and non-target stimulus types of the TOVA test (b) implementation of a wearable TOVA test

and ignore the non-targets (the black square on the bottom). Each stimulus is presented at 2000 ms intervals. A wearable implementation of the TOVA test is shown in the lower part of Figure 7.2. Participant must respond only if they perceive a haptic stimulus on the lower side of the forearm (target) and ignore the haptic stimulus on the upper side of the forearm (non-target). Similar to the desktop-based reaction time test, each haptic stimulus of the wearable device is generated for 100 ms at 2000 ms intervals.

Each setting (desktop-based vs. wearable) consisted of two experimental conditions:

1. idle condition in which the subject is just performing the reaction time test
2. cognitive load condition in which the subject is performing a cognitive performance task in parallel to the reaction time test.

As cognitive task, we employed a variant of the N-Back test, the so-called “Audio 2-Back” [1, 12]. In the Audio 2-Back task, a letter is presented to the subject via an audio message in regular time intervals. The subject must only respond if the currently pronounced letter is the same as the one that was pronounced two positions back. Since the subject has to memorize the two prior letters and has to perform a comparison with the current letter in time, this task induces additional cognitive load. Whenever a sound match occurs, the subject has to say “match”. In order to keep the subject engaged in the second task, the experiment leader continuously controls the answer of the subject and provides direct feedback to the subject about correct and false answers. A total of 20 subjects participated in our experiment. Each subject performed four different reaction time tests in the following order:

1. desktop-based in idle condition
2. desktop-based under cognitive load
3. wearable in idle condition
4. wearable under cognitive load.

Each condition lasts 10 minutes and contains 320 stimuli (160 targets and 160 non-targets).

The results showed that subjects were always significantly slower during the cognitive load condition compared to the idle condition for both desktop and wearable settings ($p < 0.001$). Second, we could observe that the variability is increased during the load condition for both settings. Subjects demonstrated higher variability under load condition compared to idle condition ($p < 0.001$). Based on the achieved results, we concluded that our wearable reaction time test is suitable to measure factors which influence length and variability of reaction times.

7.4. Case study

In this case study, we employed the wearable reaction time test tool in the daily work life of a graphic designer. Since this case study was the first attempt to collect reaction times in real life, we opted for collecting reaction times with a moderate high sampling rate. We decided to

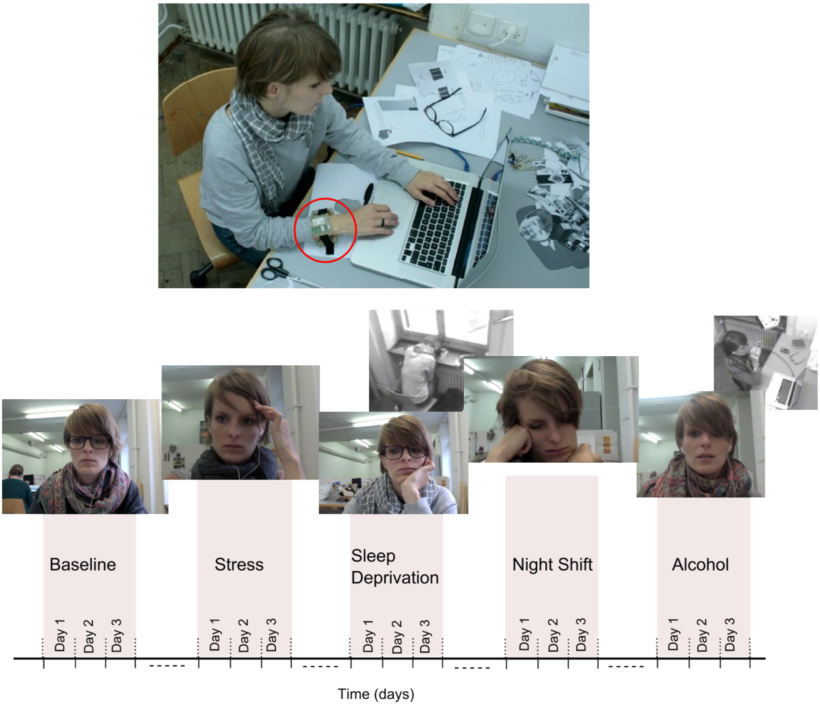

Figure 7.3. (a) Subject in her workplace, the reaction time module is placed on her dominant wrist (b) experiment procedure and screenshots taken during different factors

collect at least 20 reaction time measurements within a 30 minutes period. As a consequence, we generated haptic stimuli at random intervals varying between 60 sec and 90 sec.

Our test subject was a 25-year-old female master student of graphic design. We conducted the case study while our subject was working on her master thesis. The topic of her master thesis was to investigate the influence of typical work-related stress factors on the design process. In total, we monitored our subject throughout 15 working days. The first three working days were characterized by normal job demands, which we denote as “baseline” measurement. During the remaining 12 days, the designer was confronted with four workload factors that are common when a designer has to deliver a creative outcome in a limited time: stress, sleep deprivation, night shift, and moderate alcohol consumption. The case study pursues four main aims:

1. to continuously collect reaction times in a real-world working setting without interrupting the daily routine of a graphic designer
2. to investigate the observed variations in length and variability of reaction times regarding the four workload factors
3. to correlate the observed reaction time features with perceived workload ratings
4. to investigate the correlations between reaction time features and subjective ratings on mood.

In the following, we first explain the induced workload factors in more detail. Next, we describe the employed experience sampling method to collect subjective ratings on perceived workload and mood.

7.4.1. Workload factors

The subject was monitored over 15 working days during 6 weeks resulting in 120 hours of data (8 h x 15 days). The first three working days consisted of normal job demands, which we denote as "baseline" measurement. During the remaining 12 days, the designer was confronted with four workload factors that are common when a designer has to deliver a creative outcome in a limited time: stress, sleep deprivation, night shift, and moderate alcohol consumption. Each factor was studied on three consecutive days. The experimental procedure can be seen in Figure 7.3.

7.4.2. Baseline

The first three days were carried out as a baseline session with a typical job assignment. At the beginning of the first working day, the subject was instructed from a client to design a logo for the new established Pirate Party within three days. According to the subject's working experience, such an assignment represents a common job demand which can be easily solved within three working days. The subject was allowed to work only during the monitoring period. At the end of the third working day, she had to provide the resulting logo to her client.

7.4.3. Stress

After a break of two days, the next three working day session was carried out. At the beginning of the first working day, the subject received from the public relation manager of the Department of Information Technology and Electrical Engineering at ETH Zurich the instruction to design a new marketing booklet to attract new students. The subject was allowed to work on this task for three days. Every evening she had to provide a draft design for revision. The subject was not aware that the public relation manager was instructed by psychologists on how to create high working pressure and stress within the three days. In particular, the public relation manager was instructed to induce stress as follows:

- *Negative feedback:* The feedback on every draft design should illustrate that the client was not satisfied and that there were many open issues to be solved in the remaining time.
- *Information hiding:* Important background information like design guidelines were not given at the beginning but at the second day.
- *Work overload:* The concrete task description was slightly modified every day, e.g. the subject was told that she should especially focus on female students. In addition, the subject was asked to prepare additional material like a presentation in parallel to the normal work.
- *Social-evaluative pressure:* The subject was told that all professors of the department will evaluate her design.
- *Ego-involvement:* The subject was told that if her suggested design was satisfying to the professors, it would be selected as the new official marketing booklet of the department.

7.4.4. Partial sleep deprivation

After a break of one week, the subject was monitored during partial sleep deprivation. On the first day, the subject started to work after a night of seven hour normal sleep duration. The subject was instructed from a client to design T-shirt for the internet portal “evernote”. In the remaining two nights, the sleep duration was reduced to four hours.

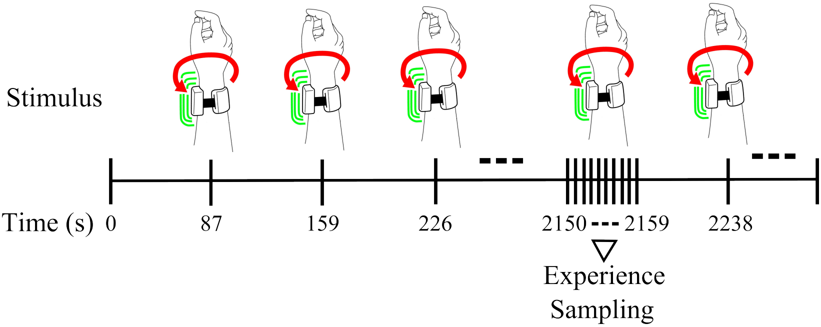

Figure 7.4. We programmed an experience sampling event in the reaction time module to signal the subject when to complete the questionnaire. Every time ten consecutive haptic stimuli (at 1 second intervals) occurred, the subject had to fill out the questionnaire items.

7.4.5. Night shift

After a break of three days, the night shift session was carried out. The subject worked between midnight and 8 AM and was sleeping during daytime. At the beginning of the first working day, the subject was instructed from a client to design a new concept for the German internet blog “Sichtweise Fotografie”.

7.4.6. Alcohol

After a break of two weeks, the next three working day session was carried out. During this condition, the subject worked between normal office hours ($\sim$ 10 AM to $\sim$ 6 PM). In the afternoon she drank 2-3 glass of wine. At the beginning of the first working day, the subject was instructed from a client to design a new logo for the Institute Human Computer Interaction of Konstanz University.

7.4.7. Experience sampling

In order to examine the relationship between reaction time data and self-reported ratings on mood and workload, we used an experience sampling methodology (ESM) known from [15, 23]. The subject was randomly probed eight times per day to complete a set of questionnaires. The minimum and maximum time interval between successive questionnaires was set to 30 and 120 minutes, respectively. The subject was prompted to fill out the questionnaires by applying ten consecutive haptic stimuli at 1 second intervals. It took approximately 2-3 minutes to complete all questionnaire items. The experience sampling approach is depicted in Figure 7.4. At each probe, we assessed subject's current mood with the presentation of 15 adjectives on a visual analogue scale (from 0 = not at all to 10 = a great deal) using a short version of the positive and negative affect schedule [25]. The 15 items consisted of seven negative ("bored", "tired", "stressed", "anxious", "angry", "depressed", and "nervous"), seven positive ("relaxed", "happy", "concentrated", "alert", "interested", "active", and "excited“) and one sleepiness item ("sleepy").

In addition, the subject was asked to assess her perceived workload regarding the last working activities. For this subjective rating, we employed the NASA task load index (TLX) from [10]. First, the subject has to rate 6 items on a scale from 1 to 20 that best indicate her 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 is systematically asked which of the items represents the more important contributor to the workload. Based on these comparisons, the total workload is computed as a weighted average of the ratings.

All questionnaire items were completed on the subject's computer. Beside the randomly generated probes, the subject filled the questionnaire also at the beginning of each day before starting to work.

7.5. Data analysis

In the following, we first provide a description of our methods to automatically detect user's hand movement as response to the applied haptic stimuli. Next, we describe how the reaction time measures were computed. Finally, we explain how we correlated the reaction time features with the self-experience data on mood ratings and NASA workload items.

In order to detect subject's hand response to a stimulus, we analyzed the gyroscope data obtained by the IMU module. First, the gyroscope data is normalized to values between -1 and 1. Second, the raw gyroscope data is smoothed using a simple moving average filtering. Then, a window of 2 seconds length is aligned on each stimulus event and the local maximum point is located within this window. If the maximum

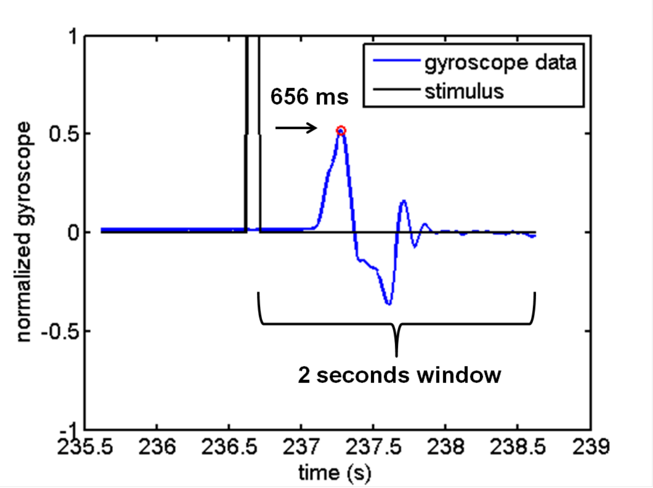

Figure 7.5. X-component (wrist-turn axis) of the normalized gyroscope data while responding to a haptic stimulus with a fast rotation of the hand. After each stimulus event the maximum peak within a fixed time window is searched. Reaction time is calculated as the time difference between the haptic stimulus and the occurrence of the peak response.

peak is larger than a decision threshold, it is classified as a correct reaction gesture; otherwise it is classified as a non-reaction, i.e. it is assumed that the user has not responded to the stimulus. The normalized gyroscope data while responding to a haptic stimulus is exemplary shown in Figure 7.5.

The decision threshold was estimated by applying a Naive Bayes classifier. In a first step, we selected a total of 120 responses and 120 non-responses by a visual inspection of the gyroscope data. This data was then used to define a decision threshold using the Naive Bayes classifier. After applying Naive Bayes, a visual inspection of 2304 reaction hand gestures revealed an accuracy of 99%.

7.5.1. Windowing for experience sampling

In order to investigate the relation between reaction times and self-rated mood and workload items, we used data segments with varying window sizes. The window size is calculated by taking the midpoint of

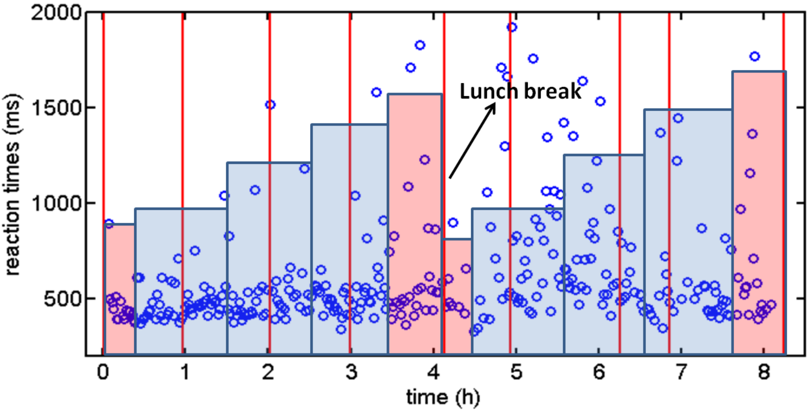

Figure 7.6. Data segments used for the correlation analysis. The reaction time features are calculated for each segment around the experience sampling events (blue segments). The calculated features are then correlated with the corresponding self-report of this segment. The red segments are ignored because of insufficient number of reaction times.

the time between the last and the preceding questionnaire. Figure 7.6 exemplary depicts the segments used for the correlation analysis. The reaction time features are calculated for each segment around the experience sampling events (blue segments in Figure 7.6). The calculated features are then correlated with the mood and perceived workload ratings of this segment. The first and the last segment of every four hour session (red segments in Figure 7.6) are not used in the correlation analysis since the amount of data here is only half as much as in the other segments. For each segment, the following features are calculated: mean reaction time, standard deviation of reaction time and number of omission errors defined as the number of missing responses to the stimuli.

7.6. Results and discussion

Our first goal was to show that a continuous collection of reaction times in a real-world working setting is feasible without interrupting the daily routine. In the presented case study, we were able to show that even when employing a moderate high sampling frequency (60-

90 s stimuli interval) our wearable reaction time test tool allows to continuously collect reaction times throughout 15 working days of a graphic designer. Our test subject was able to successfully complete all job demands over the complete monitoring period while performing the wearable reaction time test in parallel without interrupting her daily working routine. Hence, we were able to present a successful example showing that our wearable reaction time test tool is appropriate to continuously collect reaction times in real world working settings.

In the following, we present the results of our remaining three goals: variations in length and variability of reaction times regarding the four workload factors, correlations of reaction time features with perceived workload ratings, and correlations between reaction time features and subjective ratings on mood.

7.6.1. Reaction time measures under workload

Figure 7.7 shows mean, standard deviations and omission errors of the reaction times for all five working conditions under investigation. The measures were computed in time windows around the questionnaire probe as described in Section 7.5.1. In Figure 7.7, it can be observed that the variability of reaction times tends to be higher in the “stress” condition compared to the “baseline” condition. During the “baseline” condition mean reaction times on the first two days do not vary to a great extent. In the third day, mean reaction times are higher in the afternoon. A sudden increase of mean reaction times in the afternoon can be observed for the first two days of the “stress” condition. This might be explained by the fact that our subject received the negative feedback and additional tasks from the client at the end of the morning session. In the last day of the “stress” condition, the mean reaction times and the variability are decreased during the course of the day. A possible explanation is the fact that our subject was quite confident on the 3rd day to deliver all requested designs to the client. The number of omission errors during the “stress” condition is almost 3 times higher than during “baseline” ($N_{baseline} = 35$, $N_{stress} = 113$ where N indicates the total number of omission errors). The trend of mean reaction times during “sleep deprivation” is similar to “baseline”. A possible explanation is the fact that our subject was used to work with reduced amount of sleep during her study time. The mean reaction times during “night shift” and “alcohol” are also comparable to the one obtained during “baseline”. However, the standard deviation of reaction times

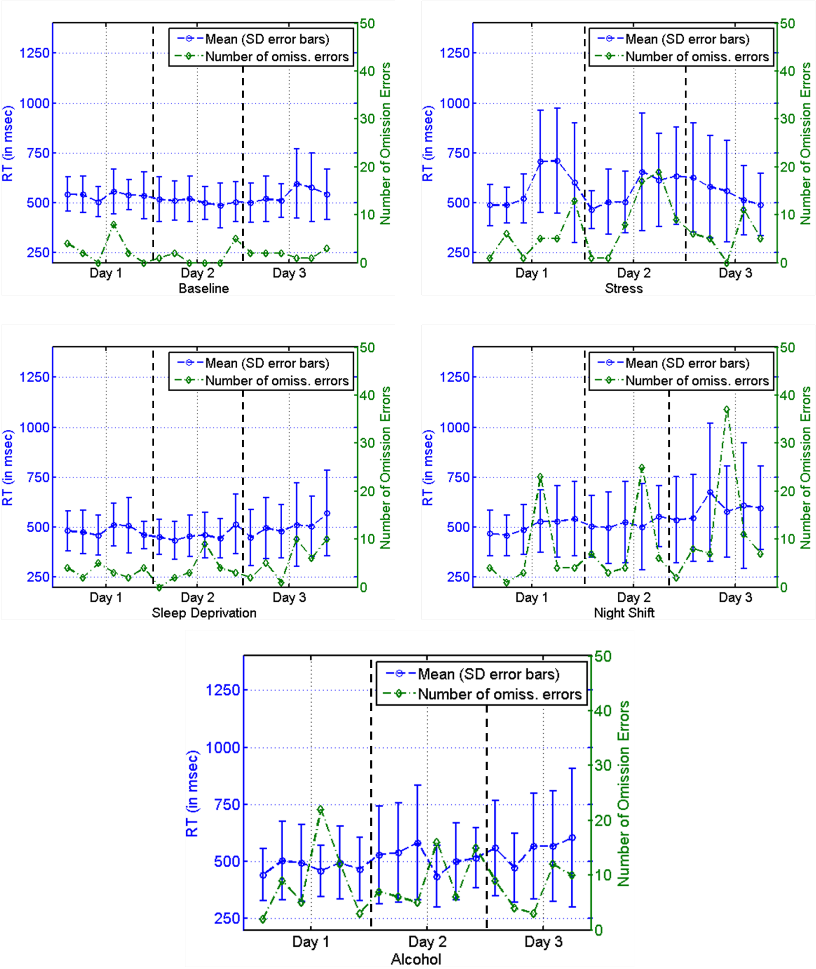

Figure 7.7. Mean, standard deviations and omission errors (missing responses) of reaction times over 15 working days

for all three conditions is always increased in comparison to “baseline”. Especially during “night shift” we can observe a steady increase on both mean reaction times and variability during the course of three days. The highest omission errors can be observed for the “night shift” and “alcohol” conditions ($N_{nightshift} = 156$, $N_{alcohol} = 146$).

These highest omission errors during night shift are caused by falling asleep from time to time. This was visible in the continuous video recording obtained from our subject’s working place. In general, an increase in the errors can be observed especially during the second half of each working day which is in line with a reduction of the concentration level and with an increase of sleepiness.

Figure 7.8 shows the boxplots of mean, standard deviation and omission errors of reaction times for each workload factor. ANOVA with pairwise comparison revealed that mean reaction times of the “stress” condition were significantly higher than those collected during “sleep deprivation” and “alcohol” conditions ($p < 0.05$). Besides, “sleep deprivation” condition showed a significant lower mean reaction time compared to “night shift” ($p < 0.05$). The mean reaction times of the “baseline” condition did not show a significant difference from other conditions. There were significant group differences in variability of reaction times for each of the conditions. Pairwise comparison showed a significant increase in variability during “stress”, “night shift” and “alcohol” compared to “baseline” ($p < 0.05$). A significant difference between “baseline” and “sleep deprivation” as well as between “stress”, “night shift” and “alcohol” was not observed. According to the pairwise comparison of omission errors, the subject showed significantly higher number of omission errors during “night shift” and “alcohol” conditions compared to “baseline” condition ($p < 0.05$). However, a significant difference was not observed for the “stress” and “sleep deprivation” conditions.

In summary, we can conclude that our test subject showed a significant increased variability of reaction times during stress in comparison to baseline. Sleep deprivation did not show a significant difference to baseline. This might originate from the fact that our subject was used to work under reduced amount of sleep during her study time. Hence, we could not observe significant differences of any reaction time measure for the “sleep deprivation” condition in comparison to the baseline session. Compared to “baseline” condition, our subject showed a significant increased variability of reaction times and a significantly higher number of omission errors during “night shift”. The high number of omission errors during night shift are caused by falling asleep from time to time as evident from the video recording obtained from our subject’s working place. Similar to “night shift”, our subject showed under moderate alcohol consumption a significant increased variability

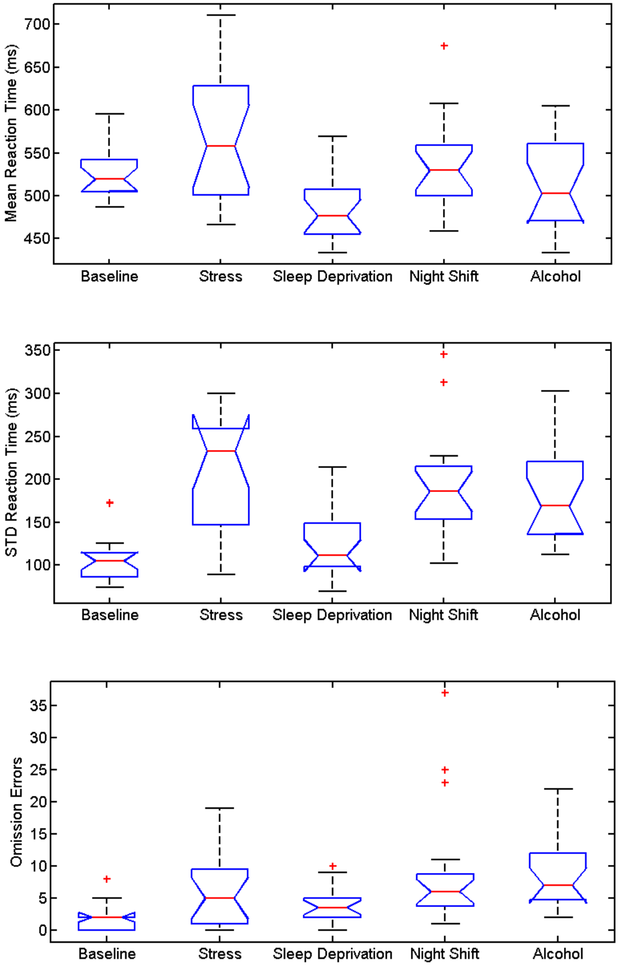

Figure 7.8. Boxplots of mean, STD reaction times and omission errors for each workload factor.

of reaction times and a significantly higher number of omission errors in comparison to “baseline”.

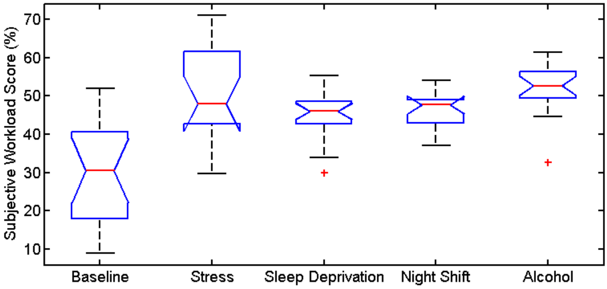

Figure 7.9. Boxplot of total subjective workload score for each workload factor

7.6.2. Correlation of reaction time measures with perceived workload

In order to perform the correlation analysis, for each day we analyzed the data from six questionnaire probes (except three days where only five questionnaires were available) as described in Section 7.5.1. With a total amount of 87 observations, the correlation for each questionnaire item was computed.

Figure 7.9 shows the box plot of perceived total workload scores for each working factor. The lowest total workload was perceived during the "baseline" condition as intended from the experiment design. ANOVA with pairwise comparison revealed that the mean total workload score of "stress", "sleep deprivation", "night shift" and "alcohol" conditions were significantly higher than those perceived during "baseline" ($p < 0.05$). However, there is not a significant difference between "stress", "sleep deprivation", "night shift" and "alcohol" conditions.

Table 7.1 shows the significant correlations between reaction time features and NASA workload items. Mean reaction time is significantly positive correlated with mental and temporal demand. Total workload showed a significant positive correlation with reaction time variability and number of omission errors. This finding is in line with the previous conducted laboratory evaluation: high cognitive load, i.e. mental demand and time pressure, resulted in slower reaction times and higher variability.

Table 7.1. The correlation between reaction time features and NASA workload items

NASA Items	*Mean RT*	*STD RT*	*Omissions*
Mental demand	*0.313***	−0.044	−0.156
Temporal demand	*0.414***	0.139	−0.087
Total workload	0.098	*0.359***	*0.235**

Correlation coefficient $**p < 0.01$, $*p < 0.05$

Table 7.2. The correlation between reaction time features and self-reported mood items

Mood items	*Mean RT*	*STD RT*	*Omissions*
High arousal			
Concentrated	*0.317***	0.140	−0.098
Alert	*0.413***	*0.237**	−0.016
Active	*0.343***	*0.231**	0.001
Stressed	*0.418***	*0.323***	0.114
Low arousal			
Bored	−0.047	*-0.373***	*-0.274**
Tired	−0.197	*-0.289***	−0.157

Correlation coefficient $**p < 0.01$, $*p < 0.05$

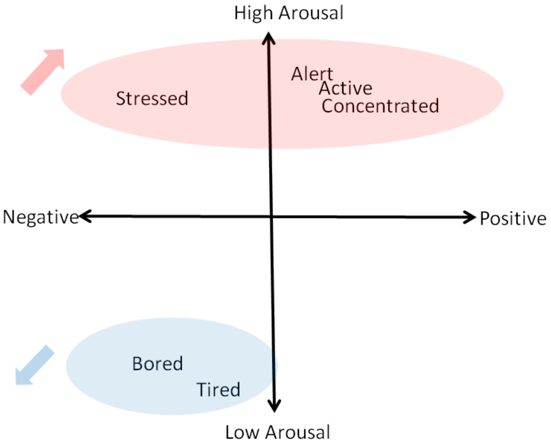

Figure 7.10. We placed the mood items which are significantly correlated with reaction time features on the valence-arousal plane. It can be observed that the most significant correlations can be found in the high-arousal plane.

7.6.3. Correlation of reaction time measures with subjective ratings on mood

Table 7.2 shows the correlation coefficients of mean reaction time, standard deviation and the number of omission error with questionnaire items on mood. As shown in Table 7.2, mean reaction time is significantly positive correlated with the high arousal items concentrated, alert, active and stressed. The reaction time variability is significantly positive correlated with alert, active, stressed and negative correlated with the low arousal mood items bored and tired. Number of omission errors show significant correlation only with the item bored. Figure 7.10 shows the significant correlated mood items on the well-known valence-arousal plane (see for example, [21]). It can be observed that a high level of arousal results in increasing of reaction times and higher variability whereas a low arousal results in lower variability of reaction times. The reason behind might be that during stress or high mental demand one is highly concentrated/alert/active on the primary task and less concentrated on the secondary task (in this case, the wearable reaction time test). Thus, it takes longer to perceive the stimulus and to react to it.

7.7. Conclusion and outlook

In this contribution, we have presented for the first time a long-term measurement and interpretation of reaction times during normal work life. We have employed a new wearable watch-like reaction time test tool which generates haptic stimuli and recognizes user's hand movements as response. In our recent work, we could show that our wearable reaction time test is suitable to measure factors which influence length and variability of reaction times in controlled laboratory settings. In this contribution, we have extended our work by transferring our approach from controlled laboratory settings to an unrestricted work environment of a graphic designer. We continuously collected reaction time data and subjective ratings on mood and workload throughout 15 working days. The first three working days consisted of normal job demands, which we denote as "baseline" measurement. In the remaining 12 days, the designer was confronted with four workload factors that are common when a designer has to deliver a creative outcome in a limited time: stress, sleep deprivation, night shift, and moderate alcohol consumption. In order to examine the relationship between reaction time data and self-reported ratings on mood and workload, we used an ESM where the subject was randomly probed eight times per day to complete a set of questionnaires. We have presented our methods to automatically detect user's hand movement as response to the applied haptic stimuli. First, a Naive Bayes classifier was used to accurately detect the response gesture from the smoothed gyroscope data obtained by the IMU module. Next, the elapsed time between a haptic stimulus and the individual's response was computed as reaction time. In order to investigate the relation between reaction times and self-reported mood items, we used data segments with varying window sizes. From the reaction time features calculated for each segment around the experience sampling events, we have computed the correlations with the corresponding self-report items.

We could observe that our test subject showed a significant increased variability of reaction times during stress in comparison to baseline. Sleep deprivation did not show a significant difference to baseline since our subject was used to work under reduced amount of sleep. Compared to baseline condition, our subject showed a significant increased variability of reaction times and a significant higher number of omission errors during night shift. The high number of omission errors during night shift were caused by the fact that our subject felled asleep from time to time. Similar to night shift, our subject showed under moderate alcohol consumption a significant increased variability of reaction times and a significant higher number of omission errors in comparison to baseline.

The correlation analysis of reaction time measures with perceived workload showed that mean reaction time is significantly positive correlated with mental and temporal demand. Total workload showed a significant positive correlation with reaction time variability and number of omission errors. This finding is in line with the previous conducted laboratory evaluation: high cognitive load, i.e. mental demand and time pressure, resulted in slower reaction times and higher variability.

The correlation analysis of reaction time measures with subjective ratings on mood showed that mean reaction time is significantly positive correlated with the high arousal items concentrated, alert, active and stressed while reaction time variability is significantly positive correlated with alert, active, stressed and negative correlated with the low arousal mood items bored and tired. The number of omission errors is significant negative correlated with the item bored. In the valence-arousal plane, it is obvious that a high level of arousal results in increasing reaction times and higher variability whereas a low arousal results in lower variability of reaction times.

Finally, we conclude that the presented case study shows for the first time that a wearable watch-like reaction time test tool enables a long-term measurement and interpretation of reaction times during normal work life activities. Thus, it becomes feasible to measure reaction times during everyday-life in order to assess variations in cognitive efficiency which are caused by influencing factors occurring in daily life like variations in cognitive load during work or activities of daily life.

In future work, we will investigate elderly people. We already started with monitoring reaction times of healthy elderly during daily life activities like reading, writing and walking. Next, in cooperation with psychologists we are planning to monitor reaction times of elderly with mild cognitive impairment during daily life activities. We will compare the achieved results with the ones obtained from healthy elderly in order enable a detection of mild cognitive impairment during daily life monitoring.

Acknowledgments The authors would like to thank our test subject Sonja Zagermann who allowed us to monitor her daily working routine throughout 15 days. In addition, the authors acknowledge the cooper-

ation with the public relation manager Andrea Holle from the ITET Department of ETH Zurich for inducing stress to our test subject.

Bibliography

[1] Brain Workshop - a Dual N-Back game. http://brainworkshop.sourceforge.net/.

[2] Pebl: Psychological test battery. http://pebl.sourceforge.net/.

[3] E. R. Braverman, A. L. Chen, T. J. Chen, J. D. Schoolfield, A. Notaro, D. Braverman, M. Kerner, S. H. Blum, V. Arcuri, M. Varshavskiy, U. Damle, B. W. Downs, R. L. Waite, M. Oscar-Berman, J. Giordano, and K. Blum. Test of variables of attention (TOVA) as a predictor of early attention complaints, an antecedent to dementia. *Neuropsychiatr Dis Treat*, 6(1):681–690, 2010.

[4] H.-Y. Chen, J. Santos, M. Graves, K. Kim, and H. Z. Tan. Tactor localization at the wrist. In *Proceedings of the 6th international conference on Haptics: Perception, Devices and Scenarios*, EuroHaptics '08, pp. 209–218, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-69056-6.

[5] B. Cinaz, C. Vogt, B. Arnrich, and G. Tröster. A wearable user interface for measuring reaction time. In *Ambient Intelligence*, vol. 7040 of *Lecture Notes in Computer Science*, pp. 41–50. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-25166-5.

[6] B. Cinaz, C. Vogt, B. Arnrich, and G. Tröster. Implementation and evaluation of wearable reaction time tests. *Pervasive Mob. Comput.*, 8(6):813–821, Dec. 2012. ISSN 1574-1192.

[7] G. B. Forbes. Clinical utility of the Test of Variables of Attention (TOVA) in the diagnosis of attention-deficit/hyperactivity disorder. *J Clin Psychol*, 54(4):461–476, Jun 1998.

[8] E. Gorus, R. De Raedt, M. Lambert, J. C. Lemper, and T. Mets. Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. *J Geriatr Psychiatry Neurol*, 21(3):204–218, Sep 2008.

[9] H. Harms, O. Amft, R. Winkler, J. Schumm, M. Kusserow, and G. Tröster. ETHOS: Miniature orientation sensor for wearable

human motion analysis. In *Sensors, 2010 IEEE*, pp. 1037 –1042, nov. 2010.

[10] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[11] A. Ivorra, C. Daniels, and B. Rubinsky. Minimally obtrusive wearable device for continuous interactive cognitive and neurological assessment. *Physiol Meas*, 29(5):543–554, May 2008.

[12] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[13] A. R. Jensen. *Clocking the mind: Mental chronometry and individual differences.* Elsevier, 2006.

[14] R. J. Kosinski. A literature review on reaction time, August 2009.

[15] R. Larson and M. Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 15:41–56, 1983.

[16] S. C. Lee and T. Starner. BuzzWear: alert perception in wearable tactile displays on the wrist. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pp. 433–442, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9.

[17] H. R. Lieberman, F. M. Kramer, S. J. Montain, and P. Niro. Field assessment and enhancement of cognitive performance: Development of an ambulatory vigilance monitor. *Aviation, Space, and Environmental Medicine*, 78(5, Suppl.):B268–75, May 2007.

[18] R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford University Press, 1986.

[19] M. Matscheko, A. Ferscha, A. Riener, and M. Lehner. Tactor placement in wrist worn wearables. In *14th annual IEEE International Symposium on Wearable Computers (ISWC'10)*, pp. 1 –8, 2010.

[20] I. Oakley, Y. Kim, J. Lee, and J. Ryu. Determining the feasibility of forearm mounted vibrotactile displays. In *Proceedings of the Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, HAPTICS '06, pp. 27–34, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 1-4244-0226-3.

[21] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[22] U. Scholz, R. L. Marca, U. M. Nater, I. Aberle, U. Ehlert, R. Hornung, M. Martin, and M. Kliegel. Go no-go performance under psychosocial stress: Beneficial effects of implementation intentions. *Neurobiology of Learning and Memory*, 91(1):89 – 92, 2009. ISSN 1074-7427.

[23] C. N. Scollon, C. Kim-Prieto, and E. Diener. Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, 4(1):5–34, 2003.

[24] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky. Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands. *Neuropsychologia*, 47(12):2389 – 2396, 2009. ISSN 0028-3932.

[25] D. Watson, L. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[26] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[27] K. Wild, D. Howieson, F. Webbe, A. Seelye, and J. Kaye. Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's and Dementia*, 4(6):428 – 437, 2008. ISSN 1552-5260.

8

Effect of Daily Activities on Reaction Times

Burcu Cinaz, Bert Arnrich, Nathan Theill, Vera Schumacher, Mike Martin and Gerhard Tröster

Full publication title: Effects of Daily Activities on Reaction Times: Comparison between Young and Elderly Subjects

Pervasive and Mobile Computing, 2013 (submitted)

Abstract

In order to enable a continuous monitoring of variations in cognitive functioning we aim for transferring reaction time (RT) tests from clinical settings into daily life. In previous work we have designed and implemented a watch-like wearable RT test which can be operated by a simple hand gesture. The aim of this work was to measure the impact of common daily activities on reaction times of younger and older adults. We have conducted an experiment in which subjects had to perform daily activities once under an idle condition and once under additional cognitive load. During performing the activities we have measured reaction times at random intervals. As evaluation metrics we considered mean RT and RT variability. We addressed the research questions how daily life activities affect the reaction time measures, whether reaction time measures differ between young and elderly subjects and whether reaction time measures differ between idle and cognitive load conditions. We could show that mean RT and RT variability were significantly affected by the type of activity for both age-groups. The increase in mean RT and RT variability from idle to cognitive load condition was significantly higher for the elderly. We performed logistic regression to investigate the discriminative power when distinguishing between idle and cognitive load conditions. The classification results confirmed the effect of activity on reaction times. In a leave-one-person-out cross validation the discrimination accuracy increased from 75% to 80.77% after adding activity class features to the classification model. When applying a context-aware RT test we achieved a discrimination accuracy of 87.5%. We conclude that a wearable RT test combined with a simple activity recognition system is feasible to detect changes in RT performance and variability during common daily life activities.

8.1. Introduction

Different studies could demonstrate that speed of information processing, measured by reaction time, is related to cognition in younger and older adults. Especially slowing in reaction time is often attended by cognitive disorders or deficits. While children and younger adults mainly suffer from cognitive disorders such as Attention Deficit Hyperactivity Disorder (ADHD) older adults tend to have cognitive deficits because of neurological diseases such as brain injury, dementia, stroke or cerebral tumor. Assessing cognitive functioning is commonly used for screening of cognitive impairment, distinguishing causes of impairment, rating of disease severity or monitoring disease progression.

Most commonly used methods for the diagnosis of cognitive deficits are neuropsychological screening tests. One simple and often used test is the Mini Mental State Examination which include questions to evaluate memory, attention, language or orientation skills [8, 31]. In general, neuropsychological screening tests are administered usually in later stages when observable changes in patient's cognitive state arise [24]. Other methods used for cognitive assessment are neuroimaging techniques or laboratory testings which measures specific hormones or proteins [2]. In recent years computerized assessments of cognitive functioning were developed which usually provide more sensitive information compared to paper-based screening alternatives. In the literature there exists several computer-based cognitive tests for detecting cognitive decline [30]. In [24] several shortcomings of the established clinical assessment tools were identified. In particular, it was highlighted that these tests are time consuming, expensive and are performed only infrequently since a trained psychologist needs to administer these tests. In order to overcome these limitations it was suggested to develop unobtrusive measurement techniques which have the potential to continuously assess cognitive functioning in daily life. One approach to obtain a continuous assessment of cognitive functioning is to use wireless sensor networks in home environment. This enables the observation of the individuals during their activities of daily living (ADL) such as physical activities, sleep quality, gait velocity, location changes or interaction with objects in the house [7, 13, 23, 25, 28]. Usually ambient sensors such as motion detectors or RFID tags are integrated into the living environment of the individual or body-worn accelerometers are used to monitor the daily activities. The main goal of these applications is to detect deviations from normal patterns i.e. to detect abnormal changes in behavioral patterns. These markers are then correlated with the cognitive performance of the subjects. However, while these approaches enable a full unobtrusive monitoring, they do not provide a direct measurement of cognitive functioning as the well-defined screening methods do.

In our approach we aimed for transferring a well-defined cognitive test into daily life in order to obtain a continuous measurement of cognitive performance. We decided to transfer a reaction time (RT) test into daily life since it is well known that RT tests offer a high sensitivity for detecting variation in cognitive functioning [17, 18]. Furthermore, they can be repeated virtually an unlimited amount of times [18]. In general, a reaction time test measures how rapidly a person can initiate a response to a designated stimulus which is also a measure of speed of information processing [18, 22]. Reaction time tests have been extensively investigated by experimental psychologists since the middle of the 19th century. Commonly, visual stimuli are generated at random time points and the subject has to respond by using a keyboard, mouse or special buttons. There are several examples of applying RT tests to assess cognitive functioning: characterization of age-associated changes in reaction times; early detection of cognitive decline such as mild cognitive impairment or Alzheimer's disease; determining the ability to manage complex activities such as piloting, or identifying children with intellectual disabilities such as Attention Deficit Hyperactive Disorder (ADHD). However, most of these tests are administered in a laboratory environment by a clinician. More importantly, RT tests are incompatible with daily life since the subject has to interrupt his daily routine for several minutes in order to perform the RT test. In order to overcome this limitation, in former work we have designed, implemented and evaluated a wearable watch-like reaction time test that can be operated throughout everyday life [6]. The wearable RT test generates haptic stimuli and recognizes the user's hand gesture as a response (see Figure 8.1). Haptic stimuli are generated from time to time by a vibration motor. The subject has to respond to a stimulus by performing a slight forearm rotation, which is recognized with an inertial measurement unit (IMU).

For the evaluation of the wearable reaction time interfaces, we conducted an experimental comparison with traditional desktop-based tests in a laboratory setting. The results showed that individual changes of reaction times occurred due to additional cognitive load were similar for both desktop-based and wearable RT test. In a case study, we monitored reaction times continuously in a real-world working setting to investigate the effects of different workload factors such as stress, sleep deprivation, night shift and alcohol on reaction times [4]. According to the results, the mean reaction time and variability were increased during high arousal conditions such as stressed, concentrated or alerted.

However, so far it remains unclear to what extent a working activity itself such as writing or reading affects the reaction times independent of arousal or cognitive load conditions.

In this contribution, we investigate how certain daily activities affect the reaction times. In order to examine both the effects of activity and cognitive condition on reaction times, each daily activity is investigated once under a control condition in which subjects just performed the activity and once under an experimental condition in which subjects were confronted with additional cognitive load while performing the activity. Since one of the potential application fields is to enable an early detection of age-related cognitive impairments in daily life, in this work we investigate an elderly collective in addition to young subjects. In addition we investigate to what extent it is possible to discriminate the cognitive load condition from the control condition when employing reaction time features as predictors.

In the following we first provide an overview about related work on reaction times. Next, we describe our methods on the experimental procedure and data analysis. In the results section we present the outcomes of the experiment. Finally we conclude our work, discuss the results and provide an outlook on future work.

8.2. Related Work

There exist mainly three kinds of reaction time tests [20]: simple, recognition and choice reaction. In simple reaction time tests the subject has to respond to one stimulus with a dedicated response. For instance the subject has to press a certain button as soon as a particular symbol appears at the screen. In recognition reaction time tests, which are sometimes called "go/no-go" tasks, the subject has to respond to target stimuli and ignore non-target stimulus types. Lastly, in choice reaction time tests the subject has to respond to multiple stimuli with a corresponding response e.g. pressing a dedicated key whenever a corresponding letter appears on the screen. A detailed series of recommendations on how to conduct experiments using reaction times and how to analyze the obtained data can be found in [18, 22, 29].

Reaction time tests have been extensively investigated for many years. There exist several studies where psychologists have identified different factors that influence reaction time. Increasing age and age-related diseases like cognitive impairment are known to influence length and variability of reaction times significantly [20]. For instance, Gorus et al. showed that reaction times and performance variability are potential markers for the early detection of Alzheimer's disease [9]. Persons with cognitive deterioration demonstrated more intra-individual performance variability and more slowing in their reaction times than cognitively healthy elderly. Braverman et al. showed that a particular recognition reaction time tests, i.e. the test of variables of attention (TOVA), is an accurate predictor of early attention complaints and memory impairments [3]. Another application area of reaction time tests is the investigation of Attention Deficit Hyperactive Disorder (ADHD) patients. In [27] the authors performed a study to examine the RT variability in ADHD using go/no-go tasks with differing levels of cognitive demand. The resulting findings showed that children diagnosed with ADHD exhibited an increased variability in both simple and more complex go/no-go task in comparison to healthy children.

Most of the studies have in common that the employed RT tests are operated with a computerized test, which requires the full attention of the subject for several minutes. Hence, most of these techniques are not feasible to be used without interrupting normal life activities. There exist only a few studies, which investigate the feasibility of measuring reaction times continuously during everyday activities. Lieberman et al. implemented a wrist-worn device to assess vigilance. The device consists of visual stimuli (3 LEDs), auditory stimuli (a miniature speaker) and two push buttons [21]. Ivorra et al. implemented a haptic stimulus to interrogate the central nervous system in a minimally obtrusive way [15]. As the response the detection of a wrist movement is defined. In our recent work [5, 6], we followed the approach of Ivorra et al. and designed a wearable reaction time device which combines the generation of haptic stimuli and the recognition of forearm rotation as subject's response. We could show that the wearable reaction time test is suited to measure changes in reaction times caused by cognitive load. In this contribution, we present an empirical study with 26 subjects in order to analyze the effects of different daily activities on reaction times. In particular, we compare the changes in reaction times between young and elderly subjects.

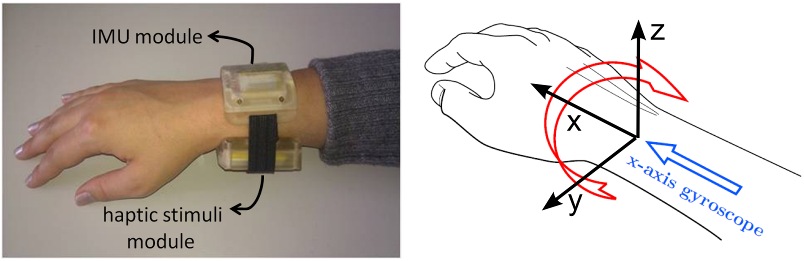

Figure 8.1. Implementation of a wearable simple reaction time test consisting of a haptic stimuli module and an inertial measurement unit (IMU) to detect subject's forearm rotation to a stimulus.

8.3. Methods

8.3.1. Objectives

The experimental procedure was designed to investigate the effect of everyday activities on reaction times of young and elderly subjects. Four common daily life activities were chosen: sitting, walking, reading and writing. As a second factor the influence of cognitive load on reaction times was investigated in a way that each activity was performed once under a control condition in which subjects just performed the activity (idle condition) and once under an experimental condition in which subjects were confronted with additional cognitive load while performing the activity (load condition). Having idle and cognitive load conditions during daily life activities allowed us to compare the changes in reaction times between idle and load condition in both age groups. In addition we investigated to what extent we can discriminate between idle and cognitive load when employing reaction time features as predictors. Consequently the experiment addresses the following research questions: (1) Do daily life activities affect the reaction time measures? (2) Do reaction times differ between young and elderly subjects for chosen activities? (3) Do reaction times differ between idle and cognitive load conditions whilst performing daily life activities? (4) How accurate can we discriminate between idle and cognitive load conditions during daily life activities?

8.3.2. Wearable reaction time test

In our previous work we have designed and implemented a watch-like wearable user interface to measure reaction time throughout daily life [5, 6]. In comparison to state of the art desktop-based reaction time tests, we replaced the visual stimuli by a haptic stimuli which can be recognized without paying continuous attention to it. In addition, we replaced the response button by detecting a dedicated rotation of the forearm as response to the stimuli. Thus, the wearable reaction time test consists of two main modules: a stimulus module generating haptic stimuli and an inertial measurement unit (IMU) module for detecting forearm rotations. For generating vibro-tactile stimuli, we used a coreless mini DC vibration motor with a resonant frequency of around 200 Hz and a vibration amplitude of 3.2 g. The IMU module consists of the so-called ETHOS which includes a three-axis accelerometer and three-axis gyroscope to recognize the subject's gesture response [11]. The accelerometer and gyroscope is sampled with a frequency of 128 Hz. An implementation of a wearable simple reaction time test can be seen in Figure 8.1. In order to react to a haptic stimulus the subject has to rotate his forearm along the x-axis. The detailed description of the hardware platform can be found in [5].

8.3.3. Participants

Overall, 16 elderly and 15 young subjects participated in our study. The young subjects were recruited through public advertising at the university. The elderly participants were recruited from a dedicated university course which offers university-level education for elderly people. One young and two elderly subjects had to be excluded from the analyses due to their insufficient number of reaction times, i.e. they omitted at least 30 of 40 stimuli during one of the four activities. In addition, the data from two other elderly subjects had to be excluded since their data violated the normality assumption required by the statistical analysis. As a result, in our data analyses we used the data from 12 elderly subjects (7 female, 5 male) with an average age of 70.17 years (±5.52) and the data from 14 young subjects (6 female and 8 male) with an average age of 25.79 years (±5.31).

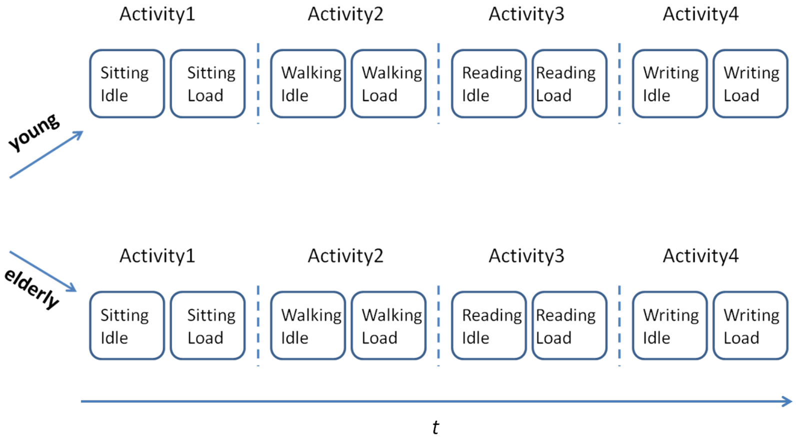

Figure 8.2. Experiment design: the four common daily life activities sitting, walking, reading and writing were investigated in both age groups. Each activity was performed once under an idle condition and once under a load condition. Each subject started with the sitting activity while the remaining activities were randomly assigned.

8.3.4. Experimental design

All participants were briefly informed about the experimental procedure and a written informed consent was signed by each subject. All elderly participants were screened for dementia using the Mini Mental State Examination (MMSE) [8]. No participant was at risk for dementia since all elderly had a score higher than or equal to 28. The experiment was designed in a way that each activity was performed under idle and cognitive load conditions. During all activities and both conditions (idle vs. load) the subject had to perform the wearable reaction time test, i.e. the subject was asked to respond as fast as possible to each haptic stimulus by performing a dedicated hand gesture. During the cognitive load condition, in addition to the main activity, the subject was asked to solve an Audio 2-Back task which is a variant of the N-Back test [1, 16]. In the Audio 2-Back task every 5 seconds a letter was presented to the subject via an audio message. The subject was asked to respond only if the currently pronounced letter was the same as the one that was pronounced two positions back. The subject responded by saying "match" whenever a sound match was detected by the subject.

In advance of the experiment, a training stage was provided to the participants that enabled them to become familiar with the wearable reaction time test and the Audio 2-Back task. Afterwards, each subject started with the baseline activity which was defined as sitting on a chair at rest and just performing the wearable reaction time test. The remaining activities walking, reading and writing were randomly assigned to each subject. Each activity started with the idle condition and was followed by the cognitive load condition. During walking activity the subject walked on a treadmill with a comfortable speed chosen by the subject himself. To obtain the right speed, each subject had a practice session to try different speeds before starting the walking session. Reading activity consists of reading a text on the table. In order to keep the subjects concentrated on the reading task, we told them that they have to answer a few questions about the text at the end of the experiment. During writing activity the subjects were asked to transcribe as much words as possible from a given text on a paper into a word document on the computer.

During all activities participants were requested to feel free in their mobility. The reaction time test device was mounted on the non-dominant wrist of the subjects. Each condition (idle vs. load) of each activity lasted about 10 minutes and we decided to collect overall 40 reaction time measurements within each condition of each activity. As a consequence, we generated haptic stimuli at random intervals varying between 5 sec and 25 sec. After each 10 minutes block, a pause was offered, if requested. Directly after each block, the subject was asked to indicate his perceived workload by completing the multidimensional assessment tool NASA Task Load Index (TLX) [12]. The rating consists of the following six scales: mental demand, physical demand, temporal demand, own performance, effort and frustration. Based on the ratings, the total workload was computed as a weighted average. The overall experiment lasted between 90 and 120 minutes. The experimental procedure can be seen in Figure 8.2.

8.3.5. Preprocessing and reaction time calculation

In order to calculate reaction time by means of subject's hand response to a stimulus, we analysed the gyroscope data obtained by the IMU module. We used the same method described in our previous work [4]. First, we normalized the gyroscope data to values between -1 and 1. Second, the raw gyroscope data was smoothed using a simple moving

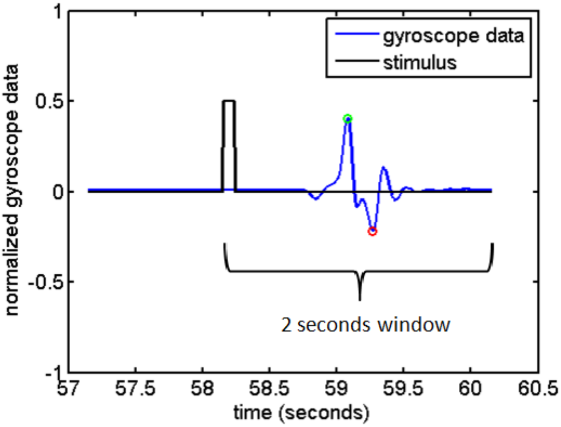

Figure 8.3. Normalized gyroscope data while responding to a haptic stimulus.

average filtering. Then, we aligned a window of 2 seconds length on each stimulus event and we located the local maximum point within this window. If the maximum peak was larger than a decision threshold, it was classified as a correct reaction gesture; otherwise it was classified as a non-reaction, i.e. it was assumed that the user did not responded to the stimulus. The normalized gyroscope data while responding to a haptic stimulus is exemplary shown in Figure 8.3.

The decision threshold was estimated by applying a Naive Bayes classifier. In our previous work [4] we selected a total of 120 responses and 120 non-responses by a visual inspection of the gyroscope data from one subject which performed daily life activities. This data was then used to define a decision threshold using the Naive Bayes classifier. After applying Naive Bayes on our data collected from 14 young and 12 elderly subjects, a visual inspection of altogether 8295 reaction hand gestures revealed an accuracy of 98%.

8.3.6. Measures

For the statistical analysis, mean reaction time and intra-individual variability measured as the coefficient of variation (CV) [19, 27] are used as evaluation metrics. For each individual, the coefficient of variation is computed as the standard deviation divided by the mean reaction time.

8.3.7. Statistical analysis

In order to analyze the effect of daily activity and cognitive load condition on reaction times for young and elderly participants, we performed a mixed 2x2x4 ANOVA. We utilized age group (young vs. elderly) as between-subjects variable while cognitive load condition (idle vs. load) and activity (sitting, walking, reading, writing) served as within-group factors. Mean reaction time and coefficient of variation were used as the dependent measures. Greenhouse-Geisser correction was performed if sphericity was violated in the repeated measures data. Benferroni corrected post-hoc group analysis was conducted to compare age group effects within each activity performed during idle and load condition. Data were tested for normal distribution and homogeneity of variance using the Shapiro-Wilk and Levene's test.

8.3.8. Classification

In order to analyze how accurate we can discriminate between idle and load condition, we performed logistic regression analysis. In all regression models under investigation, the dependent variable consisted of the cognitive load condition (1 for cognitive load, 0 for idle). In the set of independent variables the two features mean reaction time and reaction time variability were always present. The regression models under investigation differed in the amount of additional independent variables and the amount of data used for building the models. All models were built and evaluated in a leave-one-subject-out cross validation scheme, i.e. data from all subjects except one were used for building the model while the data from the omitted subject were used to compare the model prediction with the ground truth. This process was repeated until all subjects served once as test case.

In the first model, only mean reaction time and reaction time variability were used as independent variables. In the second model we considered a context aware reaction time test i.e. we assumed that we have an activity recognition system which is able to discriminate between sitting, walking, reading and writing. Therefore, we used additional activity class information as predictors together with mean reaction time and variability in the regression model.

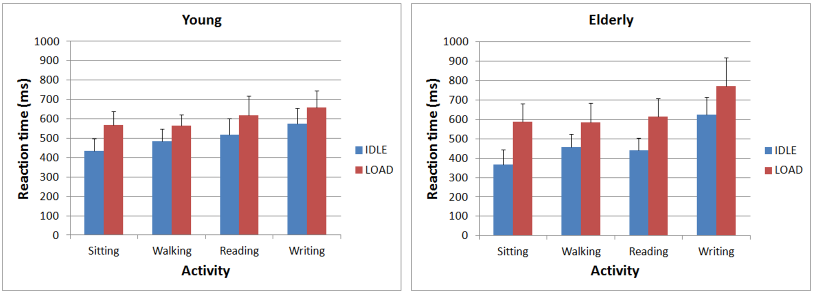

Figure 8.4. Mean reaction times of young and elderly subjects across four activities.

Afterwards, we considered a more simplified activity recognition system which could be easily implemented in the wearable reaction time test device worn on the wrist. We assumed that by using the continuous stream of IMU data collected with the wearable device, we can easily discriminate between activity classes "hand active" and "hand inactive". Therefore we built two separate models once for the activities where hand was active (walking and writing) and once for the activities where hand was inactive (sitting and reading). Consequently the reaction time features computed for all subjects during walking and writing were used for the third model and those computed during sitting and reading were used for the last regression model. The number of the observations used in the model was 208 for the first two models and 104 for the last two models.

8.4. Results

8.4.1. Reaction time performance

The mean reaction times of the young and elderly subjects collected during four different activities are shown in Figure 8.4. First, we can observe that for both young and elderly participants, the mean reaction time is increased during cognitive load condition compared to idle condition. Second, it can be observed that mean reaction time is highest in both age groups during writing activity. When performing the writing activity the elderly subjects showed the slowest reaction times both during idle and load conditions. However, during the idle condition of sitting, walking and reading activities the elderly showed slightly faster reaction times than young subjects. On the other hand, when elderly subjects experienced cognitive load, they were slower than the younger subjects during sitting, walking and writing activities. Thus, the drop in reaction time performance from idle to load condition was higher for the elderly compared to the young age group. In general, we can observe that reaction time performance varies with the type of the activity and the presence of cognitive load condition for both age groups. In the following we provide the results of the detailed statistical analysis in order to show the effect of activity and condition in more detail.

ANOVA results revealed a significant main effect of condition ($F(1, 24) = 120.03, p < 0.001$) showing that the subjects were on average slower during the cognitive load condition compared to the idle condition. There was a significant main effect of activity ($F(2.22, 53.34) = 37.97, p < 0.001$) indicating that mean reaction times of the subjects were significantly affected by the type of the performed activity. The interaction effect between activity and age-group ($F(2.22, 53.34) = 5.18, p < 0.01$) was significant i.e. the effect of the activity on mean reaction times was different for the young and elderly group. There was also a significant interaction effect between condition and age-group ($F(1, 24) = 7.46, p < 0.05$) revealing that the increase in mean reaction times from idle to cognitive load condition was higher for the older participants compared to the younger ones. Besides, there was a significant interaction effect between activity and condition ($F(2.12, 50.97) = 5.18, p < 0.01$), i.e. the difference on mean reaction times between idle and cognitive load condition varied depending on the type of the activity. The test of between-subjects effects did not show a significant main effect for age-group, reflecting the fact that there was not a significant difference between young and elderly participants in terms of the overall mean reaction time across the four activities and two levels of conditions.

In order to explore the effects further, we performed post-hoc group by group comparisons for each activity. There was a main effect of condition for all activities ($p < 0.001$) indicating that for each activity the mean reaction times were generally larger in the load condition independent of age group. There was a significant interaction effect between condition and age group for sitting ($p < 0.05$) and reading ($p < 0.05$) activities indicating that the increase in mean reaction time from idle to load condition was significantly higher for the elderly participants during sitting and reading activities. There was not a significant interaction effect between condition and group for walking ($p = 0.248$)

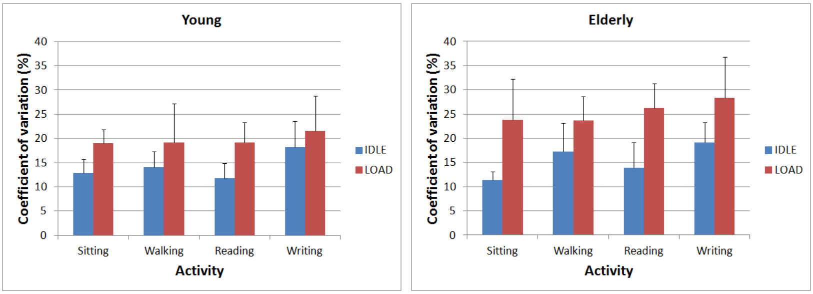

Figure 8.5. Reaction time variability of young and elderly subjects across four activities.

and writing ($p = 0.067$) activities. The test of between-subjects effects showed a significant difference of the overall mean reaction time between the young and elderly groups only for the writing activity ($F(1, 24) = 4.91, p < 0.05$).

8.4.2. Reaction time variability

Figure 8.5 depicts the reaction time variability of young and elderly subjects computed across the four activities. As observed in the mean reaction time analysis above, the group mean variability was increased during load condition compared to idle condition for both young and elderly subjects. Second, compared to the younger age group the elderly group showed higher variability under load condition during all activities. Besides, for both age groups the highest variability was observed during the writing activity. We can also observe that the amount of change in variability from idle to load condition was higher for the elderly group compared to young subjects. In order to investigate the effect of activity and condition in more detail, in the following we provide the results of the detailed statistical analysis.

The ANOVA test of within subjects effects showed that there was a significant main effect of condition ($F(1, 24) = 110.06, p < 0.001$) and activity ($F(3, 72) = 8.88, p < 0.001$) reflecting that group mean variability under load condition was higher compared to idle condition and the type of the activity significantly influenced the reaction time variability. There was not a significant interaction effect between activity and age-group indicating that the effect of the activity on the reaction time variability did not differ between young and elderly age group.

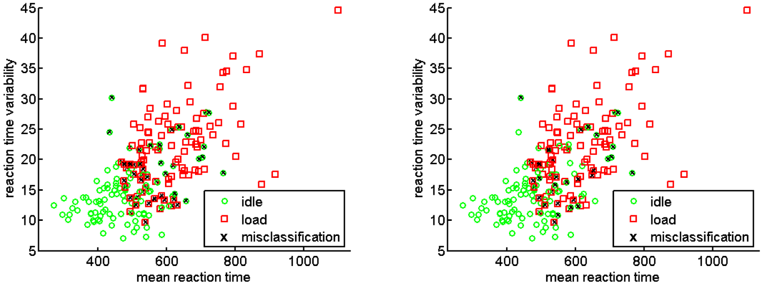

Figure 8.6. Scatter plots showing all 208 observations from all 4 activities during idle and load condition. Misclassified observations from a leave-one-subject-out cross validation are marked with crosses. Left: Logistic regression model which contains only mean reaction time and reaction time variability as predictors. Accuracy 75%. Right: Logistic regression model with additional activity class features as predictors. Accuracy 80.77%.

On the other hand, there was a significant interaction effect between condition and age-group ($F(1, 24) = 9.72, p < 0.01$) showing that the increase in reaction time variability from idle to cognitive load condition was higher for the older participants compared to the younger ones. This effect can also be seen in Figure 8.5. The interaction effect between activity and condition was also significant($F(3, 72) = 3.24, p < 0.05$), showing that the increase of reaction time variability from idle to cognitive load condition was depended on the type of the activity. The test of between-subjects effects showed a significant main effect of the age-group ($F(1, 24) = 8.97, p < 0.01$), reflecting the fact that elderly participants showed an overall higher reaction time variability across the four activities and two levels of conditions compared to the young age-group.

Post-hoc comparisons showed a significant main effect of condition for all activities ($p < 0.001$) indicating that for each activity the reaction time variability was generally larger in the load condition independent of the age group. Significant interaction effects between condition and group were found for the sitting ($p < 0.05$) and reading ($p < 0.05$) activities indicating that older subjects showed higher increase in their reaction time variability from idle to load condition compared to younger subjects. There was not a significant interaction effect between condition and group for walking ($p = 0.596$) and writing

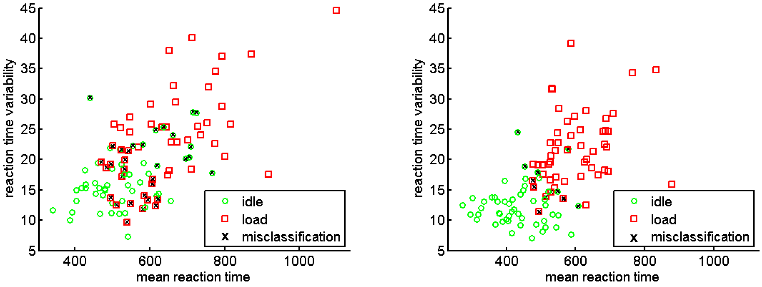

Figure 8.7. Scatter plots showing 104 observations from 2 sets of activities respectively. Misclassified observations from a leave-one-subject-out cross validation are marked with crosses. Left: Logistic regression model obtained from data when subject's hands are active, i.e. writing and walking. Accuracy 68.27%. Right: Logistic regression model obtained from data when subject's hands are not active, i.e. sitting and reading. Accuracy 87.5%.

($p = 0.054$) activities. The test of between-subjects effects showed a significant group difference in the overall reaction time variability only for the reading activity ($F(1, 24) = 10.51, p < 0.01$).

8.4.3. Classification: Idle vs. Load Condition

Figure 8.6 shows the scatter plots and the misclassified observations from the first two logistic regression models. In the leave-one-subject-out cross validation scheme the accuracy of the first model which contained only mean reaction time and reaction time variability as predictors was 75% (Figure 8.6 left). Adding the activity class features to the set of independent variables yielded an increase of the discrimination accuracy from 75% to 80.77% (Figure 8.6 right). Figure 8.7 depicts the scatter plot and the missclassifications when considering the availability of a simplified context-aware wearable reaction time test. The left figure shows the scatter plot for the regression model obtained from the data when subject's hands are active, i.e. data from the activities writing and walking. In this case, the lowest classification accuracy of 68.27% was achieved. In contrast, the right figure shows the scatter plot for the regression model obtained from the data when subject's hands are not active, i.e. data from the activities sitting and reading. In this case, the highest classification accuracy of 87.5% was achieved.

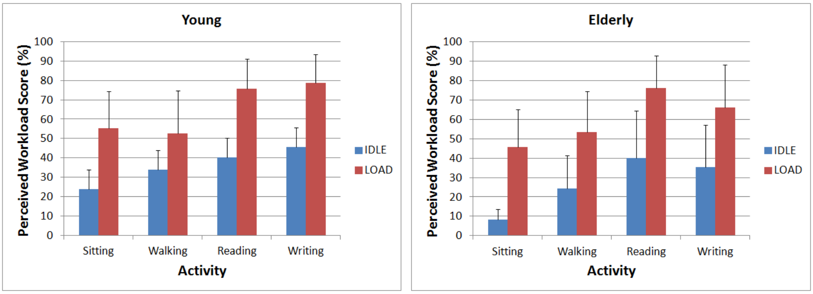

Figure 8.8. Perceived workload scores of the young and elderly subjects.

8.4.4. Subjective workload scores

Figure 8.8 shows the perceived NASA-TLX workload scores of the young and elderly subjects. First of all, as intended from the experiment design, it can be observed that for both young and elderly, the subjective workload ratings of the cognitive load condition were higher than the respective idle condition for each activity. The elderly subjects showed in average lower perceived workload than younger subjects during idle condition of all activities and during the load conditions of sitting and writing activities. On the other hand, the increase in perceived workload from idle to load condition was higher during sitting, walking, and reading activities for the elderly which confirms the similar trend in mean reaction time and variability outlined above. In order to model the relationship between NASA workload scores and reaction time measures we performed a multiple linear regression analysis combining all the data from young and elderly subjects. Mean reaction time and reaction time variability served as predictors whereas the perceived workload score was used as the dependent variable. We found that mean reaction time ($\beta = 0.085, p < 0.001$) and reaction time variability($\beta = 0.72, p < 0.05$) were significant predictors. The overall model fit was $R^2 = 0.282$.

8.4.5. N-Back scores

The N-Back scores from the Audio 2-Back task of the young and elderly subjects collected during the cognitive load conditions of all activities are shown in Figure 8.9. As seen in the figure, young subjects provided better performance in N-Back tasks than elderly subjects during all the

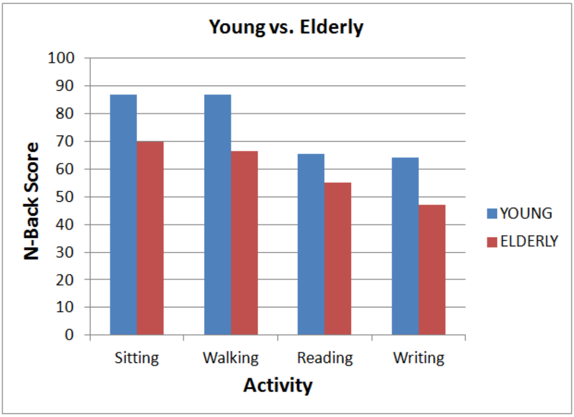

Figure 8.9. NBack scores of the young and elderly subjects collected during the cognitive load conditions.

activities. Compared to sitting and walking activities, the performance in N-Back was decreased during reading and writing activities to a great extent for both age groups. Young and elderly subjects showed the lowest performance during the writing activity which confirms the reaction time measures, i.e. the lowest reaction time performance and the highest reaction time variability was observed during the writing activity.

We performed a 2x4 ANOVA test where activity served as within-group factor while age-group was used as between-subjects variable. The results revealed a significant main effect of activity ($F(3, 72) = 26.86, p < 0.001$) on the N-Back score. There was no significant interaction between activity and age-group indicating that the effect of the activity on the N-Back score were similar for young and elderly. The test of between-subjects effects showed a significant main effect for age-group, showing that elderly participants showed an overall lower N-Back performance than younger subjects($F(1, 24) = 9.08, p < 0.01$). The post-hoc group by group comparison for each activity revealed that there was a significant difference between young and elderly for sitting ($p < 0.05$), walking ($p < 0.01$) and writing ($p < 0.01$) activities.

8.5. Conclusion and discussion

In this contribution, we have investigated how the four common daily activities sitting, walking, reading and writing affect the reaction times of young and elderly subjects. When investigating the differences between idle and cognitive load conditions we could show that mean RT and RT variability were significantly increased during load as intended from the experiment design. This confirms that cognitive load could be assessed in all daily activities under investigation.

We could show that mean reaction times and reaction time variability were significantly affected by the type of activity. For both age-groups, the highest mean reaction time and variability were observed for the writing activity. On the one side, this might be explained by the fact that typing on the keyboard is a motor-driven activity which may also cause additional delay on response times because of hand movements. On the other side, a possible explanation might be that the writing activity was a highly demanding task since it requires reading, memorizing and typing. As a consequence this finding provides evidence that a wearable RT test would allow to assess whether a certain task requires more or less cognitive capacity than another one. This would be especially important for the elderly since they may have less cognitive resources and a deviation in reaction times could indicate that a particular task is highly cognitive demanding for them. We performed logistic regression analysis in order to investigate the discriminative power when distinguishing between idle and cognitive load conditions. The classification results confirmed the effect of activity on reaction times. In a leave-one-person-out cross validation the discrimination accuracy increased from 75% to 80.77% after adding activity class features to the classification model. When considering a simple context-aware wearable reaction time test which is able to detect when the hand is not active, we achieved a discrimination accuracy of 87.5%.

When comparing young and elderly subjects, the elderly showed an overall higher RT variability across all activities. This result is consistent with other studies which have shown an increasing RT variability in the elderly [10, 14]. The increase in mean RT and RT variability from idle to cognitive load condition was higher for the elderly participants compared to the younger ones. There was not a significant difference between young and elderly regarding to overall mean reaction times although other studies showed that older adults were slower than younger ones [10]. On the contrary, older participants showed a slightly faster mean RT during the idle condition. One possible explanation could be the underlying motivational processes, e.g. it was observed that the elderly subjects were more motivated then the younger subjects during the whole experiment. Elderly perceived the accomplishment of the experiment as a sense of achievement whereas the main motivation for the younger subjects was the compensation money. The positive affective state of the elderly might have induced an increased arousal level which is a known factor that enhances the reaction time performance [18, 26]. However, the underlying motivational processes need to be investigated further.

We performed logistic regression analysis in order to investigate the discriminative power when distinguishing between idle and cognitive load conditions. The classification results confirmed the effect of activity on reaction times. In a leave-one-person-out cross validation the discrimination accuracy increased from 75% to 80.77% after adding activity class features to the classification model. When considering a context-aware wearable reaction time test which is able to detect when the hand is not active, we achieved a discrimination accuracy of 87.5%. As a consequence, this finding provides evidence that RT tests in daily life need to consider activity classes.

Furthermore, we investigated the perceived subjective workload of young and elderly subjects. Interestingly, the older participants reported in average a lower perceived workload than younger subjects. A possible explanation might be that the elderly often tend to be reluctant to admit cognitive problem because of a fear of losing their independence [24]. Finally, we analyzed the N-Back scores of the subjects. Elderly subjects showed significantly lower performance than younger subjects. In both age groups, the performance during reading and writing was lower compared to sitting and walking. This might be the result of the fact that reading and writing were cognitively more demanding.

Based on the achieved results, we finally conclude that a wearable reaction time test combined with an activity recognition system is feasible to detect changes in reaction time performance and variability during common daily life activities.

8.6. Limitations and Future work

In this contribution we have shown that a simple context-aware RT test would increase the classification accuracy when discriminating between cognitive load and idle states. In future work the context recognition should be an integral part of the wearable RT test. Thus, it would be possible to generate stimuli events only in cases when the hand is not active for a certain amount of time. More advanced, the wearable RT test could be combined with ambient sensor networks. This would allow involving more powerful context recognition systems in order to overcome the limitations of the single IMU approach we have considered so far. Thus, it would be possible to control the generation of stimuli events in a more advanced way. For example, it would be possible to activate the wearable RT test only in cases when the user is at home. Regarding design and implementation of the wearable RT test, in future work the device itself could be actually integrated into a normal watch. In addition, the device should be able to provide a feedback on the measurement results to the user. In this way, a user could be warned about an increased cognitive load.

So far, we have investigated healthy subjects and we have induced variations in cognitive functioning by applying cognitive load. In future work, wearable RT tests should be investigated for screening of cognitive impairment, rating of disease severity or monitoring of disease progression.

Bibliography

[1] Brain Workshop - a Dual N-Back game. http://brainworkshop.sourceforge.net/.

[2] E. Braverman. *Cognitive Decline of Aging: Important Neuroendocrinological Predictors of Early Cognitive Decline in A Clinical Setting.* Weill Cornell Medical Center, PATH Medical, New York, NY USA, 2011.

[3] E. R. Braverman, A. L. Chen, T. J. Chen, J. D. Schoolfield, A. Notaro, D. Braverman, M. Kerner, S. H. Blum, V. Arcuri, M. Varshavskiy, U. Damle, B. W. Downs, R. L. Waite, M. Oscar-Berman, J. Giordano, and K. Blum. Test of variables of attention (TOVA) as a predictor of early attention complaints, an antecedent to dementia. *Neuropsychiatr Dis Treat*, 6(1):681–690, 2010.

[4] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster. A case study on monitoring reaction times with a wearable user interface during daily life. *International Journal of Computers in Healthcare*, 1(4): 283–303, 2012.

[5] B. Cinaz, C. Vogt, B. Arnrich, and G. Tröster. A wearable user interface for measuring reaction time. In *Ambient Intelligence*, vol. 7040 of *Lecture Notes in Computer Science*, pp. 41–50. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-25166-5.

[6] B. Cinaz, C. Vogt, B. Arnrich, and G. Tröster. Implementation and evaluation of wearable reaction time tests. *Pervasive Mob. Comput.*, 8(6):813–821, Dec. 2012. ISSN 1574-1192.

[7] A. Coronato and G. De Pietro. Situation awareness in applications of ambient assisted living for cognitive impaired people. *Mobile Networks and Applications*, pp. 1–10, 2012. ISSN 1383-469X.

[8] M. F. Folstein, S. E. Folstein, and P. R. McHugh. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12(3):189–198, Nov 1975.

[9] E. Gorus, R. De Raedt, M. Lambert, J. C. Lemper, and T. Mets. Reaction times and performance variability in normal aging, mild

cognitive impairment, and Alzheimer's disease. *J Geriatr Psychiatry Neurol*, 21(3):204–218, Sep 2008.

[10] E. Gorus, R. De Raedt, and T. Mets. Diversity, dispersion and inconsistency of reaction time measures: Effects of age and task complexity. *Aging Clin Exp Res*, 18(5):407–417, Oct 2006.

[11] H. Harms, O. Amft, R. Winkler, J. Schumm, M. Kusserow, and G. Tröster. ETHOS: Miniature orientation sensor for wearable human motion analysis. In *Sensors, 2010 IEEE*, pp. 1037 –1042, nov. 2010.

[12] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, ed., *Human Mental Workload*, chapter 7, pp. 139–183. Elsevier, 1988.

[13] M. R. Hodges, N. Kirsch, M. W. Newman, and M. E. Pollack. Automatic assessment of cognitive impairment through electronic observation of object usage. In *Pervasive*, pp. 192–209, 2010.

[14] D. F. Hultsch, S. W. MacDonald, and R. A. Dixon. Variability in reaction time performance of younger and older adults. *J Gerontol B Psychol Sci Soc Sci*, 57(2):P101–115, Mar 2002.

[15] A. Ivorra, C. Daniels, and B. Rubinsky. Minimally obtrusive wearable device for continuous interactive cognitive and neurological assessment. *Physiol Meas*, 29(5):543–554, May 2008.

[16] S. M. Jaeggi, M. Buschkuehl, J. Jonides, and W. J. Perrig. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6829–6833, May 2008.

[17] L. H. Jakobsen, J. M. Sorensen, I. K. Rask, B. S. Jensen, and J. Kondrup. Validation of reaction time as a measure of cognitive function and quality of life in healthy subjects and patients. *Nutrition*, 27(5):561 – 570, 2011. ISSN 0899-9007.

[18] A. R. Jensen. *Clocking the mind: Mental chronometry and individual differences.* Elsevier, 2006.

[19] S. Kaiser, A. Roth, M. Rentrop, H.-C. Friederich, S. Bender, and M. Weisbrod. Intra-individual reaction time variability in schizophrenia, depression and borderline personality disorder. *Brain and Cognition*, 66(1):73 – 82, 2008. ISSN 0278-2626.

[20] R. J. Kosinski. A literature review on reaction time, August 2009.

[21] H. R. Lieberman, F. M. Kramer, S. J. Montain, and P. Niro. Field assessment and enhancement of cognitive performance: Development of an ambulatory vigilance monitor. *Aviation, Space, and Environmental Medicine*, 78(5, Suppl.):B268–75, May 2007.

[22] R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford University Press, 1986.

[23] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers. A review of wearable sensors and systems with application in rehabilitation. *J Neuroeng Rehabil*, 9:21, 2012.

[24] M. Pavel, H. Jimison, T. Hayes, J. Kaye, E. Dishman, K. Wild, and D. Williams. Continuous, unobtrusive monitoring for the assessment of cognitive function. *Handbook of cognitive aging: Interdisciplinary perspectives. Thousand Oaks, CA: Sage Publications*, pp. 524–543, 2008.

[25] J. Tung, J. Semple, W. Woo, W. Hsu, M. Sinn, E. Roy, and P. Poupart. Ambulatory assessment of lifestyle factors for alzheimer's disease and related dementias. In *AAAI Spring Symposium on Computational Physiology*, vol. 5, 2011.

[26] S. VaezMousavi, R. J. Barry, and A. R. Clarke. Individual differences in task-related activation and performance. *Physiology & Behavior*, 98(3):326 – 330, 2009. ISSN 0031-9384.

[27] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky. Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands. *Neuropsychologia*, 47(12):2389 – 2396, 2009. ISSN 0028-3932.

[28] A. Weiss, S. Sharifi, M. Plotnik, J. P. van Vugt, N. Giladi, and J. M. Hausdorff. Toward automated, at-home assessment of mobility among patients with Parkinson disease, using a body-worn accelerometer. *Neurorehabil Neural Repair*, 25(9):810–818, 2011.

[29] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[30] K. Wild, D. Howieson, F. Webbe, A. Seelye, and J. Kaye. Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's and Dementia*, 4(6):428 – 437, 2008. ISSN 1552-5260.

[31] H. J. Woodford and J. George. Cognitive assessment in the elderly: a review of clinical methods. *QJM*, 100(8):469–484, Aug 2007.

Glossary

Notation	**Description**
AD	Alzheimer's Disease
ADHD	Attention Deficit Hyperactive Disorder
ADL	Activities of Daily Living
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
CV	Coefficient of Variation
DC	Direct Current
ECG	Electrocardiography
EEG	Electroencephalogram
EMG	Electromyography
ESM	Experience Sampling Method
ETHOS	ETH Orientation Sensor
GSR	Galvanic Skin Response
HF	High Frequency
HPA	Hypothalamus-Pituitary-Adrenal
HR	Heart Rate
HRV	Heart Rate Variability
IMU	Inertial Measurement Unit
k-NN	k-Nearest Neighbor Algorithm
LDA	Linear Discriminant Analysis
LED	Light Emitting Diode
LF	Low Frequency
MCI	Mild Cognitive Impairment
MMSE	Mini Mental State Examination

Notation	**Description**
NASA-TLX	NASA Task Load Index
PEBL	Psychology Experiment Building Language
PSD	Power Spectral Density
RFID	Radio-Frequency Identification
RT	Reaction Time
STD	Standard Deviation
SVM	Support Vector Machine
SWAT	Subjective Workload Assessment Technique
TOVA	Test of Variables of Attention
TSST	Trier Social Stress Test
VLF	Very Low Frequency

Curriculum Vitae

Personal Information

Burcu Cinaz
Born 27 April 1981, Karabük, Turkey
Citizen of Turkey

Education

2009–2013 Ph.D. studies (Dr. sc. ETH) in Information Technology and Electrical Engineering at ETH Zurich, Switzerland

2004–2008 M.Sc. studies (Dipl. -Inf.) in Computer Science at University of Bremen, Germany

1999–2004 B.Sc. studies in Business Informatics at Marmara University, Istanbul, Turkey

1992–1999 High school in Büyüksehir H. Yildiz Anadolu Lisesi, Istanbul, Turkey

Professional Experience

2009–2013 Research Assistant, Electronics Laboratory, ETH Zurich, Switzerland

2004–2008 Student Research Assistant, Center for Computing and Communication Technologies, Bremen, Germany