htw saar

Systems Neuroscience
& Neurotechnology Unit

Neuroergonomic Assessment of Human Robot Interaction

Master Thesis

Systems Neuroscience & Neurotechnology Unit
Saarland University of Applied Sciences
Faculty of Engineering

Submitted by	:	Dominik Limbach, B.Sc.
Matriculation Number	:	3662306
Course of Study	:	Biomedical Engineering (Master)
Specialisation	:	Neural Engineering
First Supervisor	:	Prof. Dr. Dr. Daniel J. Strauss
Second Supervisor	:	Dr. Lars Haab

Saarbrücken, December 25, 2019

Abstract

In recent years wearables have become a staple in our society. In the form of Smartwatches and fitness-tracking devices wearables have made their way into nearly a third of german households. Aside from the basic functions of a phone or a watch these devices offer a wide array of functionalities, many of which revolve around monitoring the user's vital parameters for health reasons as well as improving their physical performance in recreational activities.
Research suggests that there is great potential in the application of psycho-physiologic monitoring in highly demanding workplaces of today's industry, an area in which wearables are still largely underrepresented.
Therefore, we built a new system based on a wrist worn device capable of monitoring workers in stressful workplaces, specifically collaborative workplaces involving Human-Robot-Interaction (HRI), and predicting their current mental state on a real-time basis. With the deployment of such a system we are able to create neuroergonomic workplaces that are sensitive to a persons mental capability and capable of eliminating stress as one of the leading causes of injury and disease in the working population.

Zusammenfassung

In den letzten Jahren sind Wearables zu einer festen Größe in unserer Gesellschaft geworden. In der Form von Smartwatches und Fitness-Armbändern haben Wearables Einzug in annähernd dreißig Prozent aller deutschen Haushalte gehalten. Neben den Grundfunktionen eines Telefons oder einer Uhr bieten diese praktischen Geräte eine Bandbreite an Funktionen. Viele dieser Zusatzfunktionen befassen sich mit dem Messen der Vitalparameter des Trägers zur Früherkennung von Krankheiten oder zur Optimierung der sportlichen Leistungsfähigkeit. Forschungsarbeiten der letzten Jahre lassen das Potential der Anwendung von psychophysiologischem Monitoring in besonders stressvollen Arbeitsplätzen erahnen, ein Gebiet in dem Wearables nur selten anzutreffen sind. Aus diesem Grund haben wir ein System für Arbeiter an besonders stressvollen Arbeitsplätzen, insbesondere collaborative Arbeitsplätze mit Mensch-Roboter-Interaktion (HRI). Das System basiert auf einem Wearable, welches am Handgelenk getragen wird, und ist in der Lage den mentalen Zustand des Trägers zu beurteilen. Durch den Einsatz eines solchen Systems sind wir in der Lage neuroergonomische Atbeitsplätze zu schaffen die sich an die cognitive Leistungsfähigkeit eines Arbeiters anpassen und somit den Einfluss von Stress, als einen der führenden Gründe für Krankheiten und Verletzungen unter der arbeitenden Bevölkerung, drastisch zu senken.

Declaration

I hereby declare that I have authored this work independently, that I have not used other than the declared sources and resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. This work has neither been submitted to any audit institution nor been published in its current form.

Saarbrücken, December 25, 2019

Dominik Limbach, B.Sc.

Contents

1 Introduction

In this chapter we will introduce the role of neuroergonomics in adaptive automation work environments and look at some of the problems that led to the current state of the field. Additionally, we will provide background information on crucial topics, such as psychophysiological measures, machine learning techniques, and emotion classification.

1.1 Introduction to the Field

Neuroergonomics is often described as the study of brain and behavior at work. As the name suggests, neuroergonomics is comprised of two disciplines, neuroscience and ergonomics (also known as human factors). Neuroscience is concerned with the structure and the function of the brain. It is a highly interdisciplinary body of research spanning disciplines such as physiology, psychology, medicine, computer science, or mathematics. Human factors on the other hand is focused on examining the human use technology at work or other real-world settings. As the intersection of these two fields, neuroergonomics addresses both the brain and humans at work, but even more so their dynamic interaction [1] . By understanding the neural bases of perceptual and cognitive functions, such as seeing, hearing, planing and decision making in relation to technology and settings in the real world, neuroergonomics strives to develop new optimization methods for various areas of applications. The additional value neuroergonomics can provide, compared to 'traditional' neuroscience and 'conventional' ergonomics, promises substantial economical benefits as well as significant improvements to health care and therefore society at large. In the scope of this thesis, we will focus primarily on applications in work settings such as modern automated systems. An area in which the effects of neuroergonomics are expected to be even greater, considering the difficulty of obtaining measures of overt behavior [1].
Automated systems, in one form or another have been present in almost every branch of industry for the better part of what has been two centuries. Automation itself can be defined as the creation and application of technology by which the production and delivery of various goods and services is controlled and monitored with minimal human assistance. Automation can be encountered in many different places, from a simple control loop in a hydraulic system up to an artificial intelligence handling emergency breaking in autonomic cars.Static automation is the original form of automation. In this form, automation is an all-or-none technology either performing a task for us or not

[**?**]. Although, there are numerous benefits (i.e. relieving workers from the strain of performing repetitive tasks) to it, there also is increasing evidence that static automation comes at the price of impaired decision making, manual skill degradation, loss of situational awareness, and monitoring inefficiency [2]. These problems stem from the significant change of roles that automation causes in a work environment. What this means is that workers, once active operators of machines and technology, are now passive monitors who often face monitoring workloads that are inherently different and often significantly higher when compared to the manual control conditions of a non automated work environment. The continuous exposure to workloads that are either too high or too low can have dramatic effects on human-system performance, thereby potentially compromising safety [3]. This has already been indicated by the work of Parasuraman et al. (1993, 1994) who tested subjects in a multi-task flight simulator with optionally automatable components and reported a substantial decrease of human operator detection of automation failures after short periods of time in a static automation scenario with constant task assignment of both operator and the automated system [2]. However, we must not forget the consequences these highly automated and ever so stressful workplaces have on human operators specifically. Studies by Cooper et al. showed that working in a stressful environment increases the risk of suffering physical illness or symptoms of psychological distress, as well as work related accidents and injuries [4].
We will now take a look at what is called adaptive automation. Adaptive automation has been introduced to resolve the abovementioned issues of statically automated systems. Whereas static automation is considered to be an agent working for the operator, adaptive automation is viewed as an interactive aid working with the operator [2]. It attempts to optimize system performance by adjusting the task assignment between the human operator and automation dynamically. This task reallocation is based on task demands, user capabilities, and system requirements. Meaning that during high task load conditions or emergencies the use of automation is increased and decreased during normal operations [3]. Another advantage of adaptive over static automation is the ability to reconstruct the task environment in terms of what is automated, how it is automated, what tasks may be shared, and when changes occur [2]. However, for an adaptive system to be efficient we require both the operator and the automated system to have sufficient knowledge of each other's current capabilities, performance and state [2]. As discussed by Byrne et al. (1996), there are three main approaches to address this issue, which either use model-based prediction or continuous measurements to determine the operators current state. In the scope of this thesis we will focus on the latter of the two. Usually, the way to measure human performance at work is to use physiological measures that reflect, more or less directly, aspects of brain functions. There is however a group of measures that is particularly favored among neuroergonomic researches. Those are the ones that are derived from the brain itself such as electroencephalography (EEG), magnetencephalography (MEG), and event-related potentials (ERPs), as well as measures related to the brain's metabolic and vascular responses such as positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) [1]. Although, there are many advantages to this group of measures, such as the temporal resolution of ERPs and the spatial resolution of fMRI, there is still one major disadvantage. In

fact, the majority of the abovementioned measures are either too expensive, impose too much restrictions on the movement of the subject, or are simply unfit to be used in a portable system, which ultimately prevents their application in real-world settings. Also, the lack of comfort and therefore low operator acceptance of a certain measure could have detrimental effects on the success of an adaptive automated system.

An alternative could by provided by systems that use psychophysiological indices to trigger changes in automation. In general, pychophysiology is focused on physiological measures and their psychological correlates [1], but there are many psychophysiological indices that reflect underlying cognitive activity, arousal levels, and external task demand. Some of these include cardiovascular measures (e.g. heart rate, heart rate variability), respiration, galvanic response, ocular motor activity, and speech [5]. Even though, research examining the utility of psychophysiology in adaptive automation has been rare, physiological measures are likely to be considered in the design of adaptive systems, either in isolation or in combination with other measures [2]. In addition psychophysiological measures are usually well accepted, due to their non-invasive nature and easy of application.

In conclusion, human-machine interaction in highly automated workplaces can be optimized by creating a work environment that is sensitive to the mental state of human operators. Using psychophysiological measures, a constant feedback in the form of a neuroergonomic assessment of the operator condition is provided to the machine agent. Consequently, adaptation mechanisms can be deployed to alter the quantity, or quality of the workload according to operator capability.

Within the scope of this thesis we designed a wearable system to facilitate this process. We built our system upon the Empatica E4 wristband, a medical-grade wearable device capable of real-time physiological data acquisition. We then conducted a pilot experiment, deploying our system in near real-world conditions. We monitored 14 subjects performing two cognitive tasks, simulating high and medium workload conditions, and two sets of visual stimulation, designed to elicit specific emotional states. Finally, we evaluated different machine learning algorithms using the acquired dataset. With our system we address a distinct need for a reliable, and truly non-obstructive method of handling neuroergonomic assessment of human-machine interaction in collaborative work environments. Eventually allowing us to create integral workplaces, that are sensitive to a persons mental capability and capable of eliminating stress as one of the leading causes of injury and disease in the working population.

1.2 Theoretical Background

1.2.1 Psychophysiology

Psychophysiology is a field of study that investigates the relationship between reasoning, feeling, behavior, and the physiological correlates associated with them. Advances in neuroscience, endocrinology, immunology, and molecular biology led to great insights in the interdependence of physiological and psychological processes. The translation of psychological functional states, emotions, and behavioral patterns into physiological reactions and processes is essentially controlled by three different systems: the autonomic nervous system (ANS), the neuroendocrine system (NES), and the neuroimmunological system (NIS). A common use of psychophysiological measures is the study of physiological correlates of emotions, attention, stress, and other cognitive processes.

1.2.2 Psychophysiological in Adaptive Automation

To fully understand the role of psychophysiology in adaptive automation we will take a look at the theoretical framework behind it. However, providing a complete overview on this topic would be far too extensive for the scope of this thesis. Therefore, we will only give a short summary of the work done by Byrne and Parasuraman (1996).
The application of physiological measures in adaptive automation is built on the premise that there is indeed an ideal mental state for human operators in a given task environment and that any deviation from this state would be detectable in the measurement. This hypothesis is based on resource and capacity theories of information processing, which suggest that humans draw from a limited pool of resources whenever they process information [2]. Over the years, many researcher delivered evidence for a connection between this resource utilization and physiological measures of activation, therefore establishing the importance of psychophysiological measures in the field of adaptive automation.
However, psychophysiological measures perform a dual role in adaptive automation systems. First, there is the investigatory role, which is often referred to as the developmental approach. This approach is focused on using the information psychophysiological measures provide on the mechanisms underlying performance changes corresponding to changes in automation, and further the development of model-based and hybrid approaches [2]. The second role, is often characterized as the regulatory approach. Here, unique information about the human operator is gathered from psychophysiologic measurements. This information is then used as input to a hybrid adaptive logic, thus allowing for dynamic restructuring of the task environment. Although, this approach seems ideal to support the operation of an adaptive system due to its immediate effect on the automated work environment, there may be years of effort and considerable maturation in technology required for it to be efficient in its application.

1.3 Psychophysiological Measures

The identification of suitable psychophysiological measures plays a vital role to the success of an adaptive work environment. Considering the dual role framework of psychophysiology, there is a distinction to be made between the two applications in adaptive automation. Because the developmental approach is in alignment with the majority of applications in psychophysiological research, the often stated criteria of specificity, diagnosticity, and intrusiveness for selecting workload assessment techniques also hold for adaptive automation [2]. On the other hand, criteria for the regulatory role of psychophysiology in adaptive automation have to be more strict. As they become part of closed-loop systems operating in real-time their potential impact is far greater, and their effects more immediate compared to when used for developmental measures. In addition, the cost in terms of intrusiveness and technical requirements have to be weighed against the explanatory power of a certain measure. If the gain in predictive value does not offset the cost of implementation, a measure is not considered for applications outside of laboratory environment.
As the recent work is determined to employ the Empatica E4 wristband, we are limited to the measures that are provided by this platform. These measures are blood volume pulse (BVP), skin response (GSR), and surface temperature.

1.3.1 Photoplethysmography

Photoplethysmography PPG is an optical measurement technique, used to detect blood volume changes in the microvascular bed of tissue [6]. To work PPG only requires a few opto-electronic components. First, a light source is used to illuminate the tissue. Then a photodetector measures the variations in light intensity associated with changes in perfusion in the catchment area. The most common light sources in PPG produce wavelengths in the red or near infrared area. This specific part of the spectrum, also referred to as the optical water window, is chosen for its ability to pass through biological tissue with relative ease. Therefore, influences associated with light-tissue interactions are widely reduced and the measurement of blood flow or volume is facilitated at these wavelengths. Even so, because the PPG is representing an average of all blood volume in the arteries, capillaries, and any other tissue through which the light has passed. The PPG signal is dependent on the thickness and composition of the tissue beneath the sensor, as well as the position of the source in relation to the receiver of the infrared light [7].

The PPG waveform: characteristics and analysis

The PPG waveform is comprised of two major components. The pulsatile component, often referred to as the "AC" component, possesses a fundamental frequency of approximately 1 Hz, and it represents the increased light attenuation associated with the increase in microvascular blood volume with each heartbeat [6]. It is superimposed onto the much larger "DC" component, which relates to the tissue and the average blood volume contained in the observation area. Variations in the DC component are slower and caused by respiration, vasomotor activity and vasoconstrictor waves, as well as thermoregulation [6].

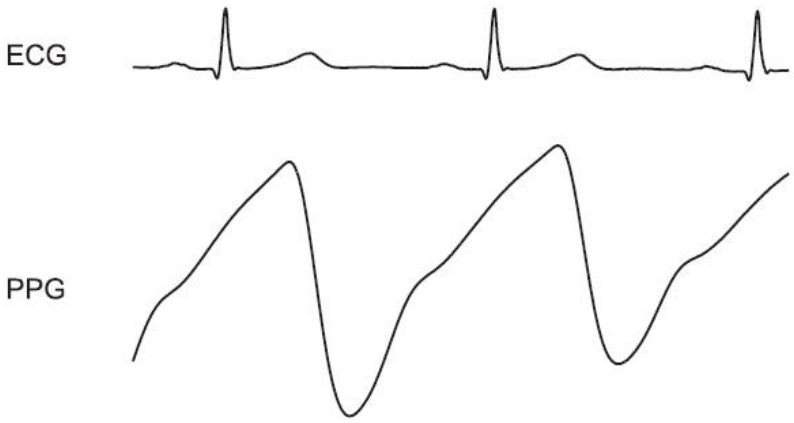

Figure 1.1: The PPG signal and the corresponding ECG. Displayed are the pulsatile AC component, which is super imposed on the much larger DC component. The PPG waveform represents light attenuation in relation to the blood volume in the tissue.

Its synchronization with the heart beat makes the AC pulse of the PPG waveform a valuable source of information on heart functions and condition. Based on the appearance of the AC pulse, two phases have been defined, reflecting its two most important properties. The first was labeled as the anacrotic phase and describes the rising edge of the pulse. This part of the waveform is primarily related to the systole. The second phase, shows the effects of diastole and wave reflections from the periphery of the vascular system. This phase is called catacrotic and can be observed in the successive falling edge of the pulse. In healthy subjects there usually is an observable dicrotic notch during this phase.

In addition to this coarse classification, a number of key landmarks have been defined to facilitate the analysis of the waveform and the underlying physiology. Depicted in 1.2 are the three main features that are derived from a single pulse. The pulse transit time to the foot (PTTf), and the pulse transit time to the peak (PTTp) are defined as the

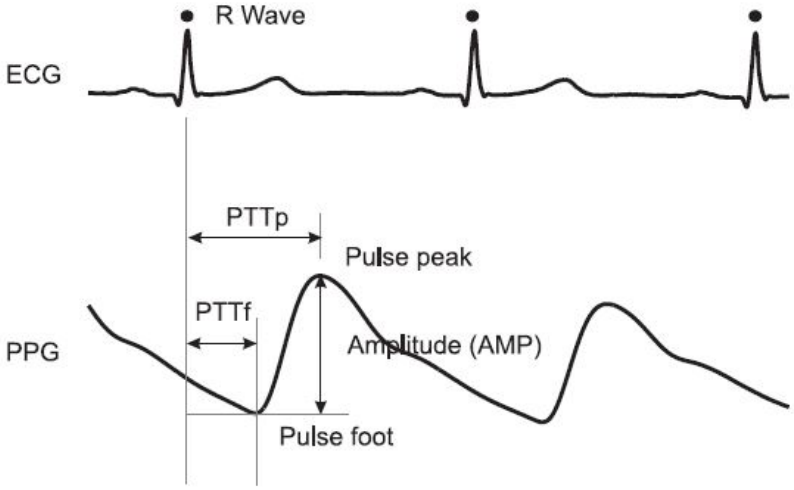

Figure 1.2: Characteristics of the PPG pulse waveform in relation to the ECG.

time delays between a heartbeat, indicated by the R-wave of the ECG, to the onset and the peak of the subsequent AC pulse. The amplitude of a pulse is determined by the absolute value of the displacement between its base and its peak, which are marked by the aforementioned temporal features.
However, the scale of these characteristics, as well as the overall appearance of the waveform are still subject to change. It is believed that these changes are largely caused by reflection of the pulse wave and the tapering down of the arteries towards the periphery [6].
Another important consideration in the analysis of PPG signals is their susceptibility to movement related artifacts. Although, there are a number of different artifacts that could occur we will only inspect the ones relative to our application. As we employed the Empatica E4, a wrist worn device, to measure BVP most of the artifacts are related to large movement of the arm and the wrist but also to small tremors in the hand or fingers. Additionally, extremes in physiological variation such as coughing and marked changes in the breathing pattern prove to be quite influential. Figure 1.3 shows an illustration by Allan (2007) depicting the effects of movement related artifacts to one minute PPG recordings that were taken at the index finger.

This concludes the section on the general waveform morphology and the reasoning for its variations. Lastly, we will take a closer look at the features we derived from the PPG measurement, specifically heart rate, and heart rate variability.

1.3.2 Heart Rate Variability

Variations in the length of the intervals between consecutive heart beats are called heart rate variability (HRV). Typically these inter beat intervals (IBIs) are determined by calculating the distance between two subsequent R-peaks of an ECG signal. However, they can also be derived from PPG signals. Again, IBIs are the time periods between the maxima of subsequent AC pulses. Since its first appreciation in 1965 HRV experienced a significant increase in popularity due to the apparent ease of derivation from widespread measures, such as ECG and PPG. In 1981, Akselrod et al. introduced power spectral analysis of HRV, contributing to the understanding of its autonomic background by relating the power content of certain frequency bands to sympathetic and parasympathetic activity [8]. In more recent years the increased interest in the application of psychophysiological measures in the field of adaptive automation led to an extensive discussion on HRV as a candidate measure. Byrne et al. (1996) argued in support of HRV as a possible index for both cognitive effort and compensatory effort. But, they also warned against the negligence of situation-related influences (e.g. the given task environment) in the process of signal interpretation. Thereby, a careless approach could have considerable consequences on the efficacy of adaptive automation.
In addition, the significance and meaning of the many different HRV measures are more complex than generally appreciated and consequently there is a high potential for incorrect conclusions and for excessive or unfounded extrapolation [8]. In 1996, the European Society of Cardiology and the North American Society of Pacing and Electrophysiology put together a Task Force to address this problem by developing appropriate standards for the acquisition and analysis of HRV. These standards remain valid until today and help preserving the integrity of HRV analysis if applied correctly.
Measures of the HRV can be divided into three general groups, the time domain methods, the frequency domain methods and non-linear methods.

HRV Measures

Of the three, time measures are perhaps the simplest to perform. These methods either determine the heart rate at any point in time or the intervals between successive normal complexes (in our case inter beat intervals) to calculate a series of statistical, and geometrical variables [8]. Statistical methods can be divided into two classes: those that are derived form direct measurements of inter beat intervals or instantaneous heart rate, and those derived from difference in inter beat intervals. Geometrical methods on the other hand, convert a series of IBI into a geometric pattern, such as the sample density distribution of IBI duration, sample density distribution of differences between adjacent IBI, Lorenz plot of IBI, etc., and then use a simple formula which judges the variability based on the geometric and/or graphic properties of the resulting pattern[8]. Figure 1.5 shows a variety of time-domain measures of HRV that have been recommended

by the Task Force of The European Society of Cardiology. Other common methods of HRV measures involve frequency-domain methods. By the means of power spectral density (PSD) analysis, these methods provide basic information on power distribution (i.e. variance) as a function of frequency. There are two ways to calculate PSD, non-parametric and parametric methods, both of which provide comparable results. In most cases non-parametric methods offer easier and faster calculation, whereas parametric methods, if applied correctly, can provide smoother spectral components which can be distinguished independently of pre-selected frequency bands. There are three main spectral components that are distinguished in a spectrum calculated from short-term recordings of 2 to 5 min: very low frequency (VLF), low frequency (LF), and high frequency (HF) components [8]. The distribution of the power and the central frequency of LF and HF are subject to changes in autonomic modulations of the heart period and therefore widely used measures in emotion recognition. Finally, we will have a look at non-linear methods. Although, the utility of these methods is still discussed by researchers, they are believed to provide valuable information for the physiological interpretation of HRV. Non-linear measures reflect the complex interactions of haemodynamic, electrophysiological and humoral variables, as well as autonomic and central nervous regulations involved in the genesis of HRV. At the moment, non-linear methods are speculated to be potentially promising tools for HRV assessment, but standards are lacking and the full scope of these methods cannot be assessed.

1.4 Machine Learning

Machine learning is a scientific study revolving around the development of algorithms and statistical models that provide computer systems with the ability to automatically learn and improve without being explicitly programmed. Machine learning techniques are commonly categorized as either supervised or unsupervised. This distinction is based on the type of input and output data they use as well as the type of problem they are intended to solve. Supervised learning algorithms require a set of data that contains both the inputs and the desired outputs to a certain problem. Based on this data, often referred to as training data, supervised algorithms derive a function that can be used to predict the output associated with new inputs. Supervised learning algorithms are generally used to solve classification and regression tasks. In contrast unsupervised learning algorithms are able to function on datasets that do not provide any output at all. They attempt to find a structure in the training data, by identifying commonalities. Based on the absence or presence of these similarities in new data they are then able to make predictions. Over the years a variety of algorithms have been developed for either category, each with its own objectives, strengths and weaknesses. Identifying an algorithm that is suitable for the desired task is one of the key components to a successful application of machine learning. Therefore, we will now take a closer look at the algorithm selection process used in the recent work.

1.4.1 Algorithm Selection

In the first step of the selection process we considered the work of previous researchers. In particular Wu et al. (2008), who presented the top 10 algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006 [9]. The algorithms were chosen according to the following steps. First they invited renowned researches of the field to each nominate up to 10 best-known algorithms in data mining. Each nomination had to provide the following information: the algorithm name, a brief justification, and a representative publication reference. After the nominations had been verified, those with less than 50 citations were removed. The remaining nomination were then organized in 10 topics: association analysis, classification, clustering, statistical learning, bagging and boosting, sequential patterns, integrated mining, rough sets, link mining, and graph mining. The final 18 candidates were then put to another vote with a much larger involvement of the research community. The results of this vote were the presented as the top 10 algorithms.
From this initial pool of algorithms we then selected those that were assigned to the topic of classification. As Python was the programming language we had agreed upon for all of our machine learning applications we selected the three algorithms that showed the highest compliance with the standard Python libraries from the subset of classification algorithms: kNN-classifiers, support vector machines, and ensemble methods. Also, we decided to include neural networks, in particular multi-layer perceptrons, into our final group of algorithms for exploratory reasons. In the following subsections we will provide a brief description of each of the final four machine learning algorithms.

1.4.2 k-NN Classifier

Nearest neighbors methods are possibly the simplest machine learning algorithms for supervised and unsupervised learning. The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The number of samples, named k, can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning). The distance can, in general, be any metric measure, but standard Euclidean distance is the most common choice. Neighbors-based methods are known as non-generalizing machine learning methods, since they do not attempt to construct a general internal model, but simply store instances of the training data. Classification can then be computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point [10].

1.4.3 Support Vector Machine

Support vector machines (SVM) are considered one of the most robust and accurate methods among all machine learning algorithms. SVM have a sound theoretical foundation, require only a dozen examples for training, and are insensitive to the number of dimensions [9]. In a learning task with two classes, SVM aim to find the best classification function to distinguish between members of the two classes in the training data. The metric that is used to identify the best classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane $f(x)$ that passes through the middle of the two classes, separating the two. Once this function is determined, new data points x_n can be classified by simply testing the sign of the function $f(x_n)$; x_n belongs to the positive class if $f(x_n)$ is greater than zero. Additionally, SVM select the best classification function from all available hyperplanes by maximizing the margin between the two classes. The margin is defined as the amount of space, or separation between the two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane [9]. The reason why maximum margin hyperplanes are emphasized this much in SVMs is that they provide the best generalization ability of the myriads of hyperplanes available. Therefore, they achieve the best classification performance on the training data, while still leaving enough room for the correct classification of the future data.

1.4.4 Decision Trees

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning a hierarchic set of simple decision rules inferred from the data features. Training a decision tree for a certain problem is equal to learning a sequence of yes or no questions that will lead as fast as possible to the right answer. In machine learning these questions are known as tests. However, real life data is rarely available in the form of binary properties, but rather as continuous attributes that require tests in the form of inequations (e.g. is the attribute bigger or smaller than a certain value). To build the tree, the algorithm will go over every possible test and select the one that provides the most information on the target value. The first node, or the root, is representative of the entire dataset. At each node a test will be administered and consequently the dataset will be divided into two subsets of data, which again are represented by nodes. This process is then repeated, growing a tree of binary decisions until the dataset is fully divided and the last nodes of each branch only contain values of one category. These final nodes are also called leaves. A leaf that only contains a single value is called pure. Predictions on a new data point can be made by examining the region of the attribute space that is associated with it and selecting the prevalent target value in this region. This region can be located

by simply reconstructing the tree. Starting from the root and following every decision until eventually one of the leaves is reached. Decision trees are easy to understand and to interpret. In addition they require very little data preparation, as opposed to other techniques, such as support vector machines. On the other hand, decision tree learners have a tendency to overfit. This means that they create overly complex trees that do not generalize well and therefore requires additional adjustments. There are two main strategies to prohibit overfitting in decision trees: Pre-pruning, which stops tree growth early on (e.g. limiting the maximum depth of the tree), and Post-pruning, which either removes or merges nodes that provide the least information.

1.4.5 Ensemble Techniques

Ensembles are methods that combine multiple machine learning models to create a new and more powerful model. Although there are many different ensemble-models, only two have proven effective on a wide variety of datasets: random forests, and gradient boosting machines. Both of these methods use decision trees as weak classifiers. Random forests are basically a combination of multiple, uniquely built decision trees. The basic principle of random forests is that every decision tree is inherently capable of making adequate predictions, but also overfits on some parts of the dataset. And because all trees are unique, their overfitting is too. If enough of these trees are combined overfitting can be reduced, by taking the average of all the predictions. In contrast to this, the gradient boosting machines use trees, that are built successively. Each tree is designed to address the failures of its predecessor. Again, the basic idea is to use a number of partially well working decision trees and combine them. Gradient boosting often applies heavy pre-pruning to create rather flat trees with a maximum depth of one to five levels.

1.4.6 Multi Layer Perceptron

Multi-layer perceptron (MLPs) are a type of feed-forward neural networks, a family of learning algorithms inspired by the human brain. Similar to their biological role model, MLPs consist of a network of artificial neurons, called perceptrons, that are arranged in multiple layers. There are at least three layers in each MLP: an input layer, a hidden layer, and an output layer. Basically, MLPs are supervised learning algorithms that learn a function $f(\cdot) : R^m \longrightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, ..., x_m$ and a target y, it can learn a non-linear function approximator for either classification or regression. The input layer is the leftmost layer of the network. It consists of a set of perceptrons $x_i|x_1, x_2, ..., x_m$ that each represent a single input feature. Next, in the middle section of the network, all of the hidden layers are located. Each perceptron of a hidden layer transforms all of the values from the previous layer with a

weighted linear summation $w_1x_1 + w_2x_2 + ... + w_mx_m$, followed by a non-linear activation function (e.g. the sigmoid function). The final layer is the output layer, which receives the values from the last hidden layer and transforms them into output value [10]. In MLP learning is achieved by adjusting the connection weights after each processing step. This adjustment is based on the size of error gained from comparing the output values with the expected results. MLP allow for complex models that capable of handling large datasets. However, a high complexity comes at the price of a high sensitivity to parameter changes, and data scaling methods, as well as longer training times.

1.5 Classification of Emotion

The means by which emotions can be distinguished from one another are called emotion classification. Over the last couple of decades this has been a strongly contested topic in emotion research and in affective science. One prominent approach to emotion classification, that has been widely accepted to this date, are dimensional models. Dimensional models of emotion attempt to characterize human emotions by determining their position in a two, or sometimes three dimensional space. After years of debate about the identity of these dimension two measures have claimed their place in most emotion models. Usually, the dimensions include some measure of valence or pleasantness and some measure of intensity or arousal [11]. For our work we employed the circumplex model by J. Russel (1980), one of the most prominent two-dimensional models for emotion classification. According to this model emotions are distributed in a two-dimensional circular space. The model space is represented by arousal (on the vertical axis) and valence (on the horizontal axis) and is centered on medium arousal and neutral valence. This allows for the possibility of emotions, or emotional stimuli that have high arousal and neutral valence (e.g. astonished, exited) [11] which separates the circumplex model from the rest of the two-dimensional models.

1.5.1 Emotion Elicitation

The American Psychological Association defines emotions as complex reaction patterns, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event. The specific quality of an emotion, such as fear or shame, is determined by the specific significance of the event. For example, if the significance involves threat, fear is likely to be generated; if the significance involves disapproval from another person, shame is likely to be generated. Emotion typically involves feeling but differs from feeling in having an overt or implicit engagement with the world. This suggests that emotions can be elicited by using appropriate stimuli. In general, there are two types of stimuli in emotion elicitation are visual stimuli (e.g.

pictures or films) and acoustic stimuli (e.g. sounds, speech, music) that are widely accepted in psychology and emotion research. During our experiments we attempted to elicit emotions by displaying images, which is one of the most practiced methods in the elicitation of emotional and affective states. As stimuli we used a selected subset of the International Affective Picture System (IAPS).

1.6 Acknowledgments

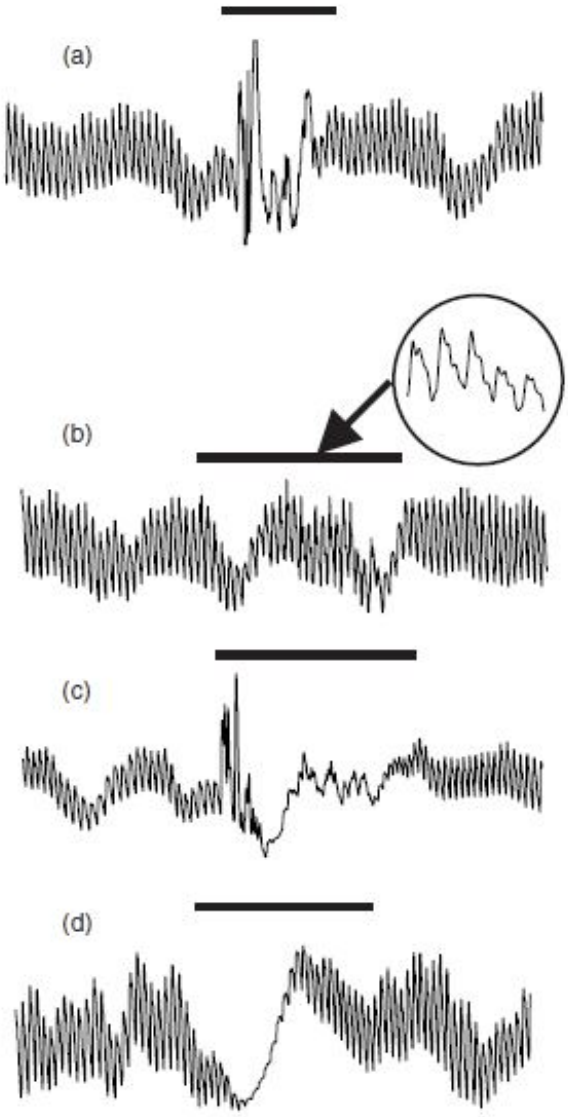

Figure 1.3: Examples of different types of measurement artifacts. All events were marked by black bars. (a) An episode of gross movement artifact of PPG probe cable tugging. (b) Hand or finger tremor. (c) a bout of coughing, and (d) marked changes in the breathing pattern (a deep gasp or yawn)

Variable	Units	Description
		Statistical measures
SDNN	ms	Standard deviation of all NN intervals.
SDANN	ms	Standard deviation of the averages of NN intervals in all 5 min segments of the entire recording.
RMSSD	ms	The square root of the mean of the sum of the squares of differences between adjacent NN intervals.
SDNN index	ms	Mean of the standard deviations of all NN intervals for all 5 min segments of the entire recording.
SDSD	ms	Standard deviation of differences between adjacent NN intervals.
NN50 count		Number of pairs of adjacent NN intervals differing by more than 50 ms in the entire recording. Three variants are possible counting all such NN intervals pairs or only pairs in which the first or the second interval is longer.
pNN50	%	NN50 count divided by the total number of all NN intervals.
		Geometric measures
HRV triangular index		Total number of all NN intervals divided by the height of the histogram of all NN intervals measured on a discrete scale with bins of 7·8125 ms (1/128 s). (Details in Fig. 2)
TINN	ms	Baseline width of the minimum square difference triangular interpolation of the highest peak of the histogram of all NN intervals (Details in Fig. 2.)
Differential index	ms	Difference between the widths of the histogram of differences between adjacent NN intervals measured at selected heights (e.g. at the levels of 1000 and 10 000 samples)[123].
Logarithmic index		Coefficient φ of the negative exponential curve $k \cdot e^{-\varphi x}$ which is the best approximation of the histogram of absolute differences between adjacent NN intervals[123].

Figure 1.4: Summary of recommended time-domain measures [8].

Variable	Units	Description	Frequency range
		Analysis of short-term recordings (5 min)	
5 min total power	ms^2	The variance of NN intervals over the temporal segment	approximately $\leq$0·4 Hz
VLF	ms^2	Power in very low frequency range	$\leq$0·04 Hz
LF	ms^2	Power in low frequency range	0·04–0·15 Hz
LF norm	n.u.	LF power in normalised units LF/(Total Power–VLF) × 100	
HF	ms^2	Power in high frequency range	0·15–0·4 Hz
HF norm	n.u.	HF power in normalised units HF/(Total Power–VLF) × 100	
LF/HF		Ratio LF [ms^2]/HF [ms^2]	
		Analysis of entire 24 h	
Total power	ms^2	Variance of all NN intervals	approximately $\leq$0·4 Hz
ULF	ms^2	Power in the ultra low frequency range	$\leq$0·003 Hz
VLF	ms^2	Power in the very low frequency range	0·003–0·04 Hz
LF	ms^2	Power in the low frequency range	0·04–0·15 Hz
HF	ms^2	Power in the high frequency range	0·15–0·4 Hz
α		Slope of the linear interpolation of the spectrum in a log-log scale	approximately $\leq$0·04 Hz

Figure 1.5: Summary of recommended frequency-domain measures [8].

2 Problem Analysis and Goals

2.1 Related Work

In the past a substantial part of research in the field of Human Machine Interaction has been focused on the use of emotion recognition to give robots the ability to perceive and appropriately react to human emotions. With the appreciation of psychophysiological signals as potent markers of emotional states, a new wave of studies has been conducted on their deployment in the recognition process of adaptive human machine interaction. We will now have a look at three studies that are closely related to ours. In this process we will investigate key aspects and show potential problems. The primary goal of all studies was to distinguish emotional states in subjects using a combination of psychophysiological measures and some machine learning technique. Measures of heart rate such as BVP or ECG, as well as galvanic skin response were used in all three as physiological indicators of emotion. In addition, respiration and EMG were used by both Maaoui and Pruski (2010) and Picard et al. (2001). Lastly, Maaoui and Pruski (2010) as well as Kim et al. (2004) employed skin temperature in their work. All measures were applied in their default measurement locations. Although, this might be acceptable for scientific settings in a lab, we argue that some of these location need to be adjusted or entirely ruled out to provide the minimum comfort to be applicable in a real-world setting. Especially, the negative effects of facial EMG, as well as the finger sensors used to measure GSR and BVP to user acceptance are easily comprehensible in an work environment where workers are physically engaged. Also, most of the sensors that were used in these studies still relied on a wired connection to deliver their data to a main processing device. A fact that further reduces practicability and in some workplaces may even increase the risk of injury. Proceeding to the classification of emotion, we find that all three studies reported above average results. A classification accuracy of 81% was achieved by Picard et al. (2001) while using Sequential Floating Forward Search and Fisher Projection methods to classify a total of 8 different emotional states. They used data they collected from a single subject over a time period of 30 days, with an average recording time of 25 minutes per day. Although, this method may account extremely well for daily changes in the recorded signal, it creates a classification that is strongly dependent on a single subject. The customization of an emotion recognition system to a single user may be of interest in the final stages of implementation, but one could argue that at this stage of research a user-independent system is far more beneficial due to its universal applicability. Next, we will consider the work of Maaoui and Pruski (2010) who achieved remarkable results.

They were able to reach a classification accuracy of 92% over 6 emotions by using SVM and Fisher Linear Discriminant analysis. However they used a set of elaborate protocols for emotion induction in combination with the IAPS and an extensive lab setting, which would again not be feasible to use in an adaptive automation workplace. Lastly, Kim et al. (2004) who attempted a user independent emotion detection system, using visually and acoustically stimuli in combination with the IAPS to elicit specific emotion in large groups (group 1: n = 125, group 2: n = 50) of children below the age of 9. Although, this strategy might be able to elicit stronger emotions that are much better represented by physiological correlates, it could be argued to be counter productive towards the generalization ability of their machine learning algorithm in an adult target population.

2.2 Problem Oriented

Section 2.1 shows that regardless of the significant accomplishments and overall advance of the field, there are still problems that are left to be resolved. As we pointed out earlier, although there were some attempts to conduct emotion recognition experiments in more realistic scenarios, the recording systems, and most of all the recording techniques (i.e. sensor locations) remained widely unimproved. Therefore, we will stray away from this pattern and use the Empatica E4, a wrist-worn system that is capable of wireless data transmission and features a single measurement site for all sensors. We believe that this is the right step towards truly unobtrusive emotion recognition allowing the field to advance by enabling more realistic experimental settings. Another important aspect we gathered from our literature research is the apparent focus on detecting only pure emotional states. Although, this may be useful in some applications, emotions are rarely presented in this form under realistic circumstances. Considering the setting of an adaptive automation workplace we will attempt to elicit and classify more generalized mental states (e.g. is the subject feeling pleasant or unpleasant?). Further, we will create our own database for classification including multiple subjects, representative of a working population to attempt a user independent system. Regarding the selection of psychophysiological measures we will comply with the studies we have presented in 2.1. However, as we have only access to a subset of these measures, we will use what is available to us: BVP, GSR, and skin temperature.

2.3 Aim, Objective, and Scope

As mentioned before, the aim of this thesis was to facilitate the neuroergonomic assessment of human robot interaction based on the real-time measurement of psychophysiological signals using the Empatica E4 wristband. Meaning, we are faced with the daunting task

of developing a compact emotion classification system that is capable of consistently gathering high quality data and reliably performing classification based on meaningful features. Therefore, we focused heavily on the development of such a system, limiting our scope to the engineering, and assembling of all system components and the final evaluation, based on system performance.
In the following we listed all the milestones that were necessary to complete this task.

MS 1: Development of a data extraction method that provides for real-time access on the raw signal data, as the E4 is actually designed for downstream data analysis only.

MS 2: Construction of a signal processing pipeline that compensates for artifacts while upholding signal integrity for feature extraction.

MS 3: Building a database for machine learning using authentic data gathered with our system in a series of measurements.

MS 4: Evaluation of system capabilities using different machine learning techniques.

2.4 Research Questions

RQ 1: Is it possible to record meaningful real-time data using the Empatica E4?

RQ 2: Is it possible to detect and distinguish different degrees of mental workload using standard machine learning techniques on three physiological signals that were recorded with the Empatica E4 in an experimental setting?

RQ 3: Is it possible to detect and distinguish two emotional states, such as pleasant and unpleasant, using the same approach as in RQ2?

3 Materials and Methods

3.1 Materials

All experiments and measurements were conducted within the facilities of Systems Neuroscience and Neurotechnology Unit, particularly the Green Lab, located at the University Hospital Saarland, and the Mindscan Lab, located at the HTW Saar (Technikum).

3.1.1 System Components

3.1.2 Hardware

Empatica E4 Wristband The Empatica E4 wristband is a wearable wireless device designed for comfortable, continuous, real-time data acquisition. It is a class IIa medical device in the EU, according to CE Cert. No. 1876/MDD (93/42/EEC Directive) and was designed for daily life usage [12].

Figure **??** shows an overview of the entire E4 wristband from either side indicating key attributes as wells as a total of four different sensors that will be discussed briefly in the following:

- **Photoplethysmography (PPG)** to provide blood volume pulse (BVP), from which heart rate, heart rate variability and other cardiovascular features may be derived
- **Electrodermal activity (GSR)** is used to measure sympathetic nervous system arousal and to derive features related to stress, engagement and excitement
- **3-Axis accelerometer** to capture motion-based activity
- **Infrared thermopile** for reading skin temperature

As the E4 is intended to be worn on the wrist these sensors are set up in a specific way

to provide for optimal use. As can be seen on **??** the majority of the sensors are located on the backside of the main unit not including the GSR-sensor, which is located on the wristband itself.

Wearing the E4 wristband is equally intrusive to wearing a watch and therefore providing a high level of convenience compared to other physiologic measures such as electrocardiogram ECG or electroencephalogram EEG.

Sampling Specifications All recordings were performed using only software licensed by Empatica. Using the approved streaming application and the compatible Bluetooth receiver, the recorded data was streamed directly to an operator's personal computer via a Bluetooth connection.

EDA sensor

- Sampling frequency: 4 Hz (Non customizable).
- Resolution: 1 digit 900 pSiemens.
- Range: 0.01 μSiemens – 100 μSiemens.
- Alternating current (8Hz frequency) with a max peak to peak value of 100 μAmps (at 100 μSiemens).
- Electrode(Placement): on the ventral (inner) wrist.
- Electrode(Build): Snap-on, silver (Ag) plated with metallic core.
- Electrode(Longevity): 4–6 months

PPG sensor

- Sampling frequency 64 Hz (Non customizable).
- LEDs: Green (2 LEDs), Red (2 LEDs) Photodiodes: 2 units, total 15.5 mm2 sensitive area.
- Sensor output: Blood Volume Pulse (BVP) (variation of volume of arterial blood under the skin resulting from the heart cycle).
- Sensor output resolution 0.9 nW / Digit.
- Motion artifact removal algorithm: Combines different light wavelengths. Tolerates external lighting conditions.

Infrared Thermopile

- Sampling frequency: 4 Hz (Non customizable).
- Range(Ambient temperature): -40...85degC (if available).
- Range(Skin temperature): -40...115degC.
- Resolution: 0.02degC.
- Accuracy ±0.2degC within 36-39degC.

Real-time clock

- Resolution(Recording mode): 5s synchronization resolution. Average of 6 seconds in 6 million seconds drift.
- Resolution(Streaming mode): Temporal resolution up to 0.2 seconds with connected device.

3.2 Methods

3.2.1 Participants

3.2.2 Stroop Test

3.2.3 Visual Stimulation

3.2.4 Procedure

One of the most important parts to this project was the collection of authentic data that could be used later on to develop a reliable classifier for our system. For that reason a experiment, specifically designed to elicit certain emotional and cognitive states in a subject, was conducted. The following section is focused on the procedure applied in this experiment.

The procedure was comprised of a total of five sessions. Every experiment was initiated with a short briefing session. Containing a short questionnaire, covering personal informa-

tion of the participant as well as habits that may have a influence on the measurement. Further a series of questions, regarding their handedness, use and frequency of use of watches or other wearables was posed, to estimate the additional influence that may be caused by wearing the Empatica E4 wristband. Concluding the first session, the participants were given a coarse outline of the experiment covering the structure and a basic description of their responsibilities.

The second session consisted of a baseline measurement used to log the participants form of the day and also to be able to account for environmental influences in the following processing steps. Before the start of the measurement the subject was placed on a chair in front of a monitor (24 inches, Resolution: 1080p) with a approximated distance of 1m. The Empatica E4 was then put on the wrist of the non-dominant hand and secured in a position that caused minimal light leakage to the PPG-sensor and provided optimal contact for the GSR electrodes. After the participants were comfortable with the device a one minute test sequence was measured to verify the functionality of the system. Consequently the paradigm was displayed on the monitor and the session was started. After reading the instructions, in which the subjects were asked to relax and remain still, and confirmation with the participant the measurement was initiated with a ten second countdown to give some additional time for preparation. During the measurement the GSR, BVP, and temperature of the subject were measured for a duration of five minutes. Afterwards, to conclude the second session, the participants had to give a subjective rating of their current mental state, regarding their stress level, ranging from 1 (completely relaxed) to 10 (stressed out).

The third session was comprised of three separate measurements, two cognitive tasks and a relaxation segment. As before a rating followed the recording. The ratings consisted of a subjective assessment by the subjects regarding their stress level. Additionally subjects had to rate the test difficulty on a scale from 1 (very easy) to 10 (very difficult) for both tasks. For the first measurement the subjects were instructed to count down aloud from 700 in steps of 7 while maintaining a certain pace. The counting rhythm was indicated by a flashing dot on the instruction screen, for a duration of five minutes. The dot's color and flashing frequency were altered during the experiment to further increase difficulty at the three and four minute mark. If the participants were to slow down or loose track an instructor would intervene to help. The second task consisted of a Stroop-Word-Color test. The test featured 11 different colors, resulting in a total of 220 trials the subjects had to work through. Although the color palette seems rather extensive when compared to the standard 3 color variation of the test, this was a conscious decision to guarantee a test time of at least 5 minutes to mitigate monotony. Each trial presented the subject with a colored word in the center of the screen and one possible answer to either side. The participants then had to choose the right answer based on the color of the word.Each decision was recorded via a key press on the keyboard.

3.2.5 Signal Analysis

Heart Rate Variability

GSR

Temperature

4 Results

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5 Discussion

future work: Existing psychophysiological research does not provide adequate information on how potential metrics might be used to regulate mental state in a closedloop environment. (byrne 1996) - and this was not the scope of this thesis we only provide the tools to measure and extract features.

6 Conclusions and Future Work

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

List of Figures

List of Tables

Bibliography

[1] R. Parasuraman. Neuroergonomics: Research and practice. *Theoretical Issues in Ergonomics Science*, 2003.

[2] E. A. Byrne and R. Parasuraman. Psychophyiology and adaptive automation. *Biological Psychology*, 1996.

[3] R. K. Mehta and R. Parasuraman. Neuroergonomics: a review of applications to physical and cognitive work. *frontiers in Human Neuroscience*, 2013.

[4] S. Clarke and C. L. Cooper. *Managing the Risk of Workplace Stress. Health and safety hazards.* Routledge. London and New York, 2004.

[5] R. Parasuraman and Rizzo M. *Neuroergonomics*, volume 3 of *Human-Technology Interaction Series*. Oxford University Press, 2008. ISBN 0195368657, 9780195368659.

[6] J. Allan. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 2007.

[7] E. Peper, R. Harvey, I. Lin, H. Tylova, and D. Moss. Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony? *Biofeedback*, 2007.

[8] Task Force of The European Society of Cardiology, The North American Society of Pacing, and Electrophysiology. Heart rate variability. standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 1996.

[9] X. Wu, V. Kumar, J. R. Quinlan, J. Gosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl Inf Syst*, 2008.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[11] D. C. Rubin and J. M. Talarico. A comparison of dimensional models of emotion:

Evidence fromemotions, prototypical events, autobiographical memories, andwords. *Memory*, 2009.

[12] *E4 wristband from empatica. User's manual*. Empatica, Via Stendhal 36, 20144 Milano (MI). URL www.empatica.com.