Leveraging Motion and Location Tracking for Supporting Cognitive State and Behavior Analysis

Vom Fachbereich Informatik der
Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Natruwissenschaften (Dr. rer.nat.)

genehmigte Dissertaton

von

Agnes Johanna Gruenerbl

Datum der wissenschaftlichen Aussprache: 20.12.2019
Dekan: Prof. Dr. Stefan Deßloch
Berichterstatter: Prof. Dr. Paul Lukowicz
Berichterstatter: Prof. Dr. Reinhold Haux

D386

Published in January 2020 by
Agnes Johanna Gruenerbl

agnes.gruenerbl@dfki.de

‘Small minds are concerned with the extraordinary, great minds with the ordinary.”

Blaise Pascal (1623-1662)

“Never be afraid to try something new. Remember, amateurs built the ark! Professionals built the Titanic!”

Unknown Source

Acknowledgment

Doing a doctoral thesis, specifically if done in parallel to a full-time research job and attempting to get a start-up running, often seems to be like a life's work. Over the years, I asked myself more than once if this will ever work out. Finally, at the end of this process, I would like to think about all those people who have accompanied me over these years and who have helped me to come this far. They are many, and my most sincere gratitude goes to all of them.

Specifically as a woman working in a male-dominated research field, it was not always easy to find the right path and the right words. Therefore, I would like to thank **Prof. Dr. Paul Lukowicz**, my supervisor first of all for giving me this opportunity to work in such an exciting field of pervasive health and pervasive computing, but also for giving me all the freedom I needed to reach the goals of the various projects on my way and not lose faith in me even when I did.
I certainly would also like to thank my second examiner, **Prof. Dr. Reinhold Haux** who was there when I took my first steps in the field of computer science. That is why I am particularly pleased that he is there when my academic path reaches its goal.

Work, as stressful as it can be is never too bad when the working atmosphere is excellent, and this mainly depends on your colleagues and people around you. In our group **Jane** (our secretary and "group mom" and over long periods the only other woman far and wide) has had the most influence on me enjoying working here. Thank you so much, Jane, for helping me so many times to get things done. I also really enjoy our occasional women's nights out :)
Some people say that one spends more time with their colleagues at work than with one's own family. This is certainly true for me. Therefore, I would also like to thank all of my colleagues. First, an foremost, **Gernot** who has shared an office with me for many years, with all the ups and downs that come with working together so closely, and also with my occasional mood swings. :)
Gernot, when someone scans through my publication list, with the number of publications you co-author my work, they can get a picture of how often you helped me out. I do not dare to start counting how often you stepped-in in an emergency to help me get things done and how often you lend me your "wits and brains" when I got stuck in my work. Thank you so very much for all your help and patience, and also thank you for every time you translated what I was thinking into something others could understand :)
To all my other colleagues **Tobi, Andi, Joey, Orkhan (and Leyla :-*), Peter (and Eva), Marco, Jens, both Geralds, Jingyuan, Vitor, George, and all the others**, our group is diverse with people from all parts of the world, and I really enjoy working with all of you.
Hamraz, I was soooo happy when finally another woman joined our group, and I will never forget our trips to Southampton, Singapore, etc., and despite the exhausting days of work, how much fun we had!
Last but certainly not least, I would also like to thank **Olga** from the travel department for taking away so much pressure with all the travels I had to do. All trips were perfectly planned and booked and especially if you have to go to cities and countries you have never been to, it is so much more comfortable and more relaxing when you can trust that things will work out. Thank you, Olga, for all your friendliness and for caring.

Working several hundred kilometers away from family and friends also takes its toll. Thus I would like to thank my family, **my mom and my dad** who in the first place have made my thesis possible by supporting my studies and endeavors over all these years. **My sister and brother, my sister in law** and my little and beloved nephews **Lumpi (Lion), Bobsi (Aaron) and Jani (Julian)** for supporting me all these years and never complaining about me coming and going as I please (or as I can).

To my friends and my second family, **Geli, Hanni**, and all you guys from the **"Oase Maria" Elfriede, Thomas, ...**, I often have less time for you than you would deserve and I would love to spend with you, but knowing that you are out there has often given me the safety I needed to go my way, I do not know what I would do without you ;-*.
Natalia, there is not much that I would not share with you and I do not know what I would do (so far away from home) without knowing that I can talk to you whenever I want, about all my joy and grief and my heartache and what moves me. I will never forget the endless conversations we have/had in your car (because no one can overhear). Even more I appreciate our trips to Medju and all the fun we have there! Because of you, I know that you can have muscle soreness from laughing! I'm forever thankful for your friendship. ;-* xoxo and much more!
Kathrin, working so far away from my friends makes it sometimes difficult to keep friendships alive. But you have always been so flexible to free time for me for a coffee or brunch when I was home, and it was never a big deal when I had no time. Thank you for your friendship (*bighug*).
Lisa, once upon a time I started to work with your husband Fritz and found a friend for life in you. Eventhough we can only meet very rarely, I have you and Leo always in my heart. I could watch you doing your dis-

sertation while having a baby and raising a child, back then you became a real role model for me and I started believing that I could do it as well. xoxos

The entire work I have done during my Ph.D. was always based on "real-life" data-collections and thus has affected the work-life of a lot of people over the years. I want to use this platform to express my gratitude to all of them, as without their patience and support none of the research presented in this doctoral thesis would have been successful. I understand how difficult it must be, especially in the healthcare sector, where work itself is very stressful, to open up to new ideas and allow researchers (with great new ideas that one day might make the life of people working in healthcare easier) to come into their stressful lives and test their theories. Thus, I am grateful to every person my studies ever have touched and whose work has been affected by them:
The study in the nursing home St. Katharina in Vienna was the first study I had ever done fresh after graduating and still learning how to find my way. In this regard, I would specifically like to thank **Sr. Jutta** (the head of the nursing home) for all her support and for always being helpful, even when I showed up several times a week because something did not work. I would also like to thank **Sr. Faustina** (who had nothing really to do with the study but was always so kind to me, young post-graduate, and offered us coffee when we had to wait before we could enter the ward). Of course, special thanks goes to all the nurses in the ward who were affected the most with searching for tags.
Since this was my first study, I often felt overwhelmed by the challenges this study imposed. Thus a very warm thank you goes to **Fritz** (my colleague from back then) who always lent me a helping hand when I did not know what to do next. I will never forget the (kind of desperate) fun we had installing the sensor systems in the nursing home in the middle of the night (in order not to distract the residents), and our "excellence list" is still a highlight.
If you want to work with real patients, specifically if they are psychiatric patients, the cooperation of a psychiatric hospital is essential. And as a psychiatric hospital is no less busy than any other hospital, participating in a study like ours certainly added to the staff's effort. Thus I would even more like to thank all the psychiatrists and nurses and the patients in the wards B1 and B2 of the psychiatric hospital Hall in Tirol for their support, which made the success of this work possible. My very special thank you goes to **Prim. Prof. Dr. Christian Haring**, the head of Psychiatry B. Without your ideas and support, this study would have never been possible. Patricia, Nadia, Raphaela I enjoyed working with you on this project. **Stefan**, doing this study was probably the most difficult one in my entire career. Without your help, I am sure I would have given up several times.
It seemed like a coincidence (actually I do not believe in coincidences) that our group was introduced to the nurse trainers at the health department of the University of Southampton. Ever since then we tried to keep this contact alive because it was amazing for us to work with people of an entirely different field who were so eager to try new things and did not mind testing sometimes crazy devices. **Mary** (Prof. Mary Gobbi) and **Ellie** (Dr. Eloise Monger) I particularity want to thank the two of you for making my life so much more comfortable over the last years, for your kindness and openness throughout all those years, for the fact that I never had to convince you or beg you to participate in our studies. I so much enjoyed working with you, and I hope that we can keep this collaboration alive over the years to come.

A special shout-out goes to all those who helped me in providing their input to this thesis itself, and those who volunteered for proof-reading. **Kai and Gernot**, thank you so much for your input and support!

It is a very rare thing, but sometimes you meet a person and you instantly hit it off. **Johannes**, even-though I just met you a short while ago, you have already had an impact on this thesis :) All the discussions we had made me (sometimes critically) think about my work, and thus influence a very particular part of this thesis. I guess you will immediately understand which one. Thank you for cross reading some parts of this thesis, I value your opinion very much. Thank you =)

Happiness is neither without us nor within us. It is in God, both without us and within us. *(Blaise Pascal)*

Abstract

Activity recognition has continued to be a large field in computer science over the last two decades. Research questions from 15 years ago have led to solutions that today support our daily lives. Specifically, the success of smartphones or more recent developments of other smart devices (e.g., smart-watches) is rooted in applications that leverage on activity analysis and location tracking (fitness applications and maps). Today we can track our physical health and fitness and support our physical needs by merely owning (and using) a smart-phone. Still, the quality of our lives does not solely rely on fitness and physical health but also more increasingly on our mental well-being. Since we have learned how practical and easy it is to have a lot of functions, including health support on just one device, it would be specifically helpful if we could also use the smart-phone to support our mental and cognitive health if need be.

The ultimate goal of this work is to use sensor-assisted location and motion analysis to support various aspects of medically valid cognitive assessments. In this regard, this thesis builds on Hypothesis 3: Sensors in our ubiquitous environment can collect information about our cognitive state, and it is possible to extract that information. In addition, these data can be used to derive complex cognitive states and to predict possible pathological changes in humans. After all, not only is it possible to determine the cognitive state through sensors but also to assist people in difficult situations through these sensors.

Thus, in the first part, this thesis focuses on the detection of mental state and state changes. The primary purpose is to evaluate possible starting points for sensor systems in order to enable a clinically accurate assessment of mental states. These assessments must work on the condition that a developed system must be able to function within the given limits of a real clinical environment. Despite the limitations and challenges of real-life deployments, it was possible to develop methods for determining the cognitive state and well-being of the residents. The analysis of the location data provides a correct classification of cognitive state with an average accuracy of 70% to 90%. Methods to determine the state of bipolar patients provide an accuracy of 70-80% for the detection of different cognitive states (total seven classes) using single sensors and 76% for merging data from different sensors. Methods for detecting the occurrence of state changes, a highlight of this work, even achieved a precision and recall of 95%. The comparison of these results with currently used standard methods in psychiatric care even shows a clear advantage of the sensor-based method. The accuracy of the sensor-based analysis is 60% higher than the accuracy of the currently used methods.

The second part of this thesis introduces methods to support people's actions in stressful situations on the one hand and analyzes the interaction between people during high-pressure activities on the other. A simple, acceleration based, smartwatch instant feedback application was used to help laypeople to learn to perform CPR (cardiopulmonary resuscitation) in an emergency on the fly. The evaluation of this application in a study with 43 laypersons showed an instant improvement in the CPR performance of 50%. An investigation of whether training with such an instant feedback device can support improved learning and lead to more permanent effects for gaining skills was able to confirm this theory. Last but not least, with the main interest shifting from the individual to a group of people at the end of this work, the question: how can we determine the interaction between individuals within a group of people? was answered by developing a methodology to detect un-voiced collaboration in random ad-hoc groups. An evaluation with data retrieved from video footage provides an accuracy of up to more than 95%, and even with artificially introduced errors rates of 20%, still an accuracy of 70% precision, and 90% recall can be achieved.

All scenarios in this thesis address different practical issues of today's health care. The methods developed are based on real-life datasets and real-world studies.

Zusammenfassung

Die automatische Erkennung von Aktivitäten stellte in den letzten zwei Jahrzehnten ein großes Feld in der Informatik dar. Forschungsfragen von vor 15 Jahren haben zu Lösungen geführt, die heute unser tägliches Leben unterstützen. Insbesondere der Erfolg von Smartphones oder anderer intelligenter Geräte (z. B. Smartwatches) beruht auf Anwendungen, die Aktivitätsanalysen und Standortverfolgung nutzen (Fitnessanwendungen und Karten). Heute können wir unsere körperliche Gesundheit und Fitness nachverfolgen und unsere körperlichen Bedürfnisse unterstützen, indem wir lediglich ein Smartphone besitzen (und verwenden). Die Qualität unseres Lebens hängt jedoch nicht nur von Fitness und körperlicher Gesundheit ab, sondern zunehmend auch von unserem geistigen Wohlbefinden. Da wir gelernt haben, wie praktisch und einfach es ist, viele Funktionen, einschließlich der Gesundheitsunterstützung, auf nur einem Gerät zu haben, wäre es besonders hilfreich, wenn wir bei Bedarf auch das Smartphone zur Unterstützung unserer geistigen und kognitiven Gesundheit verwenden könnten.

Das ultimative Ziel dieser Arbeit ist es, sensorgestützte Standort- und Bewegungsanalysen zu verwenden, um verschiedene Aspekte medizinisch valider kognitiver Beurteilungen zu unterstützen. In dieser Hinsicht baut diese These auf der Hypothese 3 auf: Sensoren in unserer allgegenwärtigen Umgebung können Informationen über unseren kognitiven Zustand sammeln, und es ist möglich, diese Informationen zu extrahieren. Darüber hinaus können diese Daten verwendet werden, um komplexe kognitive Zustände abzuleiten und mögliche pathologische Veränderungen beim Menschen vorherzusagen. Schließlich ist es nicht nur möglich, den kognitiven Zustand durch Sensoren zu bestimmen, sondern auch Menschen in schwierigen Situationen durch diese Sensoren zu unterstützen.

Im ersten Teil der Arbeit geht es darum, mentale Zustände und Zustandsänderungen zu erkennen. Hauptzweck ist es, mögliche ertse Ansatzpunkte für Sensorsysteme zu evaluieren, um eine klinisch genaue Beurteilung der mentalen Zustände zu ermöglichen. Diese Bewertungen müssen unter der Voraussetzung funktionieren, dass ein entwickeltes System in der Lage sein muss, innerhalb der vorgegebenen Grenzen einer realen klinischen Umgebung zu funktionieren. Trotz der Einschränkungen und Herausforderungen realer Einsätze war es möglich, Methoden zur Bestimmung des kognitiven Zustands und des Wohlbefindens der Bewohner zu entwickeln. Die Analyse der Daten liefert eine korrekte Klassifizierung des kognitiven Zustands mit einer durchschnittlichen Genauigkeit von 70-90%. Methoden zur Bestimmung des Zustands von bipolaren Patienten liefern eine Genauigkeit von 70-80% für die Erkennung verschiedener kognitiver Zustände (insgesamt sieben Klassen) unter Verwendung einzelner Sensoren und 76% für die Zusammenführung von Daten von verschiedenen Sensoren. Methoden zur Erkennung des Auftretens von Zustandsänderungen, ein Highlight dieser Arbeit, erreichten sogar eine Genauigkeit und Recall von 95%. Der Vergleich dieser Ergebnisse mit gängigen Standardmethoden in der Psychiatrie zeigt sogar einen klaren Vorteil der sensorgestützten Methode. Die Genauigkeit der sensorgestützten Analyse ist 60 Prozentpunkte höher als die Genauigkeit der derzeit verwendeten Methoden.

Der zweite Teil dieser Arbeit stellt Methoden vor, um das Handeln von Menschen in Stresssituationen zu unterstützen, und die Interaktion zwischen Menschen bei Hochdruckaktivitäten zu analyisieren. Mithilfe einer einfachen, beschleunigungsbasierten Smartwatch-Anwendung mit sofortigem Feedback konnten Laien lernen, wie sie im Notfall im Handumdrehen HLW (Herz-Lungen-Wiederbelebung) durchführen können. Die Bewertung dieser Anwendung in einer Studie mit 43 Laien ergab eine sofortige Verbesserung der HLW-Leistung von 50%. Eine Untersuchung, ob das Training mit einem solchen Sofort-Feedback-Gerät ein verbessertes Lernen unterstützen und zu dauerhafteren Effekten für den Erwerb von Fähigkeiten führen kann, konnte diese Theorie bestätigen. Da sich das Hauptinteresse am Ende dieser Arbeit von einem Individuum zu einer Gruppe von Menschen verlagert, stellt sich die Frage: Wie können wir die Interaktion zwischen Individuen innerhalb einer Gruppe von Menschen bestimmen?, wurde durch die Entwicklung einer Methode zur Erkennung der nichtverbalen Zusammenarbeit in zufälligen Ad-hoc-Gruppen beantwortet. Eine Auswertung mit Daten, die aus Videomaterial abgerufen wurden, liefert eine Genauigkeit von mehr als 95%, und selbst bei künstlich eingeführten Fehlerraten von 20% kann eine Genauigkeit von 70% und ein Recall von 90% erreicht werden.

Alle Szenarien in dieser Arbeit befassen sich mit verschiedenen praktischen Fragen der heutigen Gesundheitsversorgung. Die entwickelten Methoden basieren auf realen Datensätzen und realen Studien.

Contents

Recognizing Cognitive State Supporting Activities in Cognitively Stressfull Situations

Over the last decades, life in our society has become rather hectic. Competition and pressure in our work environment have increased dramatically. At the same time, more and more people are suffering from “burnout" or other mental affective disorders like depression. Whiteford et al. [1] have estimated that by 2030 affective disorders will contribute to the highest disease burden in the developed world. According to the World Health Organization, in 2018, more than 300 million people worldwide suffered from depression-like disorders [2]. Further, a study by Oelseon et al. [3, 4] projected in 2010, the total cost of brain disorders (mental and neurological) in Europe to be €798 billion. 40% of these costs were determined to be indirect costs. Which means they are not costs for medical treatment, but costs of incapacity and loss of workforce, suicides, unemployment, and others.

These numbers listed above, which were published by official bodies, public organizations, and health insurances, call for action. Nevertheless, the question is, where to start. Illnesses such as burnout or depression hardly come overnight. Most manifestations are the result of ever-increasing pressure in the work environment, but they can also be the result of mental stress and pressure in our society. To combat this issues, applications that, on the one hand, would be able to help to monitor and to measure our mental burden or track and assess our cognitive and psychological states, which are undoubtedly complex and thus tricky to self-assess, would be a significant support to prevent crises. On the other hand, once a crisis (e.g., burnout) occurs, there is a specific need for additional support to keep the adverse effects minimal and moreover prevent such crises from happening ever again.

A similar pattern of issues arises in elderly care. Significant developments in medicine contribute to the fact that our life expectancy is steadily increasing. With people living longer though, age-related diseases like dementia are on the rise. At the same time, our family structures are changing. Today older people tend to live alone or in small communities instead of within larger families where, in the case of dementia, they could get the necessary care. Understandably, of course, people who have lived all their lives independently, want to stay this way as long as they possibly can. Thus, applications that would allow determining the well-being of dementia patients would help to assure an extended independent life.

A third example, highlighting why the assessment of cognitive state is becoming more relevant today, again regards our work society. In general, in many professions the developments go in the direction of more control and liability. This issue has become imminent in recent years after some incidents (reported in the news). These have stirred public discussions in some areas about how to make sure that teachers are fit to teach young adolescents and handle potentially tricky situations.
The same is true for nursing schools. Teachers are trying to figure out how they could objectively judge whether a trainee nurse is ready to care for real patients [5]. For both groups, it is essential to ask these questions, as failure can have severe consequences. Thus, research is challenged to find methods to objectively but also emphatically monitor and analyze the way young trainee nurses or soon to be teachers or other young groups act and interact in their professions.

1.1 Motivation

These three examples are practical cases, taken from our society. Though they might seem random at first, they do have essential aspects in common. All three examples display situations that currently have to rely on the experience of experts and their subjective perception of whether something is happening or not. E.g., whether someone is ready to act, or someone is depressed, etc. Thus, all three cases miss the opportunity of objective and independent assessments. To gain such objectivity would require methodologies and devices that were able to assess complex cognitive states and understand behavior on a long-term basis.

Of course, there is EEG (electroencephalogram), the typical way to measure brain wave patterns. Mobile versions of EEGs are already available [6, 7]. Nevertheless, EEG is not practical in the long-run and everyday life. Regarding the three examples provided above it would require many people, in the range of hundreds or thousands, to wear EEGs all day long for extended periods. An everyday situation where lots of people are walking around wearing electrodes on their heads would seem odd, like a scene from a bad SciFi movie. More importantly, though, the visible presence of such devices can be traced back to a mental or cognitive disability very easily. This would significantly violate a person's privacy and a person's right to decide about their health. Thus, there has to be another way to empower people to manage their cognitive states. Today, 2019, there is research ongoing to make EEG less obtrusive, as in Kosmyna et al. [8], for example, who inserted the relevant technology into smart-glasses. However, these experiments are still in a developmental stage. In addition, various studies have shown that patient compliance is highly influenced by the need to use additional equipment. Thus, the use of smart glasses can only be guaranteed by people who need glasses.
Research in the field of long-term cognitive assessments is challenged to find sensor solutions that can be integrated into people's lives unobtrusively and invisibly! Ideally, such sensors would fit smoothly into the user's life, and even "vanish" from the user's perception. Thus the behavior of the user would not be influenced or biased by the sensors. Moreover, these sensor set-ups would have to be able to work on a long-term basis, but all the same, only require minimal maintenance, since increased required effort notoriously goes together with reduced compliance.

Glancing over into activity recognition reveals that research in the past few decades already has achieved for our physical health, what we now require for our cognitive well-being. Twenty years ago researchers started to attach single IMUs to the leg or foot of a person. A direct outcome of these endeavors are conveniences like step-counters. Today, these conveniences are default applications on smartphones or smart-watches. Particularly, smart-phones or more recently other smart devices owe their success to applications that leverage activity recognition (various fitness applications) and location tracking (maps, outdoor fitness). It is safe to say, today we can track our fitness and support our physical health individually. All we need is to own (and of course use) a . Very recent research, even started using low-cost motion sensors to provide opportunistic heart-rate assessments from ballistocardiographic signals during restful periods of daily life (see Hernandez et al. [9]).
Precisely such functionalities were and are still missing for our mental and cognitive health. As hinted above, never before mental disorders and cognitive stress have had such a defining role with such effects as in our society. Nevertheless, for many reasons, mental health has not yet been promoted in research and our society the way physical well-being has. As of now, we have learned from activity recognition how practical and easy it is to have many functions, including health support on just one device. Thus it would be specifically helpful if we could also use our smart-phone or at least devices that already exist to support our mental and cognitive health.

1.2 The Overall Thesis Objectives

Given the increasing need to address pressure and its impact on our lives, this dissertation aims to find sensor-based solutions that help people manage their cognitive health.
More specifically, our everyday life has already been equipped with a variety of sensors and smart devices. These sensors will not disappear soon. So all these "omnipresent" sensors could be used to gather information about our cognitive states and (possibly unhealthy) state changes when we need them.

In addition, the sensors in our environment are becoming smaller and more pervasive. Often, we no longer perceive them as sensors, which in turn means that we naturally act in their vicinity. Natural action is an essential criterion for this work! When people believe they are being monitored (by someone or something), they tend to act more consciously and less naturally. However, to analyze cognitive behavior, it is important to analyze unbiased behavioral patterns. To ensure a natural behavior, sensor systems should be used that have established a broad availability in our lives. If such sensors are not available, the sensors to be used should at least be integrable into our lives.

Following these preconditions, the work of this dissertation is based on three main hypotheses:

- **Hypothesis 1:** Devices in our pervasive environment can collect information about our cognitive state. Furthermore, it is possible to extract this information from these sensors.
- **Hypothesis 2:** Given that the first hypothesis is correct, it is also possible to infer complex cognitive states and possibly predict or detect pathological changes from information extracted from pervasive sensors.
- **Hypothesis 3:** Again, assuming a positive result of hypothesis one and two, not only can the cognitive state be determined by given sensor systems, but also sensors (or suitable devices including such sensors) can be used to assist people in difficult situations.

After formulating these hypotheses, the goals of this thesis details as follows:

1. Identify scenarios where cognitive state assessment could bring real benefits. Determine appropriate sensor systems that promise to hold information, relevant to these scenarios.
2. Qualitatively analyze the information extracted from the sensor systems determined above. Evaluate what information retrieved from these sensor systems can contribute to cognitive state assessments.
3. Based on sensor systems and information identified, develop and evaluate clinically valid methods to infer cognitive state and state change.
4. Identify scenarios where people in stress situations could benefit from direct support from sensor systems. Other potentially interesting scenarios are those in which sensor use could provide new relevant insights into a situation or a specific scenario.
5. Identify sensor systems suitable for helping humans in the scenarios identified above, and develop and evaluate appropriate methods.

1.3 State of Research

At the beginning of the research, from which this dissertation would ultimately emerge, many related research areas were in the process of being formed. Most of them, however, were only in very early stages of development. Parallel to the efforts in this dissertation, these research fields developed in the following years to become extensive research areas.

This section on the state of research attempts to give an idea of what the initial situation in the research looked like when certain parts of this work began. This section is less about how the field looks today. Please note, therefore, that this section should provide a better understanding of why the topic of this dissertation was formed and what, moreover, specifies its work and its relevance. This section also attempts to focus on the overall picture and therefore does not address the individual parts and topics of this thesis. Detailed related work on each specific topic is provided in each chapter.

1.3.1 Activity Recognition

During the last 10-15 years, activity and location tracking has been at its height. Liao et al. [10] provide one of many approaches regarding location-based activity recognition. Their approach takes high-level context into account to detect significant locations of a person's day. From them, they infer low-level activities. On the front of human activity recognition with sensors, Lara et al. [11] provide a more recent survey. It compares 28 respective systems qualitatively by a number of essential design issues as response time, learning approach, obtrusiveness, accuracy, and others. Work in this field, however, started much earlier, in the early 2000s. In 2001 for example, Mantyjariv et al. [12] introduced initial experiments with acceleration sensors for recognizing human activities. Then, in the first decade of the 21st century, groups were still using self-made sensors, e.g., as in Choudhury et al. [13]. Today, these kinds of sensors come as a "side-effect" with all smart-phones.

Activity recognition over the years has been evaluated in various states of complexity and various settings. If activity recognition is deployed for real-life applications, then simple sensor setups are more suitable, most of the work thus focuses on recognizing basic motions like "walking", "sitting", "standing" or "running" (Biever et al. [14]) or fitness and gym-machine exercises (e.g., Muehlbauer et al. [15] or Seeger et al. [16]).

To recognize more complex activities, e.g., in maintenance work, often a more significant number of sensors is required to get satisfactory results (see Ogris et al. [17] and Zinnen et al. [18]). In an even more complex environment, sensor information alone is not enough and requires additional sources of information. Thus methods become necessary to combine different sources of information. Examples here are, hidden Markov models (HMMs) [19], conditional random fields (CRFs - Hue et al. [20]) or probabilistic plan recognition (Geib et al. [21]). These use the model-based information to support the process. Other methods from more recent years transfer the knowledge learned from one application to another (Blanke et al. [22]).

1.3.2 Activity Recognition in Health Care

Activity recognition and the use of pervasive technology in healthcare also have become a wide field. A variety of publications has tried to provide an overview of potential applications of activity recognition in healthcare.

These include Lukowicz [23] or Orwat [24], for example. In healthcare specifically, pervasive computing and context recognition generally includes context-aware systems, which intend to provide information, e.g., for care documentation. For example, Agarwal et al. [25] describe a prototype context-aware information system to capture and interpret data in an operating room of the future. In [26], Cheng et al. introduce a study about a nursing support system in a series of laboratory experiments under simulated conditions. Naya et al. [27] describe a sensor network system for supporting context-awareness of nursing activities in hospitals.

1.3.3 Assisting An Aging Population

Within the middle of the first decade of the third millennium, work towards assisting the aging population with cognitive impairment appeared. Pollack [28] for example summarizes the, by 2005, existing technologies for supporting the elderly. These include assurance as well as guidance and assessment systems, and also management schedules. From this time onwards, most research in this area of intelligent technology for the elderly was mainly going in the direction of smart-home based approaches. Earlier work in the late 2000s starts with Hayes et al. [29], over Kaye et al. [30] to the more recent publications of Dawadi et al. [31] in 2013.

1.3.4 Cognitive State, Emotion and Stress

Scanning through literature with key-words like “assessing cognitive state and behavior" reveals different kinds of publications. First of all, there are many somewhat older (clearly before 2010) publications about measuring cognitive workloads and states. In this regard Burken et al. [32] and Marshall et al. [33] are work from the early 2000s on measuring cognitive work loads. While Marshall’s work[33] uses a pupil dilation measurement technique, Brunken et al. [32] discuss different methods of assessing cognitive load with visual monitoring. Still, both have in common that neither is mobile or fitting for long-term everyday use.
Even earlier, in 1997, Picard [34] hinted that computers might soon be given the ability to have emotions. Today, 20 years later this has not yet been fully achieved. Nevertheless, many research groups have worked on affective computing and recognizing stress and emotions. So for example, Kort et al. [35] introduce work on a digital learning platform that is supposed to track the affective state of the learner and respond correspondingly. In [36], Picard et al. argue for the importance that machine intelligence should incorporate emotional skills and be able to recognize human emotions and emotional communication between humans. They also admit though, that one of the main difficulties in this field is gathering data that would represent real human emotions (not “faked” emotions by hired actors). Thus in [36] they also try to provide different factors that could help in this process. Picard et al. go on analyzing different aspects of emotions and also presenting a classification of emotions based on physiological data. To evaluate these results, they used several weeks of sensor data from one person collected by a number of physiological sensors: electromyogram, facial muscle tension, photoplethysmograph, blood volume pressure, skin conductance, and electrodermal activity.

Following the work on recognizing human emotions Healey et al. [37] have collected different physiological parameters of car drivers (ECG, EMG, skin conductance and respiration) during real-life car drives in order to determine their stress level with very high accuracy.
A further system in this line, which is supposed to be both, capable of long-term monitoring and also being mobile, is the Physiological Sensor Suite introduced by Matthews et al. [38]. This platform intends to measure physiological and cognitive states by providing "wearable" sensor platforms incorporating ECG, EMG, EOG and EEG sensor. Despite being small enough to be wearable and usable in a mobile manner, this platform is still an external sensor-system, that does not necessarily fit into a person’s life. Russo et al. [39] state that decision making and situation awareness are critical mental abilities for cognitive performance. In this line, they introduce an approach, based on physiological parameters, to examine cognitive states and predict operator fatigue. In more recent work, Setz et al. [40] measure and distinguish stress from cognitive load based on electrodermal activity (EDA).

A field towards emotion recognition that started around the middle 2000s has set as its goal to determine emotion from motion. One example is Barry et al. [41] and [42] who inferred emotions from Butoh Dance performances. Butoh Dance is a form of dancing that intends to express emotions via the entire body. Barry et al. used acceleration and magnetic field sensors, placed at the arms and legs of Butoh dancers and recorded their performances. Using HMMs, they were then successfully able to determine the particular emotion that was expressed by the dancer.
Another group around Crane et al. [43] has shown in a laboratory setting that basic emotion can be determined by movements and that body movement are affected by the emotion the person feels. This work is based on video and motion capture data that was collected from many test subjects. The videos were shown to other test-subjects (that had not been part of the emotion-walk recording), who had to determine the emotion (pick one out of ten) that the presumed to have seen in each of the videos. In [44] McDonnell at al. analyze the emotional content of motions with real and virtual replicas of an actor exhibiting six basic emotions. In addition to the actual actor, the actions were applied to five virtual body shapes. From participants asked to rate the emotions expressed, it showed that the perception of emotional is independent of the character’s body type.

An interesting, quite recent (2017) take to influence cognitive performance, comes from Amores et al. [45]. This paper discusses the role of smell to affect one’s mood and cognitive performance while being asleep or awake and introduces a first pervasive device to manually or au-

tomatically to release subtle bursts of scent. However, this work is at its very beginning and is not yet in the stage to be integrated into a persons live-style.
Also a relatively recent approach on the cognitive "quantified self" has been introduced by Kunze et al. [46]. With focus on reading habits, a prototype cognitive activity recognition system was developed that monitors what and how much users read, as well as how much they understand. They also hypothesize that such systems could revolutionize teaching, learning, and assessment both inside and outside the classroom.

1.3.5 Face Expression Recognition

Research on automatically detecting face and facial expression already started in the early 90's. Samal et al. [47] provides a survey of early algorithms. These early works though, generally focus on recognizing faces out of images. Later in the decade and the early 2000's work in face recognition progressed to recognizing emotions in facial expressions. Pantic et al. [48] for example compare different, in the early 2000 ongoing approaches of facial expression detection and classification in static images and image sequences. In further research, face expression recognition progressed from a simple static image to moved image face recognition. Busso et al. [49] for example, compare emotion recognition with marker-based motion capture to simultaneous speech recordings. The results show that facial expressions have a better performance. They also evaluate the performance of emotion recognition by combining both modalities, leading to improved robustness for emotion recognition.

All the above systems are more or less systems that analyze physiological parameters to determine cognitive states. When it comes to reviewing the literature on the analysis of cognitive state in the sense of analyzing how a person feels or behaves, there is significantly less work available. Today, 18 years into the 21st century, indeed more research can be identified in this field then five years ago, but still, most of them have just started. One example is introduced by Masai et al. [50] and [51], who present a novel smart eyewear that recognizes a wearer's facial expressions with the still distant goal to detect facial expressions related to cognitive loads such as attention, interest, fatigue, and concentration.

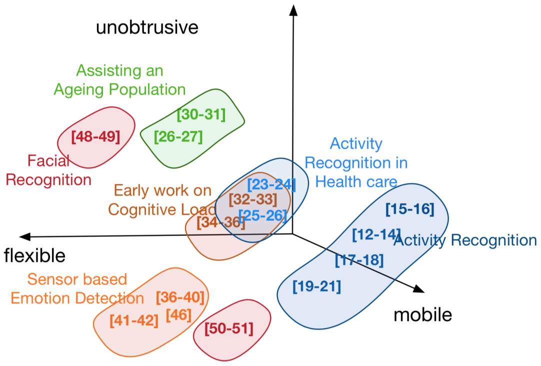


Figure 1.1: Available research and its location in regards of flexibility, unobtrusivenes and mobility at the beginning of this work.

1.4 Thesis Contribution

Despite the diverse research touching the field of cognitive state assessments, none of the applications and papers described in the State of Research above match the exact goals and purposes of this dissertation. The main drawback of the applications described above is that none of them is designed to fulfill all the essential requirements of this thesis. See Figure /reffig:relwork to get a picture where research was located in terms of flexibility, mobility, and unobtrusiveness.

Different methods of monitoring cognitive state or assessing emotions are stationary, like the face recognition or emotion recognition via camera. Thus they cannot accompany a person throughout their day. Other systems, like measuring physiological parameters are in theory wearable or apt for real-life, e.g., while driving a car or when using wearable ECG. Nevertheless, in practical terms, they are just not applicable in the long run and everyday life, at least not in the state as they are in now. On the other hand, systems designed for long-term use, like support systems for the elderly, are not intended as monitoring systems. Hence, they can support a person's ability to handle activities of daily living but do not address the monitoring of cognitive state.

However, the aim of the present work is to develop methods for the long-term support of cognitive state management, which in turn would also fit the actual lifestyle of a person. In this dissertation, especially in the chapters where the mobility of the sensor system is less critical, some of the above applications will be considered. Nevertheless, all work done must meet the user's life requirements. The sensor systems to be used must, therefore, be **flexible, mobile, and unobtrusive**.

Following these requirements and following extensive literature analyzes, it was decided to use location and activity tracking in particular. Both are already part of our lives because of their availability in our smartphones. They are also mobile, unobtrusive, and flexible. All the work in this dissertation will be built up step by step. It begins with coarse detection of dementia conditions with an indoor location sensor system and within a limited area. The detection of state and state changes of affective disorder patients follows with more complex sensor systems. In the second part, this thesis goes as far as assisting people in emergencies. The main contributions of this work thus can be summarized as follows:

- A large part of this thesis deals with extracting medically relevant information from commercially available sensor systems. These sensor systems, however, have not been designed for the use in medical environments but everyday life. Hence, there is no guarantee that they will work as intended.

In addition, many scenarios from health-care and mental-care can hardly be simulated. For example, it is a great challenge even for actors to reliably portray a cognitively impaired person. Therefore, laboratory simulations in such medical environments cannot provide the required data quality. Moreover, a psychiatrist will hardly trust a system that was developed using data of people pretending to be manic or depressed. Thus, most parts of this thesis require data acquisitions in real-life from actual patients. Such real-life studies and real-world data recordings, however, especially in sensitive and stressful environments as health and mental health care, represent a critical number of challenges that need to be addressed. Hence, one specific contribution of this thesis is to demonstrate **how to manage real-life data study deployments and more importantly, how to overcome their restrictions and limitations.**

Further challenges of real-life data collection studies in health care include that recorded data will be incomplete. In real-life data recordings with real patients, it is not possible to monitor whether the sensor systems will be operated correctly. Moreover, it is not allowed to force patients to use sensors in the desired frequency and manner. Therefore, another specific contribution of this dissertation is to demonstrate **how to deal with gaps in the database and how to perform statistic analyses on data that is only sporadically available or largely missing.**

- Most of this work especially chapters 2 through 4, address actual medical needs. However, research in the medical field requires an experimental and empirical strategy. On the one hand, this is a requirement in this field in order to obtain the necessary licenses for medical studies. On the other hand, at the beginning of this dissertation, there was mainly just an idea of what might be possible, and the assumption that a particular type of sensor data might contain the required information. However, this had to be confirmed. Therefore, first, it was necessary to evaluate the possible information content of available sensor data empirically.
  In this regard, another specific contribution of this paper is to demonstrate **how to analyze sensor data empirically and to identify whatever relevant information is enclosed therein.** In particular, this concerns sensor data that has been recorded in fuzzy or disruptive and error-prone environments. The evaluation following the empirical analysis then shows **how to extract the required information and provide a reliable evaluation that confirms the initial assumptions.**

- After the initial empirical approach, this thesis, step by step, provides and evaluates solutions on **how to map the extracted data onto a variety of cognitive states and disease patterns and further draw to conclusions from them.** This includes data from different sensory modalities and in various states of complexity and granularity. To provide a concrete example: this thesis starts, in chapter 2, with evaluating movement patterns of persons, grouped on one or biweekly scales, within a spatially restricted, flat like environment. It then continues in chapter 3 to extend the settings and to evaluate movement patterns of persons on a daily scale without any spatial restriction. Another example: chapter 3 will confirm that activity data, when daily recorded over weeks on a patient's, can be used to determine the mental state of a bipolar patient. However, activity data also can be used on the spot to determine if a person explicitly performs emergency heart compressions correctly.

- Chapter 4 provides a comparison of the sensor-based methodologies for determining the cognitive state of patients and the current medical standard of self-assessment. This comparison successfully demonstrates that **the sensor-based method to determine the mental state of a patient is more accurate than the currently used self-assessments.** A further highlight is that, for the first time, **a particular subjective observation made by doctors over the years but never seriously evaluated, is objectively confirmed.**

- Additionally, this thesis does not only show how to map sensors onto cognitive patterns but also introduces an example of **how to use sensors to support the skills of people to act in an emergency.** For example, work in Chapter 5, the CPR scenario, shows that persons without any medical training can perform heart compressions correctly when using a smart-watch feedback assistant. Interviews also revealed that **these persons felt more confident and were eager to act when using the assistance system.**

1.5 Outline and Structure of this Thesis

The outline of this dissertation splits into two major parts. Chapters 2-4 form the first part. They address the question of whether it is possible to extract information regarding the cognitive state of a person from location and activity data. The second part in Chapters 5-6 evaluates whether it is possible to leverage devices to support the action of persons in stressful situations.

In **Chapter 2**, the cognitive state and well-being of dementia patients are assessed. This work is based on one sensor system, Indoor Location Tracking, within a limited space. The main question in this chapter is whether it is possible to evaluate a complex construct such as the cognitive state of a person with simple sensor setups. The proposed study is able to draw from a one-year indoor location data set of 6 persons, who live in an apartment-like ward of a nursing home. Being a long-term deployment of sensors in real life, collecting sufficient ground-truths poses a specific difficulty. Thus, this chapter also deals with the challenge of dealing with the limited availability of ground-truths in real-life studies. Despite some limitations, methods for deriving the cognitive state of residents are developed. The evaluation of the collected data provides a correct classification of the cognitive state with an average accuracy of 70% to 90%, thus indicating that a stationary sensor modality could suffice to monitor the well-being of dementia patients.

Building on a successful outcome of Chapter 2, specifically on the lessons learned, in **Chapter 3**, the aim is to determine the state of patients with disorders that affect the cognitive state. This scenario comprises more complexity in the set-up as affective disorders are more complex than dementia. The sensor-setup to be used has to be mobile and must not be limited to rooms. Furthermore, location alone is not sufficient as affective disorder is not a one-directed disorder (like dementia) but includes frequent state changes. Therefore the challenge of this particular scenario is to determine in which mental state a patient is in at present, but also to detect when the state starts to change.
Thus, methods to recognize the state of people who live with bipolar disorder (manic-depressive disorder) on the one hand and to detect the onset of changes in the condition of these patients, on the other, should be developed. While the the previous chapter of this thesis works with only one sensor system within a restricted area, this chapter has to extend the number of sensors used and also to reduce the limitations in space. As a suitable sensor modality the internal sensors of the are determined. In order to achieve the goals despite the increasing complexity, tasks in this chapter evolve step by step. The efforts in this section start with a feasibility analysis that aims at evaluating the correlation of sensor traces of 6 bipolar patients with their self-assessment (the standard technique in psychiatry). In addition, various features and trends in the sensor data are empirically analyzed and related to their actual diagnosis.
Based on the findings of this do-ability study, a more extensive data collection study is conducted afterward, for ten bipolar patients and over several months. Methods, developed to determine the mental health of patients using the acquired sensor data (activity, location, and social interaction), provide a classification accuracy of 70-80% (7 classes in total) when using single sensor-traces. The classification of fused sensor-data leads to an accuracy of 76%. Methods for detecting state changes, a highlight of this work, achieve precision and recall of about 95%.

Chapter 4, is an extension of Chapter 3. The work on recognizing mental state of bipolar patients, and in particular the detection of the onset of state-changes, is of practical relevance in psychiatric care. Thus, the question arises as to whether sensor-based detection can keep up with current psychiatric standards for monitoring state and condition changes. For this purpose, similarity analysis of the two methods (sensor-based, self-assessments) is carried out, each in comparison with the actual diagnosis. The result of this analysis indicates a clear superiority of the sensor-assisted state analysis. The accuracy of the sensor-based method is 60% higher than of the self-assessment.

Chapter 5 of this dissertation leaves the field of basic recognition of the cognitive state. In the following, this work deals with the question of how sensors or portable devices can support people in stressful situations. CPR (cardiopulmonary resuscitation) performed by a layperson is such a stressful situation. A study with laypeople using a simple acceleration based smart-watch feedback application shows that even people without any prior experience or training are able to perform CPR effectively. At the same time, the users of the smart-watch application are feeling more confident in performing CPR.
The evaluation of the effect of this instant feedback device on gaining immediate skills in resuscitation has impressive results. More than 50% of the laypersons were able to gain enough skills while using the watch-assistant to perform effective CPR for more than 50% of the time. This results lead to a second question: when people can learn on the fly when using the feedback-assistance, can they also learn to perform CPR permanently. Meaning, does training with the CPR-watch stick? Thus, in a second study, it is evaluated whether people would be able to train CPR more effectively when receiving instant-feedback form a wearable device in comparison to traditional human teaching classes.

Chapter 6 changes the perspective. After recognizing the cognitive state of individual people with varying degrees of complexity, and supporting confidence and skills, the final chapter deals with how to determine how people interact in groups to perform a task in a high-pressure situation. In particular, this section is about an ad hoc group of people who need to come together by chance to solve a problem without first assigning roles and responsibilities. Trainee nurses in an emergency simulation are an example of such a scenario.
The chapters of this work, so far, have shown that it is possible to determine the cognitive state of individuals using sensor traces. Therefore, this part deliberately does not focus on recognizing the behavior of the individuals in the group but evaluates the larger image to see when people interact or work together to achieve a goal. The method for determining collaboration is based on a hierarchical tree model. It is being evaluated with an increasing error in the underlying detection of basic actions. The results of detecting collaboration between persons are in the range of 70-90% depending on the recognition error of the basic per person actions (up to 20% error).

This dissertation closes with a general discussion of the results of this thesis in the context of pervasive computing and practical application in our society and will also provide an outlook on how this work should and will be pursued.

1.6 Overview of Selected Publications per Chapter

The following table lists my most important publications on which each chapter builds. For a complete overview of relevant publications, consolidate each chapter individually:

Chapter 2	Indoor-Location based Tracking of Cognitive State of Dementia Patients

- Gruenerbl A. et al. (2011), Using Indoor Location to Assess the State of Dementia Patients: Results and Experience Report from a Long Term, Real World Study. In Proc. of the 7th International Conference on Intelligent Environments, Nottingham 2011.
- Gruenerbl A. et al. (2013). UWB indoor location for monitoring dementia patients: The challenges and perception of a real-life deployment. International Journal of Ambient Computing and Intelligence (IJACI).

Chapter 3	Smartphones based Detection of State and State Changes of Psychiatric Patients

- Gruenerbl A. et al. "Towards smart phone based monitoring of bipolar disorder." Proc. of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare. ACM, 2012.
- Gruenerbl A. et al., Smart-phone Based Recognition of States and State Changes in Bipolar Disorder Patients, IEEE Journal of Biomedical and Health Informatics (J-BHI), 19, 140-148 (2014)
- Gruenerbl A. et al. "Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients." Proceedings of the 5th Augmented Human International Conference (AH 2014). ACM, 2014. - **Honorable Mention Award**

Chapter 4	Smart phone based Objective Sensing or Subjective Self-Assessment

- Gruenerbl A. et al. “Sensor vs. Human: Comparing Sensor Based State Monitoring with Questionnaire Based Self-Assessment in Bipolar Disorder Patients”. In: Proceedings of the 18th International Symposium on Wearable Computers. IEEE International Symposium on Wearable Computers (ISWC-2014), September 13-17, Seattle, Washington, USA, ACM, 9/2014.
- Gruenerbl A. et al. “Assessing Delayed Self-Perception in Bipolar Disorder Patients.” (to be submitted) International journal of bipolar disorders, Springer Heidelberg, 2018.

Chapter 5	Shaping Emergency Behavior using Smart-Watches

- Gruenerbl A. et al. "Smart-watch Life Saver: Smart-watch Interactive-feedback System for Improving Bystander CPR". In: Proceedings of the 2015 ACM International Symposium on Wearable Computers. IEEE International Symposium on Wearable Computers (ISWC), September 9-11, Osaka, Japan, Pages 19-26, ISWC ’15, ISBN 978-1-4503-3578-2, ACM, 2015. - **Best Paper Award**
- Gruenerbl A. et al. "Training CPR with a Wearable Real Time Feedback System". submitted at: Proceedings of the 2018 ACM International Symposium on Wearable Computers. IEEE International Symposium on Wearable Computers (ISWC), October 8-12, Singapore, ISWC ’18, ACM, 2018.

Chapter 6	Detecting Collaboration in Emergency Care with Activity Data

- Gruenerbl A. et al. "Detecting spontaneous collaboration in dynamic group activities from noisy individual activity data". In Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on. IEEE, 279-284.
- Bahle G., Gruenerbl A. et al. "From Individual Activity Recognition to Unscripted Collaboration Analysis". submitted at: IMWUT Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2019.

Part I

Recognizing Cognitive State

Tracking Cognitive State: Detection of Cognitive State and Well-Being in Dementia Patients with Indoor-Location Assessment

In no century our world has changed like in the last hundred years. So has health care. Achievements in electro-technology, image processing, and tele-medicine allow to detect many illnesses in their onset, and thus today many diseases can be cured, or their progress substantially slowed. Due to these exciting developments, the life expectancy of our western civilization increases steadily. People are living longer, and in many cases, people can even live a relatively healthy life after retirement. With people living longer than ever, though, age-related diseases have more and more impact on health and elderly care. The number is increasing, especially for diseases that are not necessarily due to the lifestyle, but rather to genetic aspects that progress over the years as a person ages, such as dementia or Parkinson's.

As has been pointed out before, the way our social system has developed, people live alone or in small communities (families of 2-3 people). In contrast to large families of previous decades where the grandparents helped to raise the children and in turn received care when they got sick, today's preferred lifestyle attempts to live an autonomous and self-determined life (at home) and keep this lifestyle as long as it is possible.

Especially for people showing first signs of dementia, though, this self-determined life comes with two significant issues. Dementia, in general, is mainly expressed by the patient's progressing disability to deal with everyday situations. Thus, the primary way to determine its progression is to monitor how affected people behave in their daily lives. First of all, a self-determined life for a dementia patient can only be accomplished if it is possible to monitor this person close enough to prevent accidents (that will progressively increase) from happening and to determine their state and the progression of their illness continuously. On the other hand, lacking the family-care network large families could offer, it is difficult to guarantee this essential monitoring. In turn, this means that many dementia patients have to give up their self-determined life.

The following chapter proposes a new way to achieve continuous and non-intrusive monitoring of dementia patients. The proposed method is designed to work in the patient's home or home-like environments. It should allow monitoring the progress of the disease and well-being of the patients. However, at the same time, it should not affect their lives and more importantly, not violate their privacy. Furthermore, the method only uses a rather simple real-time indoor location system that requires minimal maintenance or interaction.

The work of this chapter relies on an indoor location record of 6 residents living in a home-like ward in a retirement home. For this dataset, residents' locations within the ward were recorded daily for over a year. Despite many challenges and the minimal availability of sufficient ground-truth, it was possible to develop methods for determining the cognitive state and well-being of the residents.

The analysis of the location data provides a correct classification of cognitive state with an average accuracy of 70% to 90% and thus indicates that a stationary sensor-modality could be sufficient to monitor the well-being of dementia patients.

The author of this thesis has published this work and most contents, all pictures, most tables, and also partially text of this Chapter in the following publications. The author of this dissertation has written all the text included in this chapter herself, specifically text-passages taken from these publications. For more details about these publications please also refer to the entries in the literature list:

- Gruenerbl A. et al. (2011), Using Indoor Location to Assess the State of Dementia Patients: Results and Experience Report from a Long Term, Real World Study. In Proceedings of the 7th International Conference on Intelligent Environments, Nottingham 2011. [52]
- Gruenerbl A. et al. (2013). Uwb indoor location for monitoring dementia patients: The challenges and perception of a real-life deployment. International Journal of Ambient Computing and Intelligence (IJACI), 5(4), 45-59. [53]
- Gruenerbl A. et al. (2014) Ubiquitous Context-Aware Monitoring Systems in Psychiatric and Mental Care: Challenges and Issues of Real Life Deployments. ICCASA 2014 Conference Proceedings, (ELLCAMA-2014 Workshop), October 15-16, Dubai, UAE, ACM Digital Library. [54]

2.1 Motivation

The term ”Dementia” summarizes a set of mostly neuro-de-generative diseases, which progressively lead to a loss of cognitive abilities, and in the course of it, to the loss of the capacity to deal with basic everyday situations. Currently, different types of dementia are known. The most common type is Alzheimer’s disease, which is primary neuro-de-generative dementia. Memory deficits are indicated by cognitive handicaps (e.g., reduction of the ability to judge, the ability for scheduling or processing of information) and by changes in behavior and the affective-spectrum/emotional area (emotional instability, excitability, apathy). The cause of Alzheimer’s is known to be Plaques in the brain and/or changes in nerve cells (neuro-fibrillae, [55], and [56]). Vascular dementia, again, is primary and neuro-degenerative, but here handicaps are caused by disturbed blood flow. De-generative dementias proceed progressively, which means it is possible to observe different stadiums of worsening. In particular, some increasing behavioral disorders such as Sun Downing [57], daytime fatigue or disorientation indicate the involvement of hormonal processes and potentially affected circadian rhythms, which can be traced backed to insufficient exposure to light [56].

In 2013, Hurt et al. [58] estimated the total annual cost of dementia to the U.S. economy in the population older than 70 years of age to be $109 billion (for care purchased in the market). They further estimated a doubling of the charges by 2040 due to the aging of the population.

To judge the progress of dementia today is to monitor how a person deals with activities of daily living (ADLs). So far, this has been the motivation for a significant amount of research in the field of ADL recognition (e.g., [59, 60, 61, 62, 63]). Reliable ADL detection would, obviously, be a great advantage for many assisted living applications. However, continuous monitoring of dementia patients has been associated with great effort and has been very intrusive for patients and their privacy. Therefore, a system that could enable observations of the progression of dementia, invisibly, and unobtrusively would be desirable. This would undoubtedly be beneficial to the healthcare system and caretakers, but also for relatives and the patients themselves.

Unfortunately, most recognition systems require intricate sensor designs with limited recognition reliability. However, for useful state assessments, it is not enough to recognize activities, but it is also necessary to be able to judge how well these activities are carried out. Also, it is necessary to map such a judgment to a mental state. From a medical point of view, however, many common symptoms of dementia (e.g., disorientation, unpredictable behavior, diminished social interactions, etc.) are directly related to how a person moves and how much time they spend in certain places [57].

Therefore, this chapter proposes and evaluates an approach, developed when looking for ways to deploy a state assessment system over a long period (about a year) in a real nursing home. The idea of this approach is to map features extracted from residents’ movement patterns in areas (e.g., amount of movement, time spent in particular places, etc.) to changes in the mental state of the residents. In this context, and from the point of view of sensor-based recording, tracking of a person in rough areas with available non-obtrusive sensors can be achieved sufficiently. Examples of such systems are [64] and [65]. So far, there is little experience with the practical implementation of such systems for monitoring dementia, however. The following chapter will change this.

2.2 Related Work

Even though the research field of pervasive health is still relatively novel, a reasonable amount of work has been done already in this field. [1] A glance at the potential of pervasive computing in health-care from its early days is presented by, e.g., Lukowicz et al. [23], Teng et al. [66], Bonato et al. [67] or Orwat [24]. For pervasive computing, particularly in mental health, examples as approaches to assist older adults with cognitive impairment are presented by [68] and [69]. A distinctive part of pervasive computing in mental health is the monitoring of patients who have dementia via video footage. Megret et al. [70] describe their work of monitoring dementia patients using wearable video cameras, including a video-browsing interface so that dementia-specialists can give continuous feedback. König et al. [71] present more recent work on validating an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients. Possible techniques and set-ups needed to develop assessments for monitoring and intervention systems are presented by [72] Rebenitsch et al. where the feasibility of different approaches are analyzed, and a sample environment established in a lab is introduced.

Still, rather little work has yet been done in monitoring dementia patients using any kind of wearable sensors. The main reason for this is that problems have to expected if sensors are not hidden from the patients ([73] Alline et al.). Chen et al. [74] provide an example for deploying pervasive systems in nursing homes for analyzing dementia patients in their daily life. Their work introduces detecting of social interaction events in hallways of nursing homes by video and audio recordings.

Research with some parallels to the endeavors intro-

[1] Parts of the text of the Related Work have originate in following publications of the author of this thesis. Any text-passages taken from these papers have been written solely by the author of this thesis:
Gruenerbl A. et al. (2011), [52] and Gruenerbl A. et al. (2013), [53], please refer to the respective entries in the literature list.

duced in this chapter is described by Lin et al. [75] who propose an assessment and safety monitoring system. The goal of this system is to guard dementia patients both, indoors and outdoors (especially when trying to leave safe areas or come close to hazardous regions) by installing an RFID based alert system for caregivers and an algorithm to assess the state of dementia. However, unlike the work in this chapter, the assessment of dementia by Lin et al. [75] is performed with questionnaires, implementing standard screening procedures in a flexible XML based questioning system instead of deriving it from the sensor data.

Developments in pervasive computing have gone in the direction of smart homes and smart living space. One aspect of this research field is smart homes with protective functionalities for the elderly. Lotfi et al. [76], for example, introduce a smart home for supporting independent living by equipping homes with simple sensor-networks to monitor older people's behavior. This system includes standard sensors like motion sensors and door entry point sensors. The authors even go as far as using recurrent neural networks to predict the future values of the activities for each sensor to inform the caregiver if necessary.

2.3 Objectives and Contribution

The primary objective of this chapter is to investigate the potential of pervasive sensing technologies to monitor and assess the progress of dementia. In particular, this work should focus on the assessment of the feasibility of a suitable sensor system in a real environment. However, it should also test the usability of this sensor system when deployed over a long period. Furthermore, based on the collected data, it should be confirmed that it is possible to determine an overall state-value for people with dementia only based on this sensor data. In line with these objectives, the main contributions are:

This chapter begins with an analysis of the usability of various possible detection modalities in terms of the requirements and constraints that are relevant to the planned system and its use. In addition, all the constraint that a nursing home imposes on such a study, but also those that have emerged through observations in the home and discussions with employees, are taken into account. (Section §2.4). The system that best meets the requirements was then installed for one year on a ward in a nursing home with dementia patients. (Section §2.5).

A collection of practical problems that occurred during the study is presented and discussed below. This part has a definite meaning, as future work could benefit from these findings. (Section §2.6).
Next is a description of the extensive dataset, which represents one year of the everyday life of the nursing home. It includes a quality analysis and a discussion on how to transform standard care documentation into quantitative state assessments. These should then serve as the basis for evaluating automatic state classification. (Section §2.7).

The analysis of this sensor data demonstrates the potential of the deployed system for a reliable state assessment. The introduced method is evaluated on over 120 weeks of data, collected from 6 residents in different stages of dementia (including advanced dementia). When averaging the residents' mental state over periods of two weeks in positive or negative well-being, it is possible to achieve a recognition rate of 92%. Three of the residents even reach 100% accuracy. When introducing the possibility of a neutral state without exceedingly positive or adverse events, meaning three possible states, the recognition accuracy is 80% with one resident reaching 100%. The analysis concludes with a detailed discussion of the influence of various factors, such as feature selection, time granularity of recognition, user dependency in the amount of data used to train, and system performance. (Sections §2.10 and §2.11).

In addition to the quantitative results, some additional qualitative results are presented. Including occurrences of certain events, e.g., interactions among the residents or even aggressive behavior, which show up very clearly in the sensor data.(Section §2.9)
Finally, the analysis of the nurse's perception provides a way to understand the factors that are recommended to being considered in practice for a successful use of pervasive technology. (Section §2.12).

2.4 Choosing an Appropriate and Fitting Sensor System

Deploying a study in a real-life environment means intruding into people's lives. Therefore, a real-life study has to be carried out with special care and sensitiveness, specifically in dealing with mentally affected people. Following regulations, the study had to be approved by the relevant hospital bodies, agreed to by all the involved nurses and by the residents and, because of dementia issues, their relatives or legal representatives. All these groups of people involved (actively or passively) stated their constraints or conditions. The fundamental prerequisite was that the burden of performing additional tasks was not to be put on residents and nursing staff. Key concerns were:

- Asking the residents to wear simple sensors was possible, but there was no guarantee that they would always do so in a pre-defined way; thus it had to be expected that residents would spontaneously refuse to wear the sensors. Also, the staff immediately stated that they would not always be able to check whether the sensors were worn.

- Any infrastructure to be installed in the home had not to disrupt the day to day operations (the staff worried that ”out of the ordinary” looking sensors might have a negative influence on the residents).
- The nurses were not expected to put any effort into this study, except to make sure that the residents put on the sensor in the morning. E.g., battery changes or additional documentation was not possible.
- The study staff was allowed to conduct observations, but with the condition that the residents were not or did not feel disturbed, and the procedures at the ward should not be interrupted. Video recordings were not allowed at all.
- Also, the staff had concerns about privacy and that the sensor records would allow judging their work. As a result, the study refrained from equipping nurses with sensors. This, in turn, had a negative impact on the study, as interaction with the nurses could have been a major factor in assessing the residents.
- Finally, all equipment or other devices used in the nursing home had to be registered products with corresponding official certifications.

2.4.1 Preconditions and Considerations

Following the discussions and constraints imposed by the situation and the home's authorities the question of “what kind of technology was fitting for the study, and was acceptable for the care professionals?” had to be met by a number of preconditions:

- Any technology had not to influence or disturb the care concept. In fact, any technology had to be invisible or at least be perceivable.
- Any technology had to be able to assess the quality of life of the residents. Quality of life in this context was defined as social interactions as well as physical and, indirectly, social activity.
- Concerns (subjective or not) about technology-based health issues (e.g., “radiation of Bluetooth” etc.) needed to be considered and handled sensitively.
- Any technology had to require no or only minimal maintenance over several weeks or months.
- Any technology had not to increase the workload of the health-care professionals. This included any interaction with the system going beyond checking if residents were wearing the sensor equipment.

All considerations and preconditions could be summarized in the following system requirements:

1. The system had to be able to quantitatively capture parameters that are suitable to evaluate the quality of life of people living in a nursing home ward, entirely or at least partially.
2. Such a system had furthermore to be able to cover most parts of the ward (living space of the residents) while being able to run over a long period and also had to be able to run autonomously. Essentially, this means that the system of choice had to be either able to run with not attendance over months or had to start automatically every day.
3. Essentially, the system had not to impede with the daily life at the ward, hence had to be integrated into the wards' structure, and yet had to be removable completely after the end of the study.

2.4.2 Available Technologies

When the study was deployed (in the late 2000s) only a minimal number of potentially possible location technologies were available. Please note that the following list describes the situation in the late 2000s and not the situation of today (2019). The following technologies were evaluated [1]:

- **RFID:** RFID is a well-known system combining small wearable transponders and static readers. Depending on the modality, the range of a tag is between a few centimeters and a few meters. Using RFID technology as a location tracking system would require a large number of readers. Even though RFID technology already was used for coarse location tracking within buildings (like [77] an example relevant in 2008), by the time this study was being conducted, this kind of technology was not suitable for high-resolution location tracking within rooms. Note, after this study finished other research with RFID location tracking were performed in the context of smart-home for the elderly, see [78] as an example.
- **Ultra-Sound:** Available ultrasound systems consist of movable transponders and static receivers. The receiver recognizes if a transponder is in its vicinity. Thus, tracking of objects or movements works nicely with ultrasound in small areas [79]. Position-calculations within rooms, though, are only possible by using a sufficient number of decently placed receivers. Due to the constraint of ”technology should be invisible and not disturbing nurses' work,” the number of required receivers ruled this option out.
- **Camera monitoring:** Monitoring with cameras has a big potential ([80]). Nevertheless, camera monitoring was ruled out instantly by the given constraints. Any kind of video monitoring would violate the privacy of the residents, the nurses, and visitors. Therefore, camera monitoring was not allowed.
- **UWB Localization:** By the time the study was deployed, UWB (ultra-wide-band) radio-location was very innovative and newly released. It works with small wearable transponders (tags) and a couple of

[1] The description of available technologies origins in the following paper. All texts taken from this paper have been written by the author of this thesis: * Grünerbl A. et al. (2013). [53] please refer to corresponding entries in the literature list or beginning of this chapter

static receiver sensors and calculates the location based on the principle of time-of-flight in real-time. Depending on the complexity of a room, only a few receivers can be sufficient to cover large areas ([65]). In contrast to RFID and ultra-sound, UWB performs actual 3D real-time in-door location.

2.4.3 System Decision

The above-listed systems are a very limited variety of possible methods. Due to the imposed constraints and considerations most of them had to be ruled out. Considering the further requirements and precondition, after some initial tests, the UBISENSE UWB Indoor Location System turned out to be the only satisfactory solution, because it is designed to run on a long-term basis with only little maintenance necessary.

The Ubisense UWB System is a real-time location system ([65]) that comprises an array of a few (typically 3 - 10) ultra-wide-band (UWB) radio receivers. These calculate the location of small UWB-signal transmitters (tags). The proposed accuracy of the system (15-30cm) allows determining interaction patterns between people, with only relying on small (about 8x4x1cm) tags, which can be carried anywhere on the body.
The sole maintenance that is required by the system is to check batteries of the tags and frequently supervise whether systems were running, which can be done remotely. Besides, to guarantee the functionality of the system itself, only tags are needed to be worn.

Checking whether the residents would be wearing them was considered to be a doable task for the nurses. It should be possible to integrate this into the morning routines as the nurses were supposed to help the patients with morning hygiene and dressing in any way. This task was no additional burden on the nurses.

2.5 Deploying a Long Term Data Collection Study in Elderly Care

The data used in this chapter originates from a system deployment ([81]) as part of the "Lichtprojekt" (www.k-licht.at) in 2008. This project's goal was to enhance the quality of life of older adults with dementia through an elaborate indoor lighting system. The critical issue in the project was to evaluate the influence of different lighting setting on the residents in various stages of dementia. The goals of this project stated that only a long-term real-life installation of the enhanced lighting would allow getting the required results. Thus, a home for the elderly, the "Home St. Katharina" in Vienna was part of the project consortium.

2.5.1 Study Background and Set-Up:

During a complete restoration of the St. Katharina home for elderly, one ward was remodeled into a residential ward for dementia patients according to a new care concept called Maieutics. [82].

Facility and Care Concept: Next to eight single and two double rooms, the remodeled ward itself consists of a big living area including a kitchen, an area for eating, and a living room. (Figures §2.1 (left) and §2.7). The re-design of this ward aimed to create a family-like environment needed for the Maieutics nursing concept. ([82], [83], [84]). [1] The rooms and the living area were furnished in a style fitting the resident's customs.
According to the care concept, the residents had to receive as much freedom as their state allowed to define their daily routines. Equally, the residents were encouraged to socialize and engage in a broad range of activities. In the morning, they would get up at the time they preferred and either dress themselves or get help with putting on clothes. Afterward, they would go to the kitchen area, where they would get breakfast. Some residents preferred to take the meals in their rooms. Between breakfast and lunch, the nursing staff would clean the rooms and provides drinks. At around eleven o'clock, lunch would be served in the kitchen/living area. Afterward, most left for an afternoon nap. During the afternoon, the nursing staff again would provide coffee and snacks. From time to time, also depending on the staff present, nurses would offer activities like playing ball, painting, or doing handicrafts. Around five o'clock, dinner was served. After dinner, most residents retreated to their rooms. Others stayed watching TV together.

Residents - Study Participants: Overall, 13 different residents lived in the ward during the study period (11 female, two male, the average age was 88.6 years at project start, and 87.4 years at projects end). At the beginning of the study, every patient received an MMS (Mini-Mental Status) score. According to this score, all residents were classified as dementia patients. The degree of dementia, however, varied from mild to very advanced with severely constrained cognitive abilities. The mini-mental status examination is a frequently used cognitive screening measure to identify dementia. Folstein and McHugh introduced it first in 1975 [85].
Common additional illnesses were Parkinson's disease, Diabetes Mellitus, and high blood pressure. Psychotropic drugs were prescribed for ten residents and soporifics for six (see Appendix for a detailed description of the relevant subjects).

Nursing Staff: During the study, a total of 10 nurses worked at the ward with some personal fluctuations at the beginning of the study. In the second half, staff composition remained stable. Based on the Maieutic care

[1] Note that the Maieutics care concept is a regionally implemented method in Germany/Austria and the Netherlands and therefore, only references are available in the German language.

concept, the nursing personnel was supposed to support the resident's abilities by playing games, cooking easy meals, doing simple housework, and more. Due to disagreements among the nursing personnel, Maieutic activities were not promoted very well. However, generally during the morning shift, 4-5 nurses were present and helped the residents getting ready for the day.

2.5.2 System Installation

As a matter of respecting the privacy of the residents, it was decided not to install sensors in the bedrooms but only cover the common areas with the Ubisense System. Anyway, most interactions and activities relevant to state-assessment will take place in the common areas (communication, movement, etc.). With sensors covering parts of the hallway, it should be possible to infer when the subjects were in their bedrooms.

For the Ubisense System, the number of sensors needed to cover an area depends on the size and structure of this area, but also on how many larger structural and/or metal items are placed within the area to cover. In an empty rectangular space, four sensors can be enough to cover several hundred sqm. If the area includes reinforced walls or for example have a kitchen block somewhere within the area, probably significantly more sensors will be needed. For the study presented in this thesis, six sensors mounted on the walls or in the corners of the living space, as shown in Figure §2.1(left) did suffice to cover the common area and parts of the hallway (see Figure §2.7). After some pre-tests and several discussions with the nurses and home authorities, it was decided to attach the tags in the form of a necklace around the patient's necks (see Figure §2.1 (right).

2.6 Challenges and Issues of Deploying Sensors in a Home of the Elderly

Numerous groups work on integrating technical systems into everyday life. Most groups like [86] prefer to work in laboratory-like settings where different parameters can be influenced. Certainly, this has a reason. Deploying technology in real-life is difficult, as, in the real world, various issues can complicate the work. In real life, studies have to be executed without or with as little interference with the regular work/life of the study subjects as possible.
Furthermore, in real life, every study has to deal with factors, which can influence the success of the study. Often those factors are not expected or even possible to consider beforehand. However, by careful preparation and consideration of potential challenges, real-life deployments indeed can be successful. In the following section, issues and challenges that had to be dealt with are discussed. Furthermore, it is shown how it was possible to overcome these issues. Even though some of them might seem to be obvious from an outside perspective, in the planning phase of the study, however, the most obvious aspects are easily forgotten. [1]

2.6.1 Technical Issues

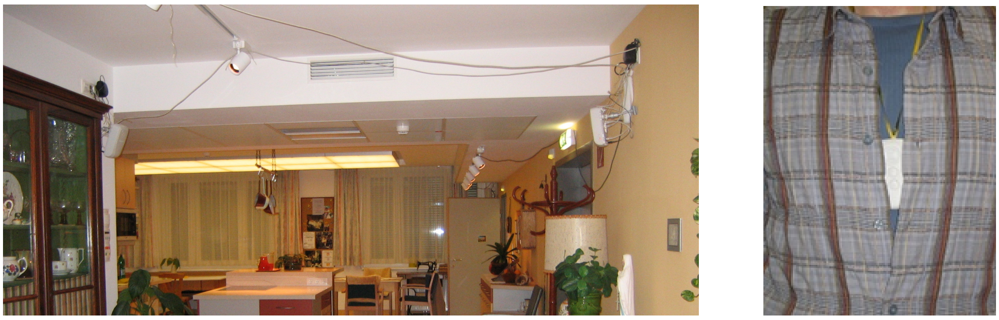
Many technical issues can be eliminated beforehand by carefully testing everything. Still, not all potential technical problems can be considered in advance. In the nursing home in Vienna, specifically in the beginning phase of the deployment, four technical issues arose:

Using newly released Technology: At the time the study started, the first certified version of the Ubisense System was only newly released. Regarding the project, it was required to use this newly certified version, which, contrary to the previously un-certified releases required additional timing-wires. As the necessity of these additional wires was not anticipated (specifics were not available, and for previous versions, timing wires were not essential) those extra wires were not included in the structural changes of the ward. Therefore, the cables had to be laid provisionally after installing the sensors, see Figure §2.1. In fact, such "visible and potentially disturbing installations" were opposing to the requirement of ”sensor-systems not to disturb visibly, but fortunately was accepted by the staff.

Missing Network Connection: Long-term operating systems, especially new edge technology like the Ubisense System, have to be supervised and maintained frequently. Since it was not possible to visit the ward daily, a remote supervision system was necessary. Even though the nursing home authorities had promised Internet connection, due to some ongoing house-internal shifting of powers, no Internet connection was available. To solve this issue without losing time (while waiting for the internal disagreements to be settled), a UMTS-router (with static IP-address including a DynDNS service), was installed and all sensor systems were connected to this router.

Instabilities in power network: During the operational tests in the first weeks of the deployment, instabilities in the power network of the nursing home were revealed. Those instabilities were due to frequently power-ups/downs of high power consuming devices in the nearby hospital sharing the power network with the nursing home. These instabilities caused significant crashes for the sensor systems, and it took some weeks at the beginning of the study to find the cause of these crashes. Eventually, the instabilities in the power network were only discovered by chance as there was no way for us to gather knowledge about the power-

[1] Vast areas of the description of the particular challenges and issues faced during the study have been taken from the following papers. All texts taken from these papers have been written by the author of this thesis:
* Grünerbl A. et al. (2013). [53] and Grünerbl A. et al. (2014) [54], please refer to respective entries in the literature list.



(a) Sensor installment in the living room. Provisional additional wires between the Ubisense Sensor Beacon were needed.

(b) Resident wearing a tag

Figure 2.1: Installed UBISENSE Location System

infrastructure to which the nursing home is connected. In fact, we would have never anticipated power fluctuations in the network of such extent. Once the problem was determined, the issue was solved by installing a UPS (uninterruptible power supply) for all critical systems. In addition to the UPS, a timer switch was set to frequently shut down and restart critical parts of the system in the very early morning (at a time it could be assumed that all residents were sleeping in their bedrooms). This restart was done to guarantee a working system as soon as the first residents showed up every day.

When Sensors do not work as expected Initially, the study included the measurement of sleeping quality [81]. Following the idea that restless sleep would be expressed by much movement during sleep, the plan was to use inertial sensors (accelerometer and gyroscopes) and attach them to the slatted frame of the resident's beds. These sensors were meant to detect any movement in bed, and the tests at the researcher's homes worked well. In the home for the elderly, though, some of the resident's beds were of particular orthopedic design, partially with solid metal panels instead of a slatted frame, which suppressed the forwarding of any movement to the sensor. Other patients received a rather high dosage of sleeping medication and were, therefore, hardly moving at all. For these reasons, the sleep monitoring with simple inertial sensors failed and was discarded from the study.

2.6.2 Handling Mentally Affected People

It is a fact, that technical issues, as just described, have a considerable influence on the success of a study. Nevertheless, dealing with human beings is a yet even more uncontrollable factor. It becomes even more critical when a study has to rely on mentally affected study participants. Generally, the success of the study presented here depended on residents - suffering from dementia - accepting and wearing a sensor-tag every day for about a year. To guarantee this, it took some effort on the nurses. Specifically, convincing the study participants day by day to put the tags on was an effort for them. Regardless, it was entirely clear that it was not allowed to force residents to participate. The willingness of these mentally affected people to cooperate in the study had a direct influence on the amount of data that could be collected.

Res.	1010	1032	1031	1041	1060	1090	Per.	Avg.
Nov.	0	12	13	6	0	8	4	**9.8**
Dec.	0	7	7	0	0	0	2	**7**
Jan.	15	16	10	12	7	14	5	**12.3**
Feb.	21	19	20	20	18	1	6	**16.5**
Mar.	23	26	30	18	19	12	6	**21.3**
Apr.	18	20	18	1	0	0	4	**14.3**
May	19	9	19	0	0	9	4	**14**
June	8	0	0	0	0	9	2	**8.5**
July	20	24	11	7	18	14	6	**15.7**
Aug.	15	14	0	13	10	5	5	**11.4**
Sept.	7	12	0	0	0	0	2	**9.5**
Mon.	**9**	**10**	**8**	**6**	**5**	**5**	6	7.17
Days	149	159	128	77	72	56	6	106.8

Table 2.1: # of days with available data per resident per month and for how many persons data is available. Some months show more data gaps than others.

Eventually, some residents refused the tag. In total, only six out of 13 residents would wear tags often enough to provide a sufficient amount of sensor data. In the beginning, all 13 residents started to wear them, but some quickly refused to continue. Moreover, three of the residents died during the study year, and another resident got hurt in an early phase of the study and was bedridden afterward. An interesting side effect (of the light-projects) was that the different lighting-phases seemed to influence the willingness of residents to wear tags. In some light-phases, most residents wore their tags, and in others, most did not.
Further reasons for the decimation of data was that some residents partially refused the tags for some weeks, very likely due to their mental states, and later started to wear the tags again. Thus, in total, for six residents out of 13 a sufficient amount of data was collected. Within these six residents, the amount of available data ranges vastly between five and ten months.

Table §2.1 shows the amount of available data-sets per resident per month. It also provides an overview of which phases tags mostly were worn. For example, during July only one person provided data, while throughout January nine residents wore the tag. The absolute number of data-points per patients lies between 6 and 159. It was decided to use only data of patients with a minimum of 50 days of data or more and a distribution over at least 4-5 months (which should include several changes in state). Thus, in summary, data of only six residents was used (marked bold in Table §2.1).

2.6.3 Limitations of the Ground Truth

One disadvantage of the study, partially a direct effect of the nurse's struggle, was the fact that it was not possible to repeat the MMS test regularly. Moreover, at the time of the study, the keeping of health records in homes for the elderly was not regulated clearly, which means that no systematic health record was kept. Contrary to standardized health record of hospitals, the "resident's health documentation" contained reports of events and observations that the nurses considered noteworthy.

These records and observations were put down frequently, but not daily, and were in style and tone unique to the respective nurse in charge of the day. This way of record-keeping was a typical documentation style in Austrian nursing homes back then (refer to [87]). Furthermore, the written observations and reports of events were not expressed in a numerical score or some standardized, fixed scheme, and, as already mentioned, differed significantly in quality and quantity of information.

Moreover, detailed human observations on a larger scale (throughout the entire study) were not possible or doable. Note, using of cameras and video recording was not allowed. Thus, the health records were essential for this study's requirement to get ground-truth. Nevertheless, the way they were kept proved to be an additional and specific challenge in this study. The health records of the nursing home, no matter how detailed or not detailed they were, had to suffice as the only possible source for ground-truth!

2.7 Data Collection Study and Data Quality Assessment

From November 2007 until September 2008, the Ubisense System was running 24/7 and gathering data from those tags being worn. The Ubisense system, though accurate up to 15 cm (in 3 dimensions) in a perfect environment, can easily be disturbed by corners, niches and larger metal constructs in the furniture.

To assess the influence of the home environment onto the sensor system, and the reliability of the data gathered, an observation lasting a few days was performed in the ward. Since video-recordings were not allowed, a human observer spent several days at the ward documenting everything that happened, including every single location change of each resident.

2.7.1 Qualitative Data Quality

Figure §2.2 (left and right) shows the data gathered by the sensor system in comparison to the observation. Specifically in Figure §2.2 (left) both agree very well. Irregularities were mainly caused by the impossibility to fully synchronize the computer with the human observer, as the human observer was not able to document activities of more subjects at once to the split second, and in the evening a tag was laid aside by its resident, which went unnoticed.

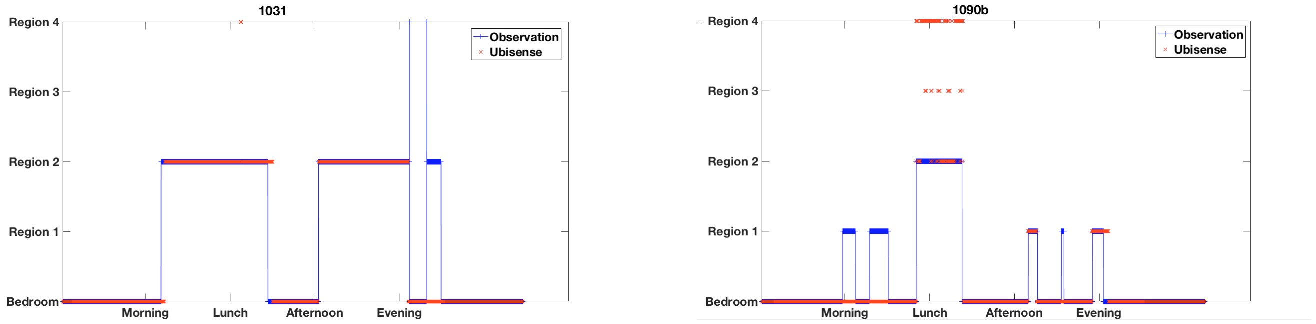
In Figure §2.2 (right) coarser errors appear to be visible. However, these faults can also be explained and can be assigned to known error sources. Therefore, these inaccuracies are not indicative of malfunctions of the localization system. As an example: resident 1090b (data shown in Figure §2.2 (right)) arrives in the living room in the morning but does not carry the tag with them (which could not be observed by the human observer as the tags were worn mostly beneath clothing). This situation explains the deviations between observation (blue) and measurement (red). Resident 1090b picks up the tag around noon and sits in region 2 (at the small table) just at the border to regions 3 and 4 (see also figure §2.7 for a better understanding). Movements in the chair, and sliding around with the chair (a habit of this particular person, that was heightened during some phases), caused the system to partially allocate the resident to regions 3 and 4 (which, regarding hard borders would be correct). The observer, on the other hand, also correctly, did not determine any change in location. It must be added that the sensor system naturally loses its accuracy due to metal components in furniture, which in this case could have increased the described effect.

A further interesting example of data irregularities was given by one resident who had to use a walking aid and was moving very little but, during the first days, seemed to be among the most active subjects in the data-set. It turned out that the nurses were attaching the tag to the metal walking aid, which caused the system to provide highly unstable data.

After the tag was placed on the body of the resident, the problem disappeared. Additionally, effects as the bouncing between different areas were filtered out as far as possible for further evaluation in order not to allow such effects to be included in the calculation of motion. The observations themselves also revealed (not surprisingly) some limitations of a human observer in comparison to what a sensor system can do. The following aspects are comprehensive sources of errors in the comparison of human observation and sensor readings:

1. For a human person, it is impossible to monitor 13 people every second as precisely as a sensor system. As a result, observation will always deviate, and sensor measurements only can be synchronized with the human observations to a limited extent. E.g., the human observer was focusing on documenting movements of persons A and B in one corner and thus, missing person C moving in the opposite corner. The location change of C was documented as soon as noticed.

2. For the human observer, it was only possible to esti-



(a) Occurrence of tag laid aside unnoticed in the evening.

(b) Jumping of sensor values between 3 regions.

Figure 2.2: Data gathered by the sensor system (red) in comparison to the human observation (blue) over one day for two residents. Both sensor and observation match for most parts of the day, but also show explainable differences.

mate the coarse ranges and borders of the areas and locations, while the sensor system obviously can calculate them precisely. Thus, inaccuracies and errors in the allocation of locations for persons, particularly at the junctions between areas, are very likely.

3. Since some residents wore their tags beneath their clothes, it was not always possible to ascertain whether the residents did carry their tags or not or had put them aside for a short time. Therefore, sometimes laid aside tags would go unnoticed and thus could also lead to differences between the localization data and the observer's recordings.

2.7.2 Quantitative Quality Assessment

Table §2.2 shows the quantitative results of the human observations. During an average time of 29.3 hours per patient, the sensor data correlates with the observation by 95.7% on average. Most error rates over 5% were explainable by the error sources listed above. Differences in the amount of time every single person could be observed were resulting from patients' habits to lay the tag aside or to forget to wear the tag which then had to be handed to the resident later. During the observation, a total of 69 events (= changing location between regions) were documented. Of those, 55 events could also be extracted from the sensor data.

DS	1060	1032	1090	1010	1031	1041	mean
match %	87.4	98.9	94.8	97.8	96.7	98.6	**95.7**
error %	12.6	1.1	5.2	2.2	3.4	1.4	**4.3**
time in h	48	30.7	34.3	24	24	14.7	**29.3**
lost	-	3x	1x	-	-	-	4x
handed	-	1x	3x	-	-	-	4x

Table 2.2: Data Quality: Sensor-data and human observation match by 95%.

2.8 Handling Ground Truth with Limited Availability

In the year the data collection study took place, the care laws for the elderly and regulations were about to change. Thus some provisions valid for hospitals were not yet implemented in homes for the elderly. One of these regulations was the health documentation. On the contrary to hospital patient records, the standard documentation practice for residents of nursing homes (as described e.g., in [87] which was the recognized nursing handbook in Austria) was in the form of qualitative analysis through observations and documented events, done frequently but not necessarily daily.

Furthermore, the entries were not expressed in numerical or fixed schemes, but more in the way the nurse in charge presumed to fit. Thus those records included information about visual appearance and (seldom) screening tests. These entries were the primary basis for determining ground-truth about state and state changes (e.g., to determine a patient's medication needs). The kind of documentation in nursing homes back then was presumed to be sufficient for experienced doctors and nurses who would be able to look at a set of entries and put them in relation to what they knew about the patient. This way, they would be able to come up with a subtly, differentiated metal state assessment.

For the meanings of this study we had neither the possibility of having such an analysis done and particularly not for each resident, and throughout the entire year of data recording. Nor were we actually interested in such a detailed and subtle diagnosis. This apparently would have been beyond what could be expected to be derived from a simple location tracking system. Instead, we were interested in broader classes of well-being, whether the resident's state was either positive or negative. Therefore, the challenge and at the same an important goal was to find a way to relate such classes to what was written in the records. To do so, again, the only available sources for a coarse ground-truth were these loosely kept resident's documentation records.

A close study and thorough review of these records revealed that most entries could be classified into four different categories:

- Positive or negative entries about the resident's physical state. These entries included phrasings like ”resident was tired all day” or ”resident was restless” or ”resident was quite fit today.”

- Positive or negative entries about the resident's mental state, which included statements like ”resident seemed to be happy” or ”resident was really aggressive and tried to hit others.”
- And either positive or negative entries about social abilities, like ”resident was unwilling to attend the joint games” or ”enjoyed talking to others.”
- Any signs of other illnesses and issues.

2.8.1 Ground-Truth Metric

All these entries were extracted from the records, based on the Behavioral and Psychological Symptoms of Dementia Educational Pack and its description of the assessment of quality-of-life [88]. Further, discussions with nursing staff and nursing faculty professionals helped to enclose it to the best fitting set of entries.

All entries of the health records were extracted one by one and assigned to one of the four categories. About 80% of the extracted entries could be classified this way. Not all of them were explicitly positive or negative, in any case. A statement like ”the resident was funny” (in context of the language used) could either mean that the resident made jokes - an apparently positive entry, or that the resident acted strangely - a somewhat negative entry. Months or weeks after the entry was made, it was impossible to determine which meaning was intended. Thus, unclear entries were not considered for ground-truth to avoid biasing by subjective interpretation.

2.8.2 Ground-Truth Resolution

After extracting all relevant entries from the records and classifying them accordingly, the next step was to quantify the resident's well-being over the study year. Unfortunately, as has been described, the health records would not standardly include entries for every single day. A closer look revealed that not every entry could actually be assigned to a specific day, and sometimes entries were delectably made on a wrong day. For example, one entry about a person said on one day: ”Resident fell and hurt their hip,” but the fall-report (by the ambulance) had the date of the day before. As such misalignments of days and entries could be detected, this also meant that other such misalignments likely remained undetected. Hence day-specific labeling of ground-truth would have been problematic. Furthermore, all entries were rather patient-specific, and thus, the extraction of the resident's state parameters had to be done uniquely for each resident. In order to find suitable ways to extract sufficient ground-truth, nursing professionals were consulted.

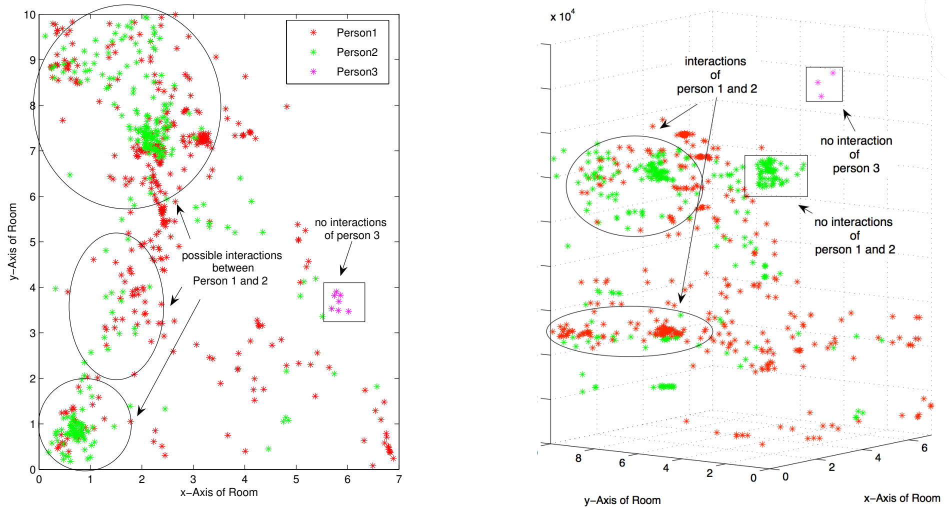
After various discussions and different ideas, we settled on grouping the total number of positive and negative entries for the areas of physicality, mentality, and sociability by a period of two weeks. This way of grouping would lead to a negative value for this particular two-week period if negative entries did come in higher number or to a positive value respectively. Eventually, to avoid dealing with negative numbers, the per 14-day entry sum was normalized (between the maximum number and the minimum number of entry sums). The resulting value for each fortnight, a value between 0 and 1, therefore was ranging between the patients worst and best state, respectively. This approach was acceptable inasmuch, as both physical behavior and health record entries are patient specific and not combine-able in-between patients. Thus, a certain number of entries or a particular value could have an entirely different meaning for different patients, always according to a patient's character. Therefore, the state-index is patient specific and had to be calculated for each patient uniquely. Regarding the classification (see in the following sections), for each day (or each week depending on the granularity of the recognition system) a label was assigned, derived from the corresponding fortnight-period.

Regarding the resident's state, two different modalities for rating the states were pursued. To recap, the state values range between the worst and best state of each resident. Therefore, the first modality to rate the state values were in two classes: positive state (0.5-1) and negative state (0-0.49). As this modality ignores the possibility of a normal/neutral state of well-being, the second modality should include this normal state. Hence, this second modality opted for using following three classes: negative state (0 - 0.33), neutral state (0.33 - 0.65) and positive state (0.66 - 1).

2.9 Qualitative Analysis of Cognitive State

Before starting a systematic quantitative analysis, first, a qualitative evaluation should reveal whether the recorded sensor data would include information that could be mapped onto disease relevant observations from the health records. Social interactions are an essential aspect of assessing the state of dementia patients. The term “social interaction” has a broad meaning, because even looking at each other could mean to interact on a social level. In order to limit what is possible to infer with location data, social interaction was defined as “staying in the vicinity” of other persons. This is reasonable, as a person that stays in the vicinity of other people broadcasts the desire to interact on a social level, even be it in the form of enjoying other peoples company.

Figures §2.3 (next page) show the raw data-points of three residents for one day. On the left picture, the data is displayed in 2D. However, even here, it is clear that resident 3 does not co-locate with any other person and is only present for a brief period. On the other hand, residents 1 and 2 spread over the entire area and are often co-located. Looking at the 3D plot that has time added as the third dimension, this impression of co-location of residents one and two are confirmed mostly. Residents 1 and 2 are often co-located both location and time wise,



(a) co-location in 2D

(b) co-location with additional 3rd axis time

Figure 2.3: Social interaction: co-location with possible interaction, showing residents staying in the same area at the same time

but at one point resident 2 stays alone. Resident 3 is confirmed to be alone the entire time.

Even though the health records would not provide sufficient information about detailed social interaction and only coarse information about sociableness, there were particular aspects stated in the health record that seemed worth to evaluate them in a qualitative manner. For these qualitative evaluations, the location data was used to calculate the co-location of persons, meaning different residents staying in the same area for a certain amount of time. Following social aspects, retrieved from the health records could be analyzed qualitatively:

Periods of Aggressive behavior The health records indicated that resident 1041, who was profoundly demented, acted in a slightly aggressive manner during some periods, but did not during others. This can be seen in the amount of time that resident was co-located with others. Figure §2.4 (left) shows the periods of non-aggressiveness where resident 1041 spends time in the vicinity of others, while during aggressive periods no other resident is co-located (right picture).

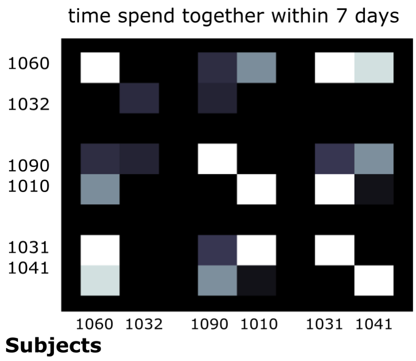


Figure 2.5: Personal relationship - co-location with possible interaction residents 1010 and 1031, 1031 and 1060 who reportedly liked each other's company.

Personal Relationships The health record of resident 1010 indicates in several entries that resident 1010 enjoys the company of resident 1031. This observation can also be detected in the sensor data. Figure §2.5 left, summarizes the time the different residents stay in each other's proximity over one week. While 1031 spends time together with different other residents, 1010 mainly spends time along with 1031. Still, the color indicated that they did spend a lot of time together.

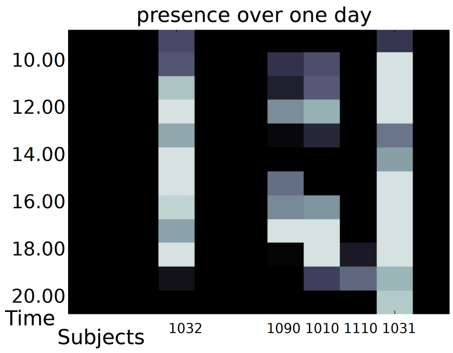
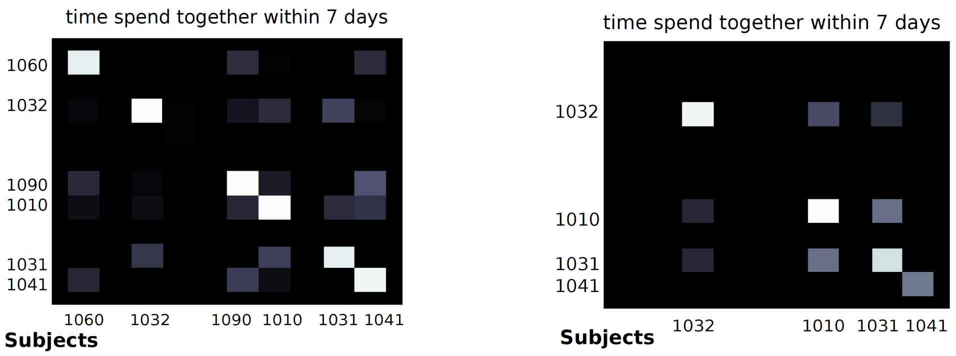


Figure 2.6: Daily habits - presence in the social areas reflects known habits. 1031 and 1032 stay the social area, 1090 and 1110 leave for a midday nap.

-

Daily Habits The Maeiutics care concept of the ward encouraged the residents to design their day according to their preferences and the way they were used to spend the day. Thus the daily habits of the different residents and their presence in the social area differ in-between residents. While some of the residents, especially 1031 and 1032, did spend the better part of a day in the social area, other residents, like 1010 or 1090, commonly were present in the social areas only when taking the meals. Figure §2.6 provides the distribution of the daily presence of all residents accumulated over one week per person. Light colors in the figure indicate higher values than darker colors.

The residents 1031 and 1032 show light color in this plot during almost the entire day, just with some darker



(a) co-location in normal periods

(b) not co-location in aggressive periods

Figure 2.4: Social interaction (co-location) and behavior (light colors mean higher values than dark colors).

color during lunch break (probably with an occasional midday nap). Other residents like 1090 and 1010 mainly have a lighter color in the plot during breakfast, lunch, and the afternoon coffee/early dinner time but return to their rooms in-between.

Resident 1110 (got ill soon after the study started and was excluded from the study), initially took only the dinner with the other residents in the ward's social area. All other meals were consumed by 1110 in the nursing home's Mensa. This agrees with Figure §2.6 where resident 1110 is only visible in the evening.

2.10 Quantitatively Measuring the State of Dementia Patients

The primary objective of this endeavor was to recognize the state of well-being of the residents living in the ward, only by analyzing their location data. In this regard, it should be evaluated if and to what extent the states extracted from the health records could also be acquired by performing classification of these location data-traces.

As is done in most standard classification processes, first adequate features had to be determined and extracted. These features then had to be evaluated to make sure they sufficiently incorporated all essential and relevant information. In the following step, parts of the features were then used to train and test a machine learning classifier.

2.10.1 Feature Computation

Some different features and parameters were first evaluated. These tested features included frequency of movement and time of stay in the overall common area. Eventually, however, it proved to work best, not to see the monitored area as a whole but divide it into different meaningful sections and analyze movement within and in-between these areas.

Data Pre-Processing: In this regard, the first preprocessing step for extracting the adequate features was to divide the monitored area into a total of ten meaningful sections. These were (see also figure §2.7):

- big table
- small table
- kitchen
- walk-ways
- entrance
- table
- couch
- balcony door
- setee

The raw data from the Ubisense system is presented in the form of x, y, and z coordinates. Thus, these coordinates were mapped onto the areas, meaning a feature space reduction of the raw data, resulting in areas assigned to the patient in one-second intervals.

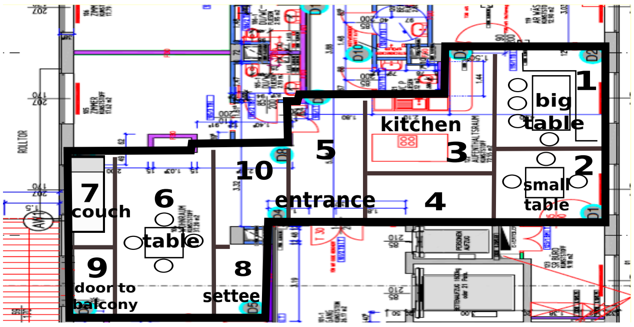


Figure 2.7: Area covered by sensors and its division into 10 semantic regions

As described in the data quality assessment §2.7.1, fluctuations of the sensor signals would occur if patients were staying at a border between areas or when staying in the vicinity of disturbances. Thus the sensor-data was smoothed further. This smoothing was done by basically ignoring fluctuations (changes between areas) if they were smaller than 10 seconds. By considering the somewhat impaired physical conditions of the residents, in combination with the size of the area itself, this strategy of smoothing the data was not found to distort the data, as it could be assumed that the residents were not able to actively move between areas within less than 10 seconds. An additional aspect of using semantic regions, besides the reduction of the amount of data, was to measure general movement.

Regarding looking at a location as semantic regions, changing between regions can be considered as a movement (e.g., switch from table to kitchen). As the regions were defined according to a clear semantic and practical meaning, staying in one region should not be considered as a movement. For example, if a person was staying in the area “kitchen” for washing dishes, it would be very likely that this person would change their position within the kitchen area. Meaning, the person would put the spoons into the cutlery’s drawer, and afterward step back to the sink, which in itself is not being considered as movement!

Calculation State Parameters: Once the data was preprocessed, relevant parameters were calculated. These parameters were determined to be: [1]:

- **Total daily duration of stay in the social areas:** Staying in the social areas generally indicates the willingness of the resident to be in the vicinity of others and to get in contact with them. Therefore, the daily duration of stay in the social area, and specifically its progress/changes over the weeks is a valuable parameter in regards to (mental) well-being. This parameter itself was calculated as the sum of seconds the tag, which was carried by the respective resident, was visible by the Ubisense sensors.
- **Total daily number of changes between regions:** A further indication of (physical) well-being is the mobility or the amount of movement of a person. Regarding the location system, mobility could be measured by the number of changes between the semantic regions. This parameter was determined by counting the times a subject crossed the border between two semantic regions! For example, if the respective resident walked from the ”big table” to the ”kitchen” and back two border crossings would be counted.
- **Index for movement:** “Duration of stay” and “changes between regions” in combination provide more in-depth and more detailed information about the value of both parameters. E.g., three residents have a total number of changes between regions of 10. However, while person 1 had ten changes within 5 hours, person 2 had ten changes within one hour. Person 3 stayed in the social area all day long. This means that person 3 was rather calm, while person 2 was somewhat unsettled. Therefore, the movement-index should give an insight into the ratio between stay and changes. It was calculated by dividing the number of changes between areas by the total sum of seconds the resident stayed in the social area.
- **Distribution over areas:** This parameter provides the information in which areas the subjects remained during the day. It was calculated by counting how often a respective resident entered each of the ten regions during each day. This parameter implicitly includes information about the number of changes. Nevertheless, it does neither include information about the amount of time spent in each specific area or the amount of time spent in the social area in total.

2.10.2 State Recognition Methodology

With mapping the raw sensor data onto semantic regions, the amount of data was already reduced from three dimensions to one. By extracting different relevant features and thus increasing the feature space, the amount of data for the classification process to handle was again raised. A commonly used method to, once again, reducing the feature space is the linear discriminant analysis (LDA [89]). This method was performed in the feature space, and the data-set was transformed accordingly. Subsequently, different standard classification algorithms were investigated. All steps of the classification processes were performed with WEKA [90]. The tested classifiers were the following:

1. a Bayesian classifier,
2. a k-nearest neighbors classifier (kNN,k = 3),
3. a decision tree classifier (J.48)
4. a conjunctive rule learner

The evaluation of these possible classifiers showed that both the KNN and the Bayesian classifier perform equally very well with more than 90% accuracy. The Bayesian classifier was slightly 2% worse than the KNN on average but still performed very well. All other tested classifiers, the decision tree, and the conjunctive ruler learner produced far worse results. Details about these results can be found in Table §2.3.
Since the KNN-classifier performed best, the KNN-classifier was used for all further evaluations. The classification itself was done in a ten times cross-validation with a 66/33 percentage split strategy, meaning 66% of the data-sets were used to train the classifier and 33% were then used to test.

Resid.	**knn**	**bayes**	**j48 tree**	**conJR**
1060	100	100	73.33	90
1032	100	100	91.67	100
1090	100	86.67	33.33	40
1010	88	86	90	78
1031	80	84	84	64
1041	92.5	92.5	70	67.5
Mean	**93.42**	**91.53**	**73.72**	**73.25**

Table 2.3: Performance of different classifiers. Knn and Bayesian Classifier perform equally well.

2.10.3 Evaluation Methodology

Regarding the scenario and available data-sets, three different factors seem to influence the possible recogni-

[1] The features and relevant state parameters have been described in publications of the authors of this thesis. The text in the following listing is taken in parts from this publication. All texts taken from this publication have been written by the author of this thesis in the first place:
* Grünerbl A. et al. (2011), [52], please refer to corresponding entries in the literature list or beginning of this chapter

tion accuracy: (1) the feature set, (2) the temporal granularity of recognition and (3) the number of recognized states. These three factors were varied and evaluated:

Feature Sets: The different features, as described above (subsection §2.10.1), and the classification was performed with different combinations of them. This provides detailed information on the location-related behaviors of the respective resident, relevant for recognizing the state of well-being.

Temporal Scale: As described earlier, to eliminate bias by wrongly dated entries in the health records, the ground-truth was extracted on a 14-day basis. Thus it would make sense, to also evaluate the classification on a 14-day block-set. Nevertheless, the classification evaluated on different temporal granularity:

1. 14-day periods, aligned with the health records and the available label-sets (note in a 14-day period sensor data was not necessarily available for each day): for the classification, results of each day were averaged over these 14-day periods.
2. periods of one week (7 days): labels for each week were derived from the available 14-day label-sets; again classification results were averaged over the seven day periods according to their availability;
3. on a single day granularity: again the labels were derived from the available 14-day label-sets (note, in this granularity the label-data, have the lowest accuracy as daily fluctuations cannot be pictured in this form of ground-truth);

Number of States: Also, as described before, during the extraction of ground-truth, two different sets of labels were generated. One for two possible states (positive or negative) label-set, and another three-state (positive, negative, or neutral) label-set. In this regard, both options were used for the classification. Therefore, for each resident, the classification was performed in the light of two and of three possible states. Specifically considering the basis for ground-truth - the health documentation, would not have allowed going for a finer subdivision. As a reminder, the two different state sets were defined as follows (please note that ground truth was normalized between 0 and 1 - see also §2.8,):

- Two State Label Set: the state-indices where split at 0.5 into ”positive state” for values $0.5 - 1$ and ”negative state” for values $0 - 0.49$.
- Three State Label Set: the state-indices where split into thirds. Class ”positive state” for values $0.66 - 1$, ”normal state” for values $0.33 - 0.65$ and ”negative state” for values $0 - 0.32$.

2.11 Results of State Recognition of the Residents

After the qualitative analysis of the sensor data has shown clearly distinguishable effects of the residents' cognitive state or well-being in the sensor data, a quantitative evaluation should confirm this.

2.11.1 Effect of Different Features

Out of the tested features, the “normalized distribution” of the ten different areas per day together with the “sum of entering areas” over the day worked well for all patients. In term of normalized distribution, normalization is meant as dividing the raw values for each area by the sum over all areas. The other features (except “change between areas”, as this is already implicitly included in the ”distribution over areas”) did work for some residents but did not for others:

For example, “Duration of stay” improves the classification result by 8.2 percentage points (pp) for three residents, but for all others, the classification accuracy drops by 26.1% (average -8.9%). The same way, the “index for movement” enhances the classification for three residents by 9.1pp, yet has either no impact (on two other residents) or actively worsens the classification by 16.67% for the sixth resident. If the features “duration of stay” or “movement index” are used uniquely or together (not in combination with “normalized distribution” or “sum of entering areas”) the classification results are terrible (average 35.22 %, 44.5 % and 36.7%).

Performing classification with all available features together enhances the result for three residents by 8.78% on the one side, but do worsen the outcomes significantly for the other three residents by 27pp. As a consequence to the results of the evaluation of different features and combinations of them, any further analysis was based on the combination of the normalized location distribution over the different semantic areas and the number of visited areas.

2.11.2 Two States Analysis

In the two-state analysis, the overall state of the respective resident was determined as either positive or negative. Respectively, the classification was performed with data-sets grouped in 14-days, 7-days, and single day scales. The recognition of two general states is summarized in Table §2.4. For the 14-days periods' analysis, in 3 out of 6 residents, 100% of the states were recognized correctly. In total, the accuracy of the two-state 14-day scale is a high 93.4% (worst accuracy at 80%) with a low standard deviation.

The accuracy drops to 84.1% in the one-week scale analysis. This would not be overly surprising, considering the granularity of the ground-truth. With the ground-truth being only available on a two-week scale, a one-week analysis would likely include outliers not covered by the ground-truth. Nevertheless, interesting to consider here is that the recognition is near to failure for one subject (1031) with a 54% recognition rate, just above random. Looking into this resident's details, revealed that their dementia status was severe with an urge to walk, and the behavior of this resident would vary daily. Excluding this resident from the data-set shows an average recognition accuracy for the other five residents at a high 90.1% (standard deviation 11.3), thus being only less accurate than the 14-day scale results. Still impressive is the fact that one of the residents (1090) remains at 100%. Additionally, it is also noteworthy that one resident (1041) even performed better, and another (1010) almost equal (2pp less) than in the 14-day analysis.

A day by day analysis provides a decrease in recognition accuracy to 70.9% (standard deviation 12.1). This result is not surprising. The assumption, a 14 day period label could be applied to 14 individual days would require a very stable behavior, which is not very likely for most people. Even a generally stable person will likely have a not so good day within two weeks. Such outliers will have affected the results of this daily basis analysis. This also means that the recognition result of an average of 70.9% is the worst case result, and with more detailed ground-truth values might be better. Also, in the daily analysis, again, the recognition for resident 1031 fails at 52.95$. The same argument, as explained above, is valid here. A test subject with a severe state of dementia and the urge to move with many variances is challenging to evaluate.
Hence, the classification on a daily basis, but based on labels that were derived from 14-days period values, will very likely be influenced. Excluding this resident from the analysis draws a slightly better picture with a recognition accuracy in the worst case at 75% (standard deviation 9.3). Likewise, though, a resident with a somewhat stable condition would perform still well, even with fuzzy labels. Resident 1060 is such an example, even daily recognition providing an accuracy of 83.95%.

residents	single		weekly		14-days	
	# DS	2 classes	# DS	2 classes	# DS	2 classes
1060	67	83.91	16	93.33	8	100
1032	159	69.82	34	72.5	17	100
1090	40	78.57	12	100	6	100
1010	120	60.73	28	86	14	88
1031	128	52.95	26	54.44	13	80
1041	73	79.6	20	98.57	10	92.5
average	**98.8**	**70.9**	**22.7**	**84.1**	**11.3**	**93.4**
std dev.	44.8	12.1	8.2	17.7	4.1	8.2

Table 2.4: Results of classification of state with 2 possible state classes (positive state or negative state) for all six residents, with different granularity and respective number of available data-sets (DS)

2.11.3 Three State Analysis

The results of the tree-state analysis with respect to the different time scales are summarized in table §2.5. The average recognition rate of the 14-day periods is 80.94% (standard deviation 11.1). This result is less accurate than the two-state analysis. Still, with three classes an accuracy of 80% is respectable and far above random, with no resident performing worse than 70%. A noteworthy aspect of this analysis, is provided by resident 1032, who performed worst in the two-state case (only 80% accuracy in the 14-day analysis), improves to an impressive 100% correct recognition in the 14-days periods. This is insofar remarkable, as this resident basically failed in the two-state analysis for any other than the two-week case (see also section 5.5 on this).
Looking at the one-week period analysis, the results on average are equal to the 14-day periods. With an accuracy of 78.4% (standard deviation equal to 11.4), it is only approximately 2pp less accurate given the fuzziness in the ground-truth.

Another interesting aspect in the 3 class weekly analysis is the fact that two residents (1060, 1010) perform equally concerning the 14-day analysis and two residents (1090, 1041) perform even better than in the 14-day analysis. These are more or less the same residents that also performed equal or better in the 2 class analysis.
In the daily analysis, the accuracy clearly drops to 55.4%. Still better than just random on average, the results of two residents (1031, 1010) are close to average anyway. Remarkable though, resident 1090 manages to keep the accuracy of 71%, which is slightly better than the 14-day result of this resident.

residents	single		weekly		14-days	
	# DS	3 classes	# DS	3 classes	# DS	3 classes
1060	67	56.52	16	88.33	8	88.33
1032	159	64	34	67.5	17	76.67
1090	40	71.43	12	84	6	70
1010	120	40	28	76	14	76
1031	128	38.18	26	63.33	13	100
1041	73	68	20	91.43	10	75
average	**98**	**55.4**	**23**	**78.4**	**11**	**80.94**
std dev.	44.8	114.3	8.2	11.4	4.1	11.1

Table 2.5: Results of state classification with 3 possible state classes (positive state, normal state or negative state) for all six residents in different granularity and respective number of available data-sets (DS)

2.11.4 Influencing Factors

Looking at the best performing case, the two states in a 14 days classification, two additional factors, which in general are essential for practical system usability, were investigated: the user dependence of the classifiers and the required training set size.

User Dependence From the observations in the nursing home and a general understanding of dementia, it became clear that symptoms of dementia and worsened well-being can be expected to vary from person to person significantly. The results of user-independent classification confirm this. When trained on sets of 5 residents and tested on the sixth, the average recognition rate was just vaguely above random at 54%. Two groups of residents, however, can be distinguished: In the user-independent classification, three residents show results above random with 1031 even reaching 77% accuracy, while the other three were below 50%. This indicates that some residents/dementia patients exhibit symptoms that are more "general" and applicable to other patients. For these residents, user-independent training might be possible.

tested	training set	#DS	corr. class. %	total #DS
1 1060	2,3,4,5,6	7	43.8	16
2 1032	1,3,4,5,6	18	52.9	34
3 1090	1,2,4,5,6	7	58.3	12
4 1010	1,2,3,5,6	13	46.4	28
5 1031	1,2,3,4,6	20	76.9	26
6 1041	1,2,3,4,5	9	45.0	20
Mean			53.9	

Table 2.6: Inter Subject Classification, trained with 5 subjects, tested with the remaining dataset

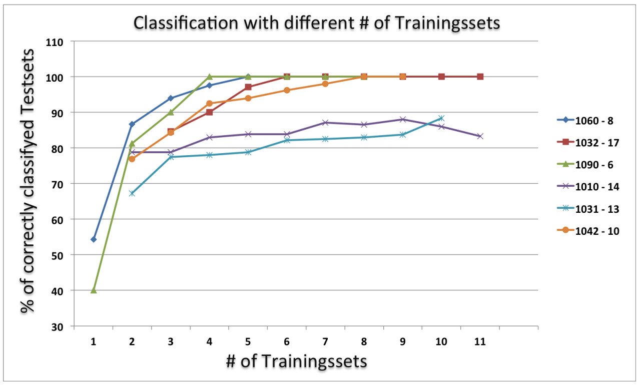


Figure 2.8: Classification accuracy with an increasing numbers of training sets, saturation at 5 training sets.

Required Amount of Training Data Given the results of the user-independent training, it is not possible to develop a plug and play pre-trained system for any new user. Thus a further crucial question to understand is the required amount of training data that is needed to train the system for a particular individual user. Figure §2.8 investigates this question. The figure shows the recognition accuracy for each of the participants as a function of the number of training samples. As can be seen, the "elbow" of the curve (the point where the curves level off) lies at four samples (of 14 days periods) for most participants, and most are reaching saturation after five samples. This indicates that running the system for approximately two months while documenting the state of the users (residents) for every 14-day periods would be enough to reach a sufficient classification accuracy in most cases. This seems like a reasonable effort.

2.11.5 Discussion of Classification Results

Overall the results indicate that the analysis of location traces is well suited for

1. assessment of the state of dementia patients with a limited number of coarse states,
2. on a user-specific basis, and
3. over periods of several days.

Although the 2-states-classification seems to perform better than the 3-state-classification, the available data does not allow to conclusively answer the question, whether 2 or 3 states are more fitting to the resident's actual well-being.

Anyway, this might be a matter of the personality of the respective patient/resident. Some of them will likely have a more two-state behavior (either good or bad) others might also and mainly act neutrally (neither incredibly good or bad but just average). Especially the fact, though, that the performance-decrease is less pronounced for the weekly accumulation of data (where there is more training data) seems to indicate that the problem may not be as dependent on the number of states but the amount of available training data.

2.12 Nurses Reception of a Real-life Study in Health Care:

Despite the challenges a long-term study will imposes to the technology, it also specifically affects the life and work of the people who have to deal with the study to a very distinctive level. In the nursing home, the affected persons were not only the residents but especially the nursing personnel. As already mentioned earlier, at the beginning of the study year, the nursing personnel was not only skeptical about the reasonableness of the study but also were even a bit afraid of the technology. Sentences like "will this technology be able to monitor us?" or "isn't radiation like Bluetooth or UWB harmful?" or "oh these sensors and wires do not look nice," could be heard rather frequently during the first weeks.

Nevertheless, the nursing staff had to learn to deal with the study, and for nearly one year these nurses handled residents refusing to wear tags, and searched for tags that were laid aside, and more. So at the end of the study year, the nurses had collected experience in dealing with technology in a way researchers typically will not be able to. This experience can mean valuable insight into handling long-term studies with patients being monitored and interacting with patients who have to deal with wearing technology.

Thus, it was reasonable to give the nursing personnel the possibility to share their experiences. This way, we would hopefully learn some valuable lessons and draw some conclusions for any future deployments.

2.12.1 Questionnaire

So, to assess the nurses' view on technology and its potential in health-care and nursing, we designed a simple questionnaire. This questionnaire covered topics like: "amount of additional work for them personally", "disturbance by visible parts of sensor installation", "effect of having project personnel around frequently", "impact of wearing a Tag for the residents", "personal view about sensor technology and its benefits in health care and the project itself."

Modality The questionnaire itself was close-ended for the most parts, in ordinal-polygamous (partially dichotomous) form, but it also included some open-ended questions to provide the nurses an option to give their personal feedback. In terms of the close-ended questions, four (for some distinct questions two) ordered options were provided. These were: yes, a little/possible, hardly, and absolutely not. A neutral option was intentionally left out as the nurses should be made to deliberately give either a more positive or a more negative answer and avoid central tendencies. The questions themselves were all formulated neutrally to avoid tendencies.

Contents The details in the questionnaire and each question can be inspected in the Appendix §7.3. The questions covered following topics:

- Additional workload caused by the study.
- Visible sensor installations.
- Impact of sensors/technology on involved persons (patients, nurses).
- Impact of the need to "wear" sensors, e.g., convincing residents to wear them.
- Potential of the study and the study's topic for nursing and health-care.
- General potential of supporting technologies for nursing and health-care.
- Communication policies of study and authorities.
- Dealing with not medical study personnel.
- Personal opinion about different aspects.

2.12.2 The Nurse's Perception - Response to the Questionnaire

Over the entire study period, in total, 14 nurses were involved. In the first months of the study, the nurses changed frequently. During the second half of the study year, the nursing personnel remained stable. Four of the 14 nurses had left the home during the first months and thus could not be reached for filling in the questionnaire. Thus the questionnaire was handed to ten nurses. Out of those ten nurses, 9 filled it in and returned the questionnaires, meaning a high return rate of 90%. This high response shows that the involved nurses were eager to pass on their experiences. [1]

The first part of the questionnaire accumulates questions about the deployment of the study and its impact on the nurse's work and the way the residents were able to deal with the study and its implementation. Overall the responses were somewhat more positive than negative. Negative replies primarily were related to the frequent necessity of having to look for laid aside tags and problems in convincing the residents to wear the tags. This was explicitly stated for the beginning of the study. Details about questions and explicit responses can be found in Table §1 in the Appendix.

Additional Burden: On average, the nurses stated that it took them approximately 15 minutes of extra work per day to make sure the residents were wearing their tag, or search for them. Nevertheless, two third (66%) of the responses stated that they did not or barely experience the extra work (explicitly stated was the supervision of wearing the tags) as an additional burden. Only 22% expressed that it actually was an additional burden for their work, which shows that small amounts of extra work do not seem to bother the affected people very much.

Visual Installations: About the visual and partially provisional installation, 62.5% of the respondents stated that it actually did disturb them in the beginning. Anyway, all nurses (replies were: 78% yes, 22% mostly) got accustomed to it, because the installations (wall mounted sensor, wires) themselves did not or just hardly affected their actual work. An effect this study could profit from is the fact that humans tend to get accustomed to things they see frequently. This means that luckily, the provisional placement of additional wires, in the beginning, had no adverse effect on the actual success of the study.

Influence on the Patients: In regards to the influence the study had on the residents, 50% stated that the sensor tags, in fact, had some impact on the residents (some refused to wear tags) and 60% said that this influence was long term but not equal for all residents (75%). Additionally, 89% expressed that there were problems to convince the residents to wear tags in the beginning, but 78% admitted that the residents got used to it. This was the most critical part of the study since the success of the study did entirely depend on wearing the tags. Fortunately, the residents got used to it. Thus, the initial resentments (specifically of the nurses) only partially influenced the outcome of the study. The statements of the nurses, in general, reflect the situation we could find in the amount of recorded data.

Influence on the Nurses: The question, how the study influenced the nurses in a personal way, was replied by

[1] Most texts in the following section and subsection about the Nurse's Perception have been taken in part from the following paper. All texts taken from this paper have been written by the author of this thesis:
* Grünerbl A. et al. (2013), [53] please refer to respective entries in the literature list or beginning of this chapter

75% as having some impact on their very own work (again the frequent searches for laid aside tags were mentioned) but just 56% felt that it had some influence on them personally (mood, burden). Furthermore, 89% stated that neither the residents nor the nursing staffs were much disturbed by the technical study personnel that had to maintain the sensors and technical installations frequently. The second part of the questionnaire focused on the nurses' personal opinion about technology in general and the potential of technology to assist them in their everyday nursing work.

Personal opinion: In parts surprisingly, after experiencing this one-year study with all the ups and downs, the nurses mostly had a positive attitude towards technology in health care. 93.75% of the respondents considered the topic of the project capable of enhancing the life-quality of the elderly. All (100%) considered it somehow possible to draw useful conclusions from sensor data, and 78% expect to get a benefit for their work if such a system would be installed permanently. These three last statements show that the nursing staff is open-minded for new developments and the usage of technology to assist their work. It furthermore shows that, despite the additional burden, they were able to see the benefit of such studies, even though they personally did not benefit.

Additional readings stressed once more that the significant effort for the nursing staff was to convince the residents to wear the tags and to search for tags laid aside which, as expressed before, was feasible for them in the amount the study required. One hopeful message of the answers provided by the nurses was the positive overall perception of the involved people. Even expected issues and concerns from our side, e.g., searching for laid aside tags, turned out to be acceptable for the nurses. Furthermore, the general unexpected positive attitude of the nurses towards technology in health care encourages to proceed this way.

2.13 Discussion and Future Suggestions

This thesis chapter has introduced a long-term, real-life deployment of a pervasive indoor-location monitoring system for the assessment of dementia patients. Despite the challenges imposed by the real-life deployment, the results strongly support the notion that such technology can be deployed successfully on a large scale. Obviously, this would be a precondition if such a system was to be installed professionally for monitoring the state of dementia patients. Thus, the results provided by this thesis have a strong indication that such systems are possible. The major strength of the study includes the realistic and real-life setting, a set of study subjects with different symptoms and severity of dementia and the extended period of observation.
The Maieutics care concept (several residents living in an apartment like a joint family), implemented in the ward is not only becoming more popular but also resembles a somewhat home-like environment. Thus, it is reasonable to defer from the results of this study, that such a system would also work in the private home of a dementia patient.

The main limitation of the study, though is the ground truth. Only basing any analysis on coarse entries of health records kept at the ward leaves a fuzziness to the results. Despite, as explained before, based on nursing literature and discussion with several nursing professionals, I am convinced that within the type of analysis that has been performed (two or three states, averaging over 14 days) the ground truth can be trusted. Moreover, even though the impossibility of additional daily observations over the long study period, seemed to be a disadvantage in the first place, it still allows a further interpretation. Because, as the outcome of the analysis provide excellent results based on the record entries, this also means that the sensor system is equally good as the health records, or at least do not stand behind the current method of documentation (where the most diagnoses of the doctors will be based on).

However, having a much more detailed ground truth (preferably systematic assessment for each day) would have provided more options to understand the system performance and potential better, and would have allowed going into more details in the analysis. Thus, if this work were to be seized further, it would be interesting to know whether and how the system could be used for a daily rather than weekly or bi-weekly state analysis. Also, it would be interesting, whether it was possible to detect changes in well-being within a day (as much as many mental disorders have a morning/evening behavior expression).

In general, the feedback from the nurses regarding the deployment of a study in the daily life of a home for the elderly has been very positive. Even though there were various issues during the starting phase of the study and some extra work and additional burden for the nurses, the overall perception was positive. This feedback should definitely encourage further study deployments in real-life health care.

2.13.1 Suggestions and Lessons Learned

A hospital or nursing-home ward is a bustling environment. In recent years experience has shown, that especially in health-care settings, people likely have encountered technical studies or have experienced the introduction of a new technological system, which intended to make their lives easier. Unfortunately, most of the clinical personnel will have experienced such studies and/or systems negatively. Many will have perceived such systems as additional workloads, or as systems not fitting to their specific needs. Many nurses have experienced the use of parallel-systems. Because, after implementing a new system, many wards will eventually go back to use the old system because it was more suitable for the ward's needs. Subsequently, the nurses working in the wards will have to ”suffer” the new system as redundancy in parallel (because they have to). In this regard, nursing stuff will most likely react critically or at best reserved towards a new technical system. Nevertheless, in clinical environments, the success of a study highly depends on positive interaction with the nursing personnel. Therefore, some aspects have to be considered (ideally in advance) to maximize the success of a clinical deployment:

- **Autonomous System Design:** Nursing staff usually do not have extra time to operate the sensor systems. If a system requires any kind of additional work for the nurses, it might limit the success of it, because in times of emergencies or on hectic days, even small tasks like checking batteries might be forgotten. It is therefore essential to use sensor systems that are easy to control remotely and operate on main power supply or batteries which do not need to be changed or provide the workforce to perform any additional maintenance.
- **Communication of purpose to all who might be affected:** Managing to communicate the purpose of the study to those who will be affected by it, is a main factor of success. In the presented study, the nursing personnel was not actually part of the study. Still, they were greatly affected by it. Without the nursing staff frequently checking whether the residents would wear the tags, the study would have suffered.
- **Acceptance of System and Study Personnel:** One aspect, that most technicians tend to ignore, is that the nursing staff has to accept both the technical system and the study personnel. Specifically in clinical settings, the access can be restricted and never is just simply free (because in a ward there will always be someone working, even at night).

The above can be summarized as:

an essential factor in the success of a real-life deployment is that the main actors in the study understand why the study makes sense and can see the potential benefit of it.

Then they will be willing to invest some time and effort to make the study a success. If not, researchers will have to spend much effort to keep the study running. Sensor systems should be robust and operate stably. Using cutting-edge technology may be risky. Researcher-nursing staff relation should be excellent, and the nursing staff (or any party that is affected by the study) should be informed frequently and thoroughly during the study. Additional work should be honored (maybe financially). Technical systems have to be accepted by the nursing staff - one fact that most technicians simply ignore.

Assessing Cognitive State Change: Monitoring State Progression of Psychiatric Patients With Motion and Location Traces

There are different theories why! Even at cracker barrels, this question has been discussed frequently. Still, it is a fact that the number of people suffering from mood disorders continually increases.
Estimates say that by 2030 affective mood disorders will contribute to the highest disease burden in the developed world [1]. However, already today in European countries mental disorders are the unlucky number one in newly diagnosed illnesses each year. Despite all theories about why that is so, or why these numbers are rising so quickly, and despite all efforts to eliminate the causes, mental disorders are a lingering issue in our society.

In this regard, a still not sufficiently answered question is the “how to treat people with such disorders effectively”. As has been pointed out, today’s healthcare is advancing year by year and prevention mechanisms for physical illnesses as heart attacks are giving much thought. So why is it that care and prevention of affective mental disorders do not seem to improve?
Generally, the success of today’s health care highly relies on the availability of devices that enable healthcare professionals to measure various physiological parameter objectively. These can be simple procedures like measuring temperature or blood pressure, but there is also a variety of increasingly complex image processing machines requiring high computational power. Thus, technology has found its way into any kind of health care, with one exception: psychiatry and mental care.

In psychiatric care, almost any aspect still relies on non-technical tools. As such, a psychiatrist’s or psychologist’s primary tools are interviews with patients and patients’ self-assessments. Reasons for this kind of “staying in the stone ages” simply is that mental affection happens in the mind/brain of a person, and as of now, there has no technology been developed that would make it possible to look into the mind of a person and, with high certainty, could determine that “this particular mind pattern is bipolar”. On the other side, most mental illnesses are manifested in behavior, meaning the way a person behaves and the extent of how this behavior change carries indications about mental states. However, until recently devices to measure behavior had not been developed.
With the development of wearable devices and their broad availability in our lives, this is now changing. The following chapter of this dissertation introduces all steps from an idea and the availability of smartphones to the development of a new methodology to monitor affection-related behavior and detection of mental state changes.

This chapter is split into two parts. The first part attempts to evaluate whether based sensor recordings carry enough information about the user’s mental state to determine disease relevant features and patterns. This first part relies on a small study with six bipolar patients over the course of 6 weeks. The second part, proceeding from the results of the first part, will introduce methods to determine the state of bipolar patients, only by using objective sensor-data, recorded with the patients’ smart-phones. Furthermore, a method to determine the point in time when a state of a patient starts to change will be introduced. This second part draws upon a real-life data-set of 10 patients, recorded over a period of 12 weeks (in total over 800 days of data tracing 17 state changes) by four different sensing modalities.

The work, contents, all pictures, most tables, and also partially text-passages of this Chapter has been published by the author of this thesis in the following publications. The author of this thesis has written any text in this Chapter, specifically text-passages of paragraphs taken from these publications. For more details about these publications please also refer to the entries in the literature list:

- Gruenerbl A. et al., Smart-phone Based Recognition of States and State Changes in Bipolar Disorder Patients, IEEE Journal of Biomedical and Health Informatics (J-BHI), 19, 140-148 (2014) [91]
- Gruenerbl A. et al. "Using mobility traces for the diagnosis of depressive and manic episodes in bipolar patients." Proceedings of the 5th Augmented Human International Conference (AH 2014). ACM, 2014. [92]
- Gruenerbl A. et al. "Towards based monitoring of bipolar disorder." Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare. ACM, 2012. [93]
- Gruenerbl A. et al. Ubiquitous Context-Aware Monitoring Systems in Psychiatric and Mental Care: Challenges and Issues of Real Life Deployments. ICCASA 2014 Conference Proceedings, (ELLCAMA-2014 Workshop), October 15-16, Dubai, UAE, ACM Digital Library, 2014. [54]
- Muaremi A.; Gravenhorst F.; Gruenerbl A.; et al. "Assessing Bipolar Episodes using Speech Cues derived from Phone Calls". In: International Symposium on Pervasive Computing Paradigms for Mental Health (MindCare)., May 8-9, Tokyo, Japan, Springer Link, 5/2014. [94]

3.1 Motivation

Affective mood disorders are the most common brain-related diseases, accounting for 55.1% of mental illnesses and the number of affected people has been growing steadily for the last 30 years (an increase of 37% world-wide between 1990 and 2010). Cost-effective interventions exist, but less than 1% of mental health budgets are spent on measures to prevent crises. Besides the high economic burden, the human suffering associated with them is immense. Bipolar Disorders [95] are a common and severe form of the affective mood disorders. People who are suffering from this disorder experience more or less regular successions of periods of manic, normal and depressive state. The current standard for determining the severity of an episode uses subjective clinical rating scales based on self-reporting that were developed in the early 1960s (e.g., HAMD, BRAMS scales) or more recent variations (e.g., BSDS). While the efficacy of these scales has been proven, they still are a potential source of subjectivity and additionally require the attendance of a trained professional.

The main treatment currently offered is a life of pharmacotherapy, which has to be modified according to a patient's state. Additional substances may have to be prescribed to increase the prophylactic effect of the therapy. Even so, the effectiveness of treatment strongly depends on the timing. Thus, therapeutic measures can be very effective if administered at the beginning of a patient's transition into a different state (e.g., from normal to depressive). They may be a lot less useful if severe symptoms have persisted for a significant time. As a consequence, a promising form of intervention is teaching patients to recognize and manage early warning signs (EWS). A systematic review of this approach found that 11 randomized controlled trials (RCTs) involving 1324 patients show the efficacy of interventions that include EWS self-recognition [96]. However, this involves a very significant training effort (which is difficult to finance) and strongly depends on the patients' compliance and discipline. Thus, in some cases it can be impractical or even impossible and therefore its usage has limitations. Cognitive, mental and emotional disorders are an apparent application field for activity recognition. As the symptoms of such diseases manifest themselves in changes of behavior [97], activity-aware systems could be used as core instruments for assisting diagnosis and treatment. Even more, the fact that psychiatrists currently have few objective and reliable alternatives would amplify the value of such a system. Ever since X-rays have become available, it is much easier to see exactly how extensive a fracture of a broken limb is and how best to attend it. On the contrary, most of the time psychiatrists have to rely on a patient's subjective recollection of their behavior. The closest thing to a "measurement" is self-assessment questionnaires that can be time-consuming and rely on subjective recollections and the patients' self-perception only. As consequence patients often end up visiting the doctor very late, which makes treatment more difficult and often leads to the necessity of severe measures and prolonged hospitalization. On the one hand, this can have a dramatic impact on the patient's life (long sick-leaves), and is of costly relevance to the health system.

While the benefit of a more "objective" measurement based on activity recognition is clear, developing and implementing such a system is difficult for many reasons. First, having people (who have a mental disorder) wear multiple sensors on a daily basis is often not practicable. Second, since there are no reliable automatic diagnostic instruments, getting enough ground truth for training and testing involves a considerable effort regarding long-running trials involving repeated appointments with professionals. Finally, the fact that behavior can vary strongly on a daily basis, independently of illness-based effects, makes recognition difficult. As a consequence, very little work exists on diagnostic work using pervasive sensors in real-world environments.
By overcoming such difficulties, this chapter demonstrates how smart-phone usage patterns and sensor data can be used as an objective "measurement device" for aiding psychiatric care.

3.2 Related Work

Even though the research field of pervasive health-care and more so mental-health-care is still relatively novel, a reasonable amount of work has been done already. Please note that this related work in general paints a picture of the state of work at the time the work in this chapter was done. [1]

Overviews of the usage of wearable and pervasive technology in healthcare are given by [23], [67], [66], [24] and [98]. Specific examples range from assisting older people with cognitive impairment [99], to monitoring children's developmental progress using augmented toys and activity recognition [100].

In the area of mental health, the majority of systems deployed focus on supporting self-monitoring. Systems that provide patient feedback through questionnaires or text messages are analyzed in [101], [102], and [103]. Simpson et al. [104] apply interactive voice response self-monitoring for alcohol abuse disorder patients. Nev-

[1] Major parts of the Related Work and Text in the related work have been taken from following papers of the author of this thesis. Any text taken from these papers has been written by the author of this thesis:
Gruenerbl A. et al., 2015 [91], Gruenerbl A. et al., 2014 [92], Gruenerbl A. et al., 2012 [93], and Gruenerbl A. et al., 2014 [54], please refer to respective entries in the literature list or beginning of this chapter

ertheless, despite the usability of self-assessments and questionnaires it is far more intriguing to perform diagnostics itself via a wearable technology [105]

3.2.1 Self-assessment with mobile-phones

Smartphones are playing an increasingly important role in gauging mental health. For instance, there are apps designed for self-assessment that can help patients to estimate and monitor symptoms specific to their ailment, which can then be shared with psychiatrists. For example, the eMoods Bipolar Mood Tracker app [106] provides a system that allows users to input subjective mood ratings daily and monitor them via an electronic journal. The app can also keep track of hours of sleep, anxiety levels, and medication use, which are all self-reported, and can be shared with a family member, caregiver, or clinician.
A number of other approaches have looked at incorporating Ecological Momentary Assessments (EMA) in order to gather patient state at opportune times [107] specifically for anxiety and eating disorders [108] and also provide Ecological Momentary Interventions (EMI). In this study, authors stress the use of external context clues, based on sensor data such as location and social interaction, to deliver effective interventions. Another set of studies that relied on self-monitoring of patients with severe mental illness (SMI), specifically bipolar disorder and schizophrenia are presented in [109] and [110]. Authors report evidence of short-term adherence to and acceptability of mobile devices while emphasizing that it is likely impractical for patients to respond to daily surveys, stressing that context-awareness of mobile devices and sensor sampling can provide feedback relevant to detected patterns of behavior.
Similarly, a randomized controlled trial [111] revealed that while self-reporting and self-assessment of patient state has a positive effect in increasing emotional self-awareness (ESA) in patients suffering from depression, anxiety and stress, the mental health outcomes did not improve significantly. As such, considering the impracticality of this method for long-term monitoring [109] and patients' reluctance to log information [112], there is a clear need to infer patients' states autonomously. Recent trends in this area have been pointing towards using sensors on smart-phones for collecting objective data.

3.2.2 Objective Monitoring

Objective monitoring consists of smart-phone sensors passively collecting data that can be used to infer patient state. Matthews et al. [113] outline different aspects in balancing sensing and patient's need and describe MoodRhythm, a system for tracking daily rhythms. There is far less work in automatically inferring patient state in comparison to self-reported information. One possible approach is to develop systems that predict patient state by using predefined algorithms that are initialized based on evidence from scientific or clinical knowledge [114, 115]. This has been the typical approach of systems that recognize patient activities, where algorithms make inferences regarding the patient's status by plugging in sensor data.

3.2.3 Automatic Recognition of State

Regarding automatic recognition of mental state much less work exists, in particular, work involving real-world studies and off the shelf devices like smart-phones. Massey et al. [116] describe an experimental analysis of a mobile health system for mood disorders where they introduce different possible sensors for mood detection, yet focus on technical aspects like the line of sight and reception rate, optimal coverage and optimal placement of on-body sensors.

Paradiso et al.[117] introduce a personalized monitoring system based on sensing physiological and biochemical signals. Burns et al. [118] introduce a application which tries to predict a patient's mood (of depressive patients) using machine learning models, yet (contrary to our approach) requiring constant interaction (5+ times a day) with the test patients and with no psychological assessment available.
In [119] LiKamWa et al. present an i-phone approach to infer a user's mood - yet again requiring constant mood input from the user - and present the results of a field study with 25 random test subjects. Furthermore, "TrueColours" [120] and the "Optimism App" [121] were developed to log self-reported mood, activities and quality of sleep in order to monitor depression and state changes. Two publications with similarities to the here published work are the research done by a group from Denmark [122] and the previously mentioned [118] that introduces a mobile phone application which employs machine learning models to try to predict patients' mood (of depressive patients). Here however, the ground-truth is fully self-rated, no objective psychological or psychiatric assessment is performed.

In [122], Frost et al. use a self-developed application to record subjective and objective data from patients who have bipolar disorder. Even though their main focus lies on self-reported information, they also utilize coarse objective sensor data (acceleration fragments and phone call statistics) and try to estimate future shifts in a patient's mental state. These predictions are then compared to forecasts derived from the self-reporting data. By contrast, our work goes into far more depth in the area of state classification, also uses location sensors and sound in addition to acceleration and instead of social interaction sensing compares the results to an objective, diagnostic ground-truth on a day to day basis.

3.3 Objectives and Contribution

The objective of this chapter can briefly be summarized as: find a way to objectively determine the cognitive state of patients with affective mood disorder during their everyday life, only by relying on a smart-phone. This would undoubtedly increase the complexity of the work in comparison to the work in the previous chapter. Thus the method in this chapter was to go step by step. First, a do-ability evaluation should evaluate whether the objectives of this chapter can be successful at all. With a positive outcome of the do-ability evaluation, a second and more extensive study should allow developing algorithms to determine bipolar states and predict transitions.

First of all two aspects were relevant to know: since such a system has to operate in the patient's everyday life, it was essential to understand whether a technical application would be accepted by bipolar patients and further what kind of data such a technical system should be based on? Discussions with patients revealed that smart-phones were an acceptable technology, as patients themselves were using them. Therefore the first phase of the development should comprise a real-life study where data of a small group of bipolar patients should be gathered and, in the course of this data measurement, it should be evaluated whether the test subjects were able to deal with the measuring device and would be willing to accept it. The data gathered in this first phase should suffice to detect patterns which can be correlated to the patient's mental condition and possible transitions into other bipolar episodes. See section §3.4
In this regard the main question to be answered is: do the sensors contained in a suffice to record disease-relevant information? On a second level, it should be evaluated whether it would be possible to extract enough relevant parameters, from these sensors' data. For this first part, an initial mainly qualitative study was conducted with ten patients over the course of 6 weeks per participant. The study was conducted under the supervision of psychiatrists and medical psychologists treating the patients and with the approval of the University of Innsbruck ethics board.

By drawing conclusions from the first part, and after leading various discussions with psychiatrists and other health care providers, the next step could be designed. This second part sets its goal of developing an application based on smart-phone behavior and activity monitoring that would be usable as an "objective" measurement device, which would help to determine state and detect state changes. To be more precise, the application should only and exclusively rely on objective sensor data and should work without any input or feedback from the user/patient. In order to achieve these goals, a second and larger data collection study was implemented, which would cover the entire package. See section §3.5. For ten months, a total of 10-15 patients collected 24/7 sensor-data on their smart-phones. At the same time, all patients filled-in a daily self-assessment questionnaire and were frequently checked (with the standardized methods used in psychiatry) by psychiatrists and medical psychologists. Next, the pre-processing and methodology to extract appropriate features is described in section §3.7. With the necessary features at hand state classification algorithms were applied on single sensor modalities and also an algorithm to fuse different sensor modalities was evaluated. See section §3.8. Since determining when patients changes their mental state is more important than which state a patient is in, a change detection algorithm was developed an analyzed. See section §3.9.

This chapter closes with a report on challenges that had to be faced in this particularly sensitive area (Section §3.10) and an extensive discussion of the results and possible future outlooks (Section §3.11).

3.4 Do-ability Evaluation

At the beginning of the development, quite some thoughts were invested into how such a system should look like and which steps had to be taken to realize the development of a final system. Thus, in the first step in order to understand whether the aimed at goal was possible to reach and if yes, what information was essential to collect, a do-ability evaluation data collection was performed. In this sense, a small real-life study was set up to collect the needed data and information.
The general idea was to gather real-life data from real patients, take this data and analyze it towards finding and extracting features that would resemble the patient's respective state. Still, as already mentioned the main question to this evaluation was: do the sensors contained in a smart-phone suffice to record disease-relevant information? Thus a smart-phone, used by bipolar patients during their everyday lives, should collect all data that is relevant to analyze the patients' behavior and draw conclusions about the patients' states. So first it had to be understood what this kind of relevant information would be and how to retrieve it.

3.4.1 Relevant Information

Before sitting down to create an application for recording sensor data, and deciding on an appropriate platform, a number of constructive discussions with psychiatrists were lead to get a picture of the most critical and meaningful features in the behavior of bipolar patients. These should provide information about the patient's mental condition or would indicate the transition point between episodes. The following three behavioral patterns were identified as likely being disease-relevant or indicative of an impending change in the episode were identified: [1]

Location and Movement: It is commonly known, that depressive people tend to withdraw to their homes while for manic people the world cannot be large enough. Therefore, the way a patient moves around in their surroundings along with places visited throughout the day indoors and outdoors are an essential factor in determining the state and state changes of a bipolar patient. It can display a picture of the movement patterns of a person over a day and on a more extended scale the changes of these patterns over weeks.

Level of Activity: Another indication for changes in bipolar episodes is the amount and intensity of physical activity. While depressive people often find it difficult to motivate themselves to do anything or concentrate on simple everyday tasks, manic people feel like they cannot perform enough activities.

Social Interaction: Social interaction or the possible willingness to get in contact with other people is a third indicator of a likely change in state. A manic person has a heightened urge to interact with others and possibly a reduced sense of personal boundaries, while a depressive person suffers reduced social abilities. Depending on the type of the episode (manic, depressive or balanced) and the severity of it, a person will try to interact more or less with others. At the beginning stages of a depressive episode, e.g., a patient might feel an urge to talk about their problems, but a severely depressed patient might not even be able to pick up the phone.

3.4.2 Monitoring System

After identifying the main requirements that should be covered by the sensor data, a smart-phone application had to be implemented that would allow recording just these relevant information. [1]

This application was Android-based (see fig.3.1(a)). It would work automatically on device boot and would run almost invisible to the users in the smart phone's background. Further, it did not require any user interaction. On-demand though, it provided the possibility to turn the measurement off or on. In regards to privacy and the patient's sovereignty of their data, this was an important aspect. Mentally disordered patients, but not only they, might feel the need to be in charge of the things happening around them (including the data recording). To cover the three main relevant indicators as described before (movement, activity and social interaction) following smart-phone sensor system were used to record data:

Location and Movement: Regarding recording changes in the test subject's position three different sensors could be used: GPS location: GPS was used for location tracing outdoors. To guarantee the patient's privacy all GPS sensor readings were anonymized (transformed into a neutral coordinate system) before further processing. WiFi Network: WiFi cell information and signal strength should be used to get a kind of indoor location tracking 3G Network: the 3G network could be used indoors and outdoors, primarily to assist GPS and WiFi positioning.

Level of Activity: In order to determine the amount of activity, mainly two sensor readings were to be used: Acceleration Data: Acceleration data could be used to determine motion and unmoved time-periods and might enable the recognition of specific activities. Magnetic Field Data (Compass): orientation of the magnetic field of the three axes including a time stamp should provide the heading of the movement.

Social Interaction Within social interaction in a modern world, the cell-phone itself plays an important role. Here the number of phone calls or text messages and specifically the outgoing ones give an overview of the person's desire to communicate. On the other hand an increasing number of incoming messages/phone-calls (while decreasing outgoing connections) could indicate the concern of the circle of friends or acquaintances towards the patient's state!

For security reasons, all recorded data was encrypted (AES 128bit software side encryption). The recorded data was stored on an SD card and was transferred to the study data storage during the patient's appointment at the hospital. Within typical everyday usage (including phone calls and limited web usage) the battery would last up to 10-15 hour. This did require to charge the smartphone each night, yet still was sufficient for the usage throughout the day. See figure 3.1(a)

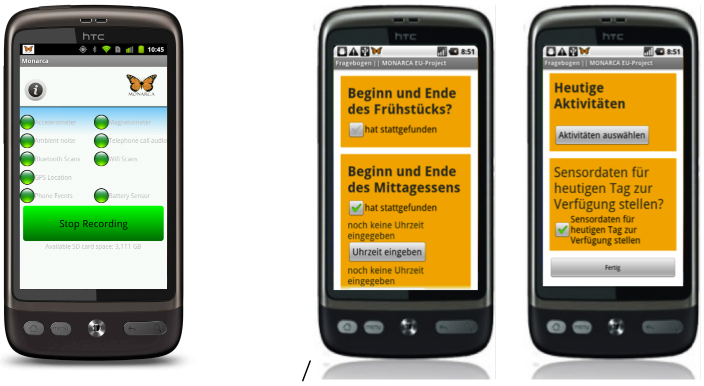
3.4.3 Ground-Truth Acquisition

This study took place not only in a real-world-like setting but actually in the every day real-lives of the participating patients. In this regard, there was no possibility to collect a ground-truth in the way a lab-study would produce. Meaning there was not even a way to document the activities the participants would perform daily, or which kind of places they would visit. In fact, the possibilities for ground-truth were minimal, even more, limited than in the previous chapter. On the other hand, as this was the goal, the study was dealing with a new way of using and interpreting sensor readings, with only limited knowledge to rely on. Therefore an elaborate ground-truth actually would have been crucial for the success. In case of this study, ground-truth did not only mean to record the reality of what kind of motion and location were happening but also to understand in which mental state the participant was in.

During this study, we had to accept that real ground-

[1] The following listing and description of the relevant parameters have been introduced in the following publications: Grünerbl A. et al., 2012 [93], and Grünerbl A. et al., 2014 [92], please refer to respective entries in the literature list or beginning of this chapter

[1] The logging app was implemented in course of the Monarca EU FP7 project by Jens Weppner and Amir Muaremi.


htc
10:45
Monarca
MONARCA
Accelerometer
Magnetometer
Ambient noise
Telephone call audio
Bluetooth Scans
Wifi Scans
GPS Location
Phone Events
Battery Sensor
Stop Recording
Available SD card space: 3,111 GB
htc
8:51
Fragebogen || MONARCA EU-Project
Beginn und Ende des Frühstücks?
hat stattgefunden
Beginn und Ende des Mittagessens
hat stattgefunden
noch keine Uhrzeit eingegeben
Uhrzeit eingeben
noch keine Uhrzeit eingegeben
htc
8:51
Fragebogen || MONARCA EU-Project
Heutige Aktivitäten
Aktivitäten auswählen
Sensordaten für heutigen Tag zur Verfügung stellen?
Sensordaten für heutigen Tag zur Verfügung stellen
Fertig
/

(a) data recording application

(b) Daily questionnaire

Figure 3.1: The smart phone application for data recording runs autonomously in the back ground and triggers the daily questionnaire in the evening.

truth was not available. Generally, in order to gather some baseline information about the participant's actual bipolar state (diagnosis) and episode changes, two ways were possible.

Daily Self-Assessment: First, to collect a basic picture of the participant's day, a short, daily questionnaire for the participants to fill-in, was developed. This questionnaire was integrated into the smart-phone application and was triggered to appear automatically at a particular time in the evening and would only require 5-10 minutes per day. Within approximately ten questions, the participants were asked to provide information about their activities of daily living (ADL). These questions were taken out of the following topics: Which meals were taken and when; what general activities were performed (indoors, outdoors, extracurricular, etc.) and when; how much time was spent on repeated activities; how many hours of sleep were taken; which common places were visited? In addition to these questions, three self-rating questions about the psychological state, the physical state and the amount of activity were included and could be answered by assigning 1 star (bad) to 5 stars (good). Note that the assignment of stars to the questions was handled in a reverse-school grade way, as this was more intuitive for the patients, rather than assigning a complex psychiatric scale. Figure 3.1(b) displays two examples of this questionnaire (in German, as the study was conducted in a Austria).

Periodic Professional Examinations: Since self-assessments are subjective, a second more objective ground-truth should be collected in the form of psychiatric assessments and psychological scale tests. Both are part of clinical standards. Thus it was necessary for the participants to come to the hospital for these assessments. The psychiatric assessment was performed by psychiatric professionals at the psychiatric hospital. They mainly focused on analyzing changes in the mental state and the respective behavior of the residents, in comparison to the previous assessment.
In a second assessment, to get a profound psychological idea of a participant's current mental health, standardized psychological scale tests were applied. Note that these scale tests are currently the main way to rate the mental state of a person somewhat objectively. Still, even though specially trained psychologists perform these scale tests, the results have a subjective quality to them. Since the answers are given by a person, even though as highly trained to be as objective as possible, still, there is no way to guarantee an equally objective attitude throughout the study. Following widely recognized scale tests were performed:

- *Hamilton Depression Scale (HAMD):* standard scale for determining the degree of depression. Performed by trained clinical psychologists.
- *The Common Depression Scale (ADS):* self-rating scale for depression. A self-assessment performed under the supervision of trained clinical psychologists.
- *The Mania Self-rating Scale (MSS):* self-rating scale for mania. A self-assessment performed under the supervision of trained clinical psychologists.

The assessment and the scale tests were performed at least three times during the study period, once at the beginning, once at the end and one after three weeks. In case the study period had to take longer than six weeks for any reason an assessment was performed latest every three weeks. Initially, the plan was to schedule assessments at least once a week, to cover the study as tightly as possible. Psychologists advised against such a high frequent scale testing. Since these tests are standardized, they include a specific set of questions that have to be asked in a specific order and in a specific way. Repeating these scales very frequently, would have caused a memory effect and thus biasing the outcome significantly.

3.4.4 Study Execution

This first do-ability study included a total of 10 patients 8 of them female, two male, in the age range of 33 to 48. Eight of the study participants were recruited during an in-patient stay in the participating psychiatric hospital. The majority of the patients (six) was discharged within two weeks after they were included in the study. Only two remained in the hospital during the entire study period.

The recruiting was done solely by the wards psychiatrists, who determined which patients would be both mentally and physically fit to deal with the requirements of the study. The primary requirement for the participants was a diagnosis of affective mood disorder according to the ICD-10. All participants were residing in rural areas or small towns. For each patient, the trial's duration was a minimum of six and a maximum of eight weeks. During the study, the application was recording 24 hours a day, seven days a week without stop.

The short self-assessment questionnaire was set to appear automatically every day at 8 pm, and the participants were asked to fill them in as frequently as possible. Frequent check-up appointments (every 2-3 weeks) were used to transfer the collected data to the study server.

3.4.5 Data and Participant Overview

In the following section, an overview of the recorded sensor data and the study participants is given. One of the patients was really motivated to participate, yet quickly realized that the study turned out to be too challenging. Thus this patient drops out after a few weeks.

Another resident had to be excluded due to smart-phone failure, that led to massive data losses. A third patient did not improve their state during the study and eventually stopped to contact the hospital. Thus the data available until the point of drop out, was not sufficient for a reliable analysis. As the time resources of the psychiatric hospital and the time available for deploying the study were limited, the loss of data could not be compensated by including more patients. Nevertheless, and even though data gaps could not entirely be avoided for all patients, out of 10 patients, the data-sets of 6 patients (P5-P10) were sufficient and were included in the analyses described in the following parts. All of the participants that could be included in the analysis had clear changes in their episodes during the study. Figure §3.2 graphically shows the state-progress of each patient during the study period. However, the degree of changes, as well as the amount of data that was collected varies greatly:

Four out of the six participants (P5-P8) started the study in a more or less severe depressive episode (see figure §3.2). All of them show improvement in the state during the study and were eventually be tested as not depressive at the end of their participation. Patient P6 alone is partially an exception of this, as their state worsened during the first half, yet eventually also showed progression up to only slight depression at the study-end. Two patients (P9 and P10) started in a more or less severe manic episode (see figure §3.2). At the end of the study, both were tested normal within the capabilities of the scale tests. Still, P9 already showed some tendencies towards depression, while P10's appointment assessment discussions with the psychiatrist hinted some mood swings in-between. Those nevertheless were could not be covered by the scale tests. The data of a seventh participant (P3) was almost complete. Still, P3 started in a mixed state. This means the behavior of the patient was showing both depressive and manic characteristics at the same time. This cluster of symptoms is an exciting case of course. However, this patient was the only participants with this form of symptoms. Thus more data of similar subjects would have been needed to draw reliable conclusions. Hence this seventh patient was also excluded from the analysis.

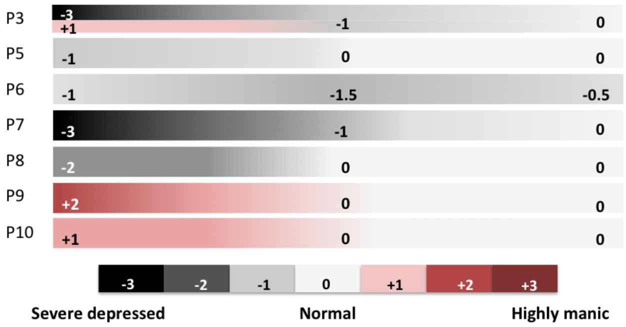


Figure 3.2: The Progression of patients (P3-P10) from depression (-3 .. -1) or mania (+3 .. +1) or mixed phase (- and +) to normal (0).

3.4.6 Evaluation and Initial Results

The aim of the do-ability analysis was mainly to understand whether sensor data collected by a patient's smart-phone would comprise disease-relevant information. Since such evaluations have not been done before, at the beginning of this evaluation, it was not clear which kind of information the sensor data would provide. Thus three evaluations were performed stepwise:

1. A qualitative analysis of visible patterns.
2. Then, a quantitative analysis of trends in the data in relation to the diagnosis.
3. Finally, a correlation analysis of sensor data and self-assessment was performed.

Qualitative Analysis

Note that this evaluation was intended to understand whether sensor data was a sufficient way to analyze affective disorders. Thus this evaluation does not intend to be extensive or complete. Also, the amount of collected data and the kind of sensor-platform used would enable a variety of possible features. Still, as the main aim was to get a first overview of the data's potential and thus the analysis and features are meant to focus on only one aspect for each of the disease-relevant behavioral patterns:

Location - Map of Movement: One outcome that the psychiatrists participating in the study were keen on was the possibility of getting a good visual overview of the patients' life, specifically during the weeks since their last appointment. An example of visual illustration they actively suggested was, what they called "a map of movement." This map should visibly show the patients' paths over some days or weeks. Figure §3.3 provides an example of such a map of movement. Please note, for privacy reasons, GPS data was anonymized before processing! The left part of Figure §3.3 displays the paths of movement for the first few days of a participant in a severely depressed state. It apparently only provides movement around the home location (0,0). The right part, in contrast, shows the paths of the same resident for one week after the state had improved. The distances, this participant has covered in a day, is undoubtedly longer.

Social Interaction - Phone Calls: An interesting and, at first even for the psychiatrists, surprising effect was that most patients had an increase in the length of phone calls and the number of calls when they were in a mildly depressive episode. This was even true in both directions. No matter if a patient was severely depressed before and increasing, which seems to be comprehensible, but also when a patient was neutral and moving towards depression. For both situations, the length of phone calls increased. On a second thought though, this effect seems to confirm the theory outlined above that mildly depressive people exhibit an increased desire to talk about their feelings. Additionally, friends and relatives could realize that a change was happening and thus be expressing their concerns by contacting the patient more often.

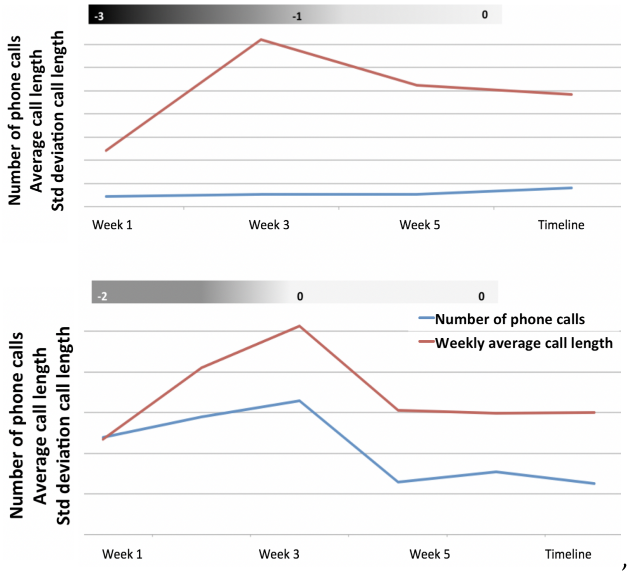


Figure 3.4: Number (blue) and average length (red) of phone calls and state profile (bar at top -3..0) of two Patients

Even though Figure §3.4 shows this trend as an example only for two patients this effect could be observed within 66% of the study participants. Both plots show an increase in the average length of phone calls with the improvement towards a mild depression. However, with further improvement into a none depressive condition, the phone call length levels off and stabilizes at a lower length. The leveled duration of phone calls in a normal state appears to be higher than in a more severely depressed state. During the normal phases, the average length remains constant. This evaluation might in a later state be the basis of the targeted diagnosis and episode prediction of depression and mania.

Trends in Relation to Diagnosis

Since the qualitative and visual analysis had revealed obvious correlations between the diagnosed state and pattern in the data, these trends in relation to the diagnosis should be evaluated. This analysis is of a qualitative nature because of both, the number of examinations per patient and the number of patients is not sufficient for statistical analysis. Nevertheless, this analysis is intended to shed light on the extent to which the effects of the patient's condition on the sensor data can be measured.

Activity - Motion Ratio: The level of activity is expressed mainly in the amount of movement. Therefore a simple and prominent feature is to calculate the ratio of time spent moving vs. time spent resting, based on acceleration. Table §3.1 shows the change in motion ratio for depressive (P5-P8) and manic (P9-P10) patients.
Comparing the values for the patients during a depressive and an increased state shows that the motion ratio increases from an average of 11.3% motion in a depressed state to 13.71% motion in an increases state. This is an increase of 21.3% for the improved state for all patients, except for P7. Even-though the effect in P7 seems to be reverse to the others, a discussion with the psychiatrists explains this effect: P7 was severely depressed at the beginning of the study, and thus, being rather young, was inclined to use the new trendy smart-phone as a quite welcome distraction. Therefore, P7 played with their new phone a lot during the first weeks of the study. Later in the study, when the state of P7 improved, the amount of phone usage leveled off to a typical amount. Removing P7 from this evaluation would increase the improvement of motion among depressive patients to 36.2%.

Not surprisingly for the manic patients, the trend is reversed. The amount of motion decreases with the progression from manic to a normal state. While on average the manic patients have a motion ration of 6% during mania, they move 4% of their day in a normal episode, a decrease of over 33%. An average amount of motion of 11% during a depressive state in comparison to an average amount of motion of 6% during a manic phase seems somewhat odd and surprising. Nevertheless, this effect has two different explanations. First, all patients were already under psychiatric treatment. This means, manic patients were told to actively try to reduce their drive to move, while depressive patients were told to increase their movement actively.
Secondly, this effect merely confirms the statement from before, saying that the characteristic of each feature is, in general, person dependent and possibly even unique. For example, the increase in depressed patients varies between 8.9% for P8 and 64.9% for P6. This means that P6 very likely express their state more through physical activity than P8. Still, even though the effects between patients vary quite a lot, there is measurable (and in trend for most patients similar) difference in the percentage of movement between bad and good state. This

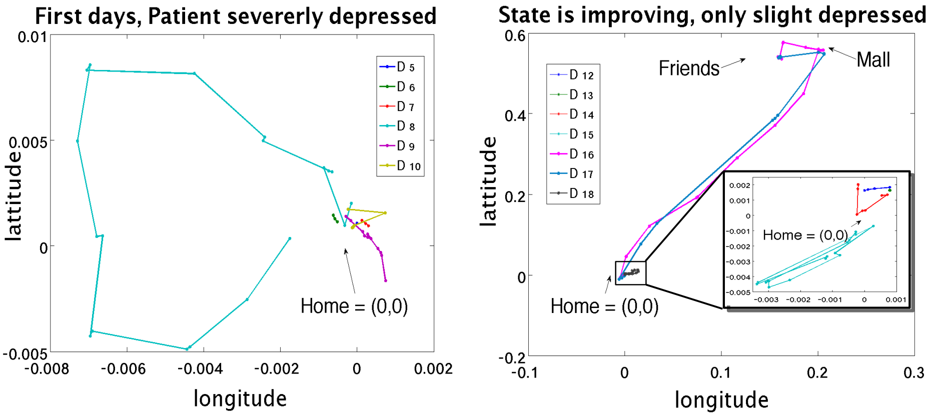


Figure 3.3: Map of movement: left - first days, severe depressed (-3), only movement around home: right - after improvement of state (-1), movement around home and first visit to friends

trend clearly indicates that the sensor data can depict effects on behavior. More precisely, within the limits of the available data, this output confirms that, with improving condition manic patients become quieter while depressed patients become more active.

Patient	depressive state	improved state	increase
P5 (scale)	13.04% (-1)	20.19% (0)	54.8%
P6 (scale)	7.32% (-1)	12.07% (-0.5)	64.9%
P7 (scale)	8.21% (-3)	4.48% (0)	-45.5%
P8 (scale)	16.62% (-2)	18.12% (0)	8.9%
mean	**11.30%**	**13.71%**	**21.3%**
Patient	manic state	improved state	decrease
P9 (scale)	3.38% (+2)	2.07% (0)	38.6%
P10 (scale)	8.71% (+1)	5.95% (0)	31.7%
mean	**6.05%**	**4.01%**	**33.7%**

Table 3.1: Increase/decrease of the percentage movement from depressive (scale: -3 to -1) or manic (scale: +2 to +1) state to improved state (scale: 0)

Location - Outdoor Ratio Depressive people are inclined to withdraw to their homes and stay inside. Thus, a first aspect of analyzing location data is the ratio of time spent indoors and outdoors. This outdoor ratio was calculated based on the amount of GPS data recorded during the day. For calculating this ratio, it was assumed that GPS only would be visible to the smart-phone sensor if the patient were outdoors. Given this assumption, the ratio can be measured easily by calculating the time the GPS signal was visible. Invisibility of GPS of more than 10 minutes was presumed not to be outdoors. For the evaluation of the outdoor rate, the daily outdoor rate was averaged over the course of approximately ten days within a particular state and over ten days within the enhanced condition.

Please, note that some patients had occasionally switched of GPS measurement off. Also not every patient provided data of 10 or more days within a specific state (some changed more rapidly). Thus, not for every patient the amount of data available was the same. All four patients that provided sufficient GPS data were depressed. For those patients, the time spent outside averages at 4.12% in a depressive episode but enhanced to an average of 12.88% in a normal state. This means an increase of 200% (between 35% and over 2700%). The range in the increase, of course, is partially due to the amount of improvement of state (e.g., severe depression "-3" to normal "0" in comparison to slightly depressed "-1" to almost normal "-0.5") but also highlights the individuality and patient-specificity of the numbers. See also Table §3.2 for details. Nevertheless, despite all individuality, this analysis could once more confirm the usability of a smart-phone to observe a change in state for affective mood disorder patients.

Patient	start (depress.)	end (improved)	increase
P5 (scale)	5.09% (-1)	12.30% (0)	140.5%
P6 (scale)	7.61% (-1)	10.32% (-0.5)	35.5%
P7 (scale)	0.698% (-3)	19.94% (0)	2759.3%
P8 (scale)	3.75% (-2)	8.98% (0)	138.9%
mean	**4.29%**	**12.88%**	**200%**

Table 3.2: Increase of time spend outside from start in depressive state (scale: -3...-1) to end in improved state (scale: 0)

3.4.7 Correlation of Sensor Readings and Self-Assessment

As has been reported before, the self-assessment of the participant, was filled-in on a daily basis, which was not possible for any form of objective assessment. Thus, the self-assessment was providing a sufficient amount of data points for statistical analysis. Still, self-rating is not objective and tends to be influenced by various factors and thus being biased and noisy. To reduced the bias effect a 7-day sliding window was applied onto the self-assessment to smooth and average the values. With theses smoothed numbers, the correlation (linear regression) of the features extracted from the sensor data (see previous sections) and smoothed self-assessment as ground truth was calculated. To verify the statistical significance of the resulting correlation a t-test of the regression was performed.

Table §3.3 lists the results of this correlation and the t-test evaluation. For all three behavioral aspects and all patients (if data of a specific aspect is available) the t-

	Activity weekly			Location weekly			PhoneCalls weekly		
Patient	T-Value	Correlation	DoF	T-Value	Correlation	DoF	T-Value	Correlation	DoF
P5	2.3918	0.36	40	2.6754	0.47	27	-7.0983	-0.82	42
P6	2.2119	0.43	24	-1.7837	-0.39	20	2.0608	0.32	20
P7	2.1273	0.35	35	2.4808	0.55	16	-0.0694	-0.02	36
P8	-8.1330	-0.87	23	-3.2647	-0.78	9	3.3761	0.62	32
P9	-0.9237	-0.2	21	-	-		4.1455	0.55	37
P10	-2.4512	-4.48	13	-	-		6.0665	0.66	41
90% C	$\lvert t\rvert > 1.79$			$\lvert t\rvert > 1.73$			$\lvert t\rvert > 1.79$		

Table 3.3: t-Value, correlation and degree of freedom (DoF) of the three features with the self-assessment

value, the correlation value and the degree of freedom (number of data instances) are listed.
In the bottom line, the table states for each feature, for the smallest degree of freedom.[1], the critical value for T to be still within the 90% confidence interval. Both t-value and correlation value show a correlation of the activity feature for all patients except for P9, who seems not to express their well-being via activity. For all other patients, the correlation is within the 90% (C) confidence level. The movement feature (location ratio) correlates for all patients within the 90%. Note, P9 and P10 did not provide enough GPS points for a statistically valid analysis, so they were left out in this analysis. The phone calls feature correlates for all but P7.

3.4.8 Discussion, Insights and Further Recommendations

This evaluation has introduced a study which served as the first analysis towards the development of a smart-phone based support system for bipolar disorder patients. The study included the collection of behavioral data from ten bipolar patients, using a commonly available smart-phone. Additionally, despite being a real-life deployment, an elaborate ground-truth by psychiatrists and psychologists was collected. The sample size of this first study is small, mainly due to due to the exploratory nature of it, as the general goal of this study was to evaluate whether such sensor data would include disease-relevant information.
Despite the small sample size, it was possible to perform a step-wise analysis including quantitative and qualitative analysis and a statistically significant evaluation of correlation. Summarizing the results of this do-ability evaluation, it is safe to say that it is possible to statistically prove a correlation between the reported state and the recorded sensor data. Of course, not all features work entirely or equally for all patients. This is not surprising, since humans are individuals and will deal with situations and states in their unique ways.

For example, not every person starts to make long phone calls when they feel sad. Also, everyone has their own unique level of activity. So the same value could mean manic for one person and depressive for another. In the same line of argumentation, a notable increase in a number cannot be mapped onto a fixed improvement-scheme for everybody. Moreover, it is very likely still, that such an increase cannot even be mapped onto the same person without further evaluation. Nevertheless, this evaluation very clearly provides an answer to the question stated at the beginning of this analysis, that sensor data indeed carry information that can be directly correlated with the patient's state.
As being a successful first evaluation, further work will continue and improve this work on a multitude of aspects: Based on this results, a follow-up study will be performed. This study will not only include an improved application but will also be deployed on a long-term, at least twice as long. All the same, the second study aims to include the double amount of patients.

An aspect, not explicitly evaluated in this section, but none the less affecting the options for evaluation was the longtime distance between objective ground-truth points. Psychiatric assessments and psychological scale tests should not be scheduled more often than every three weeks to avoid learning effects. If possible, the second study should include more ground-truth points. A possible way to achieving this could be in the form of frequent short phone calls by the psychiatrists in-between the ‚on-site assessment appointments. This way the amount of available ground truth could be doubled. Also, the self-assessment can be improved. The analysis has shown that questions like "at what places have you been today" is interesting for the treating psychiatrist, yet in the way, the data can be evaluated this information is too coarse to suffice as ground-truth. Thus the questions in the self-assessment should focus more on evaluating the patient's (self-rated) state. E.g., the daily questionnaire could be comprised of randomly selected questions out of the mania and depression catalog.
The last recommendation for future studies and analysis is to enlarge the feature set. Not only to evaluate which features work best for most patients but also to evaluate the individuality of features. The individuality of features means, to evaluate whether there are features that work for every person, or features that only work for some. Most importantly if for each person features can be extracted that work.

[1] Please note that this means, for a higher degree of freedom the t-value may be higher than the value listed in the bottom line of table §3.3 to be still within the 90 % confidence interval

3.5 The Vision - Activity Recognition Assisting Mental Care

The previous section has confirmed that smart-phone internal sensors can record disease-relevant features of affective mood disorders. In this second phase, the analysis goes more in-depth. Based on the insights of the do-ability analysis a practical and utilizable collaboration of activity recognition and mental care should be designed. The main aim of this collaboration should be to develop a smart-phone application that can monitor the behavior and activity of mentally ill persons objectively!

To specify objective monitoring: the application should be able to determine the state of an affective mood disorder patient but even more importantly should specifically detect changes in the patient's state. This, on a long run, should further enable the users and their doctors and therapists, to provide timely treatment. Additionally, as was hinted with the word "objective," the application should be able to rely only on sensor data and not on any user input or feedback, which in any form is subjective and often biased. Furthermore, the requirement of providing frequent input to a system can have negative impact on the compliance of a user. Thus, some aspects for the envisioned system should first be taken into considerations:

1. The system does not aim for automatically triggering any responses that could mean to detour a doctor, like suggesting to adapt the medication, which could have a potentially harmful outcome, but generally providing information in time and at the right time. Thus, the accuracy of the envisioned system and necessary algorithms can lie within a normal range.
2. Within affective mood disorders a condition called "rapid rapid cycler", which describes patients whose state will change within hours (oscillating between manic and depressive within one day) exists. Such a manifestation is not included in the targeted population of users. On the contrary, the major population of affective mood disorder patients experiences episode changes of 4 or less per year. This means that the reaction time for the envisioned system would be on a scale of a few days not hours.

These considerations generally mean that the system is not envisioned to substitute for a doctor or therapist, which would require very high and robust accuracy and would raise extensive ethical discussions. For the envisioned usage, the recognition should be accurate enough to be able to provide daily updates and long-term overviews to the doctors and the patients. These updates and overviews should, in turn, enable doctors and patients to interpret the current situation better. The required scale can be anything from a few days to weeks or months. Specifically essential and hence mainly desired by doctors will be trend-overviews in between appointments and the possibility to detect changes in state at onset, meaning within very few days. In context of the work in this chapter, this generally means:

1. that a robust change detection is more important than the recognition of a particular state. The system is not required to provide a detailed and correct diagnosis, but to show a trend. The diagnosis will be made by the doctor anyways, which is not only an ethical but mainly a liability question.
2. This also means that a sufficient recognition accuracy does not have to be perfect but can lie in standard ranges of 70-80% to still be useful.
3. Moreover, more important than perfect recognition results is the usability of the system and the ability to obtain fair results in the context of a realistic real-world application. This, in turn, points out that the system has to be able to deal with common effects like losing data due to empty batteries and still provide useful results.
4. Last but certainly not least, the system has to work with genuine patients, that are neither tech-savvy nor able to handle complex technical requirements. Ideally the system should be plug-and-play without requiring interaction other than typical for apps.

3.6 Long-Term Data Acquisition with Bipolar Patients

As has been mentioned, to develop a system as outlined above, data recorded under laboratory conditions would not suffice. Thus a real-life study with bipolar patients was set up as a medical trial in a psychiatric hospital with the approval of the local ethics board of the Innsbruck University Hospital. The trial was an uncontrolled (no control group), not randomized (no control-group thus no randomization), observational study with an aim at recruiting between 10 and 15 patients for a minimum of a 12-week participation per person. Equal to the do-ability evaluation, the resources of the hospital limited the number of possible participants: Not only the participating psychiatrists had to provide the time to assess a number of patients additionally to their everyday work, but also the resources for performing the scale test, which required certified programs were limited and had to be shared with the regular hospital operations.

3.6.1 Monitoring System

Similar to the do-ability study the used system for collecting data was an Android based smart-phone application [2]. The app started automatically on device boot and

[2] implemented by Jens Weppner and Amir Muaremi within the Monarca EU FP7 project

ran almost invisible to the users and requiring almost no interaction in the smart phone's background. To guarantee privacy and the patient's sovereignty about their data it was possible to switch off sensors. Again like in the do-ability study the main relevant indicators for affective mood disorders should be covered. Location and movement were once again covered by anonymized location tracking. Activity was recorded in the form of three-axis acceleration. To assess the social state, again in-going and outgoing phone calls were monitored. To put more focus on the way the patients were interacting, anonymized voice analysis was added during phone calls. Again all recorded data was encrypted (AES 128bit software side encryption), was stored on an SD card, and was transferred to the study data storage during the patient's checks at the hospital.

3.6.2 Study Participants

In line of requirements by the ethics board the inclusion criteria were as follows:

- age between 18 and 65, generally meaning no underage persons, with a flexible upper border.
- ability and willingness to deal with smart-phones.
- being "contractually capable," so no person was included that by law was not capable of deciding for themselves.
- and already having received a diagnosis of bipolar disorder categorized by ICD-10, F31 (by the International Classification of Diseases), with frequently changing episodes.

Of course, the participation in the study was voluntary and neither participating nor quitting would be allowed to affect the therapy. A total of 12 patients could be recruited in the period from November 2012 to August 2013. 11 participants were female, and only 1 was male. They ranged between the age of 25 and 65. Like in the do-ability evaluation also here the majority of participants was female. This does not necessarily mean that more female persons have a mental disorder, but that women more readily accept the fact that they need help. Similar to in the do-ability evaluation, the selection of patients was entirely up to the ward's psychiatrists and their perception of which patient was capable of dealing with the study requirements.

Two patients dropped out early (p0202 and p0602), two patients (p0502 and p0802) even extended the trial length due to different reasons. The majority of the participants started the study in a more or less severe depressive episode. Five of them were in a very severe depressive state and thus had to be handled with specific sensitivity. Three of the participants started during a manic phase, each one in a different state of mania.
All of the study participants (except for p0202 and p0402) underwent one or more clear changes in their mental state during the study. Changes per patients were between one an three. Most of them were progressions from one of the seven possible states (-3 severe depression, -2 depression, - 1 slight depression, 0 neutral, +1 slight mania, +2 mania, + 3 severe mania) to an adjacent state. This means a change was considered to be happening if a severe depressed (-3) patient progressed to depressed (-2). Thus overall 25 state changes were recorded. Figure §3.5 provides an overview of the progression of each patient. Note, since patients p0202 and p0402 did not show any changes in state during the entire study, both were excluded from the data-set.

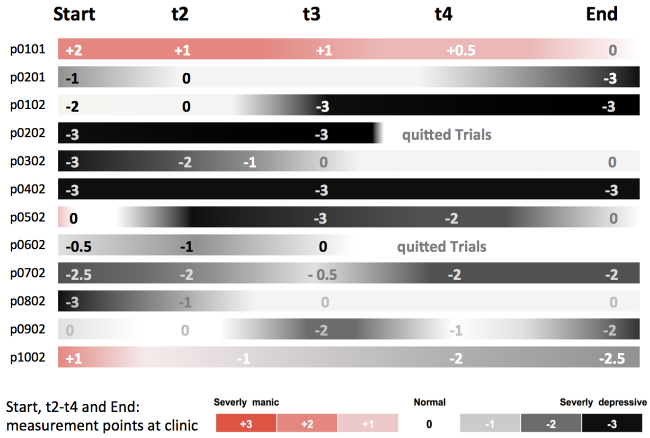


Figure 3.5: The Progression from depression (-3 .. -1)/mania (+3 .. +1)/mixed phase (- and +) to normal (0).

3.6.3 Study Implementation

Each patient was given a new Android-based smartphone running the data-logging application. The reason to give out new phones was that numerous Android-based phones already existed and it could not be guaranteed that the application would work for all of them. Using a phone version that was tested with the application should limit issues. Of course, this required the patients to adjust to a new phone, which was possible to handle by offering the participants extensive one on one time to transfer contacts and set up the new phone according to the participants' liking's.
The data recording application was designed to record all required sensor data fully automatically in the phone's background. It did not require any interaction or maintenance by the user. Due to privacy reasons, the applications even allowed the participant to refuse the usage of data for this particular day; otherwise the data was stored on the SD card. If the patient did not agree to store data of a particular day, all data collected during that day would be deleted. This protocol was implemented to fulfill a precondition for the approval of the ethics board. Anyway, during the entire trial, there was no case of a patient asking to delete data. Still, the mere option to do so was appreciated as a mechanism to feel in control. Nevertheless, most stated that, as they had agreed to participate in a study, they would not withdraw from participating on a "bad day".

The data stored on the phones SD card was transferred to the study server only during the periodic examinations. In the context of this real-live data collection study, the option via SD card and physical data transfer allowed to reduce the amount of necessary security that any wireless data transfer would have imposed. Thus, this method was sufficient and more effective than a wireless transfer with extended security requirements. All data was anonymized before processing in order to hide the patients' identity. In order to guarantee a smooth execution the trial proceeded as follows for each patient:

1. Each patient was recruited during stationary treatment at the clinic. The recruitment was done by a ward-psychiatrist, but once the patients generally agreed to have a look at the study, a person of the study personnel was introduced to the patient. Note, that stationary treatment is not equal to "lock up" treatment. Patients were not included as long as they had to stay in restricted areas. Study patients would stay in the hospital overnight (specifically during the beginning of the study) and attend various hospital therapy sessions. However, otherwise they were free to move around in the hospital compound and also in the town close by.
2. The trial started with an initial examination, once the patient had signed the informed consent. After this examination, the smart-phone was given to patients, and the collection began.
3. Each Patient was released when it was medically advisable (again, the study itself would not affect the treatment). Most participants were discharged one or two weeks after the start of the study. Still, once left, they would come back to the hospital for outpatient examination every three weeks.
4. At the end of the study, meaning after 12 weeks or later, a final examination was performed.

3.6.4 Ground-Truth:

Also similar to the do-ability evaluation the ground-truth was a collection in a detailed subjective and a coarse but objective method. As a measure of objective ground-truth, psychological state examinations (psychological standard scale tests) were performed. These were comprised of 4 standardized scale-tests, two foreign-rated and two self-rated. All of them were performed under the supervision of trained psychologists. Thus, even the self-rating scale tests can be presumed to be somewhat objective. The psychological state examination included:

- The Hamilton Depression Scale (HAMD)
- The Common Depression Scale (ADS)
- The Young Mania Rating Scale (YRMS)
- The Mania Self-Rating Scale (MSS)

Even though each scale examination has its rating scheme, in order to make them comparable to any other assessment, the psychologists were asked to transfer the results of the scale tests into the 7 part -3 (heavily depressed) to + 3 (heavily manic) scheme, as used before. Intermediate steps were: depressed (-2), slightly depressed (-1), normal (0), slightly manic (+1), and manic (+2). If it was better suited to describe the actual state of the patient, half grades were also allowed. Again, the likely memory effect prevented a more frequent execution than ever 3 weeks. Nevertheless, as suggested after the do-ability analysis, in order to shorten the time-span between measurements, specially trained psychologists agreed to talk to the patients over the phone and afterward rate their state according to the scheme. The same way as in the do-ability evaluation, the patients were asked to fill in a short questionnaire on their phone every evening as a subjective ground-truth backup.

3.6.5 Data-Set and Data-Quality

With 12 patients and 12+ weeks of 24/7 data collection, in theory, there should be more than 1000 days of data available. This seems to be a sufficient amount of data. Unfortunately, in the reality of a real-life study, many factors would influence the data collection and the amount of data at hand. Such factors were:

1. **First and foremost the patient adherence:** The study was conducted in uncontrolled conditions during the participants' normal lives. Hence, there was practically no way to guarantee the desired usage. Even when participants were explicitly asked to pay attention to charge the phone and carry it at any time, this was not happening all the time. Additionally, some patients grew a habit of switching off some sensors at certain occasions, partially to save power partially due to a heightened sensitivity to potential surveillance. After all, one goal of this work is to develop a system that would be applicable for ever day use and thus has to be able to handle such actions. Hence, the participants were told to use the phone as they would use it outside of the study.
2. **The evolution of the patient's state:** As a matter of fact, it is more likely to find patients during a depressive episode in the psychiatric hospital than in a manic episode. This is because a depression oppresses a higher factor of suffering, while during mania patients do not necessarily feel that something is wrong, on the contrary. Thus in our data (since participants were recruited during in-ward stay), there are many more depressive episodes covered. Additionally, it was not possible to predict in which individual direction states would progress. So some of the participants starting in a manic phase would move into depression while being in the study, but none of the initially depressive patients was manic at any point during the study.
3. **Availability of ground-truth:** Collecting sufficient ground-truth is one of the significant issues in a health-care real-life deployment. Objective ground-

truth, as already mentioned, only was available around the examination days. In order to extend these days, after consulting the participating psychiatrists, it was decided to additionally include a few days before the examination and a few days after, and assign them the ground-truth assessed during the examination.

In addition to the anyways limited amount of data with objective ground-truth, for a few patients, this number of available days had to be shortened even more, as partially unstable self-assessment did not allow trusting these days.

Not enough a third factor limited the amount of useful day. This factor is the availability of sensor data. For example, if examination based ground-truth was available and the self-assessment did not indicate any instabilities, but the GPS was turned off, this day was not usable.

These three factors: adherence, evolution of state, and availability of ground-truth had a limiting effect on the amount of useful sensor data. Table §3.4 provides details about the amount of useful data-points per patient. The amount days with data available can be seen in column two and ranges between 53 and 131 days. These numbers come close to the actual anticipated amount of data (84 days within 12 weeks).

The two following columns list the number of days of available GPS and Acceleration data. However, to get a picture of the usable data, the last six columns are of most interest. These last six columns depict the number of days where GPS data plus ground-truth (GT), Acceleration data plus GT, Phone behavior data plus GT, Sound data plus GT, and GT plus either one of GPS OR Acceleration or GT plus both GPS AND Acceleration were available. While the amount of usable Acceleration data is fair and above 40, and up to 70 (except for P0206. since this patient dropped out early), the available usable data for GPS drops to below 40 for most with an extreme of 19 for one patient and a highlight of 51 for one single participant. Table §3.4 also provides the distribution of the available data-points onto the different classes (in brackets) in the last six columns. A few of them only comprise 7 or fewer days of data, which is likely not enough. Still, for most cases the available data is reasonable for using standard techniques to evaluate them.

atients	# of Ground Truth Days (GT)	GPS +GT	ACC + GT	PHONE + GT	SOUND + GT	GT + ACC OR GPS	GT + ACC AND GPS
p0101	84	**26** (3/20/3)	**71** (32/27/12)	-	-	**71** (32/27/12)	23 (3/20/3)
p0201	47	**36** (10/26)	**38** (12/26)	**38** (12/26)	**38** (12/26)	**38** (12/26)	**36** (10/26)
p0102	52	**34** (21/13)	**46** (33/13)	**46** (33/13)	**46** (33/13)	**46** (33/13)	**34** (21/13)
p0302	70	**51** (11/40)	**60** (18/42)	**61** (19/42)	**61** (19/42)	**60** (18/42)	**51** (11/40)
p0502	63	**28** (5/23)	**58** (14/14/30)	-	-	**58** (14/14/30)	**23** (0/23)
p0602	41	**31** (12/19)	**21** (11/10)	**35** (13/22)	**35** (13/22)	**35** (13/22)	**17** (10/7)
p0702	53	**31** (24/7)	42 (34/8)	**47** (35/12)	**47** (35/12)	42 (34/8)	**31** (24/(7)
p0802	71	**37** (7/30)	**62** (16/46)	-	-	**62** (16/46)	**37** (7/30)
p0902	48	**41** (26/15)	**41** (26/15)	**41** (26/15)	**41** (26/15)	**41** (26/15)	**41** (26/15)
p1002	47	**19** (11/9)	**40** (29/11)	42 (31/11)	42 (31/11)	**40** (29/11)	**19** (11/9)
average	**57.6**	**33.2**	**47.4**	**44.3**	**44.3**	**49.4**	**31.8**
StD	13.76	8.76	14.72	8.48	8.48	12.36	19.73

Table 3.4: The amount of sensor data (in days) and Ground Truth days (GT) per patient (in brackets: distribution over classes).

3.7 Data Pre-Processing and Feature Extraction

The do-ability assessment has already provided important insights into the types of data and features that would probably work best. As a reminder, talking with psychiatric staff before the do-ability evaluation lead to defining three main aspects of patient behavior that should provide the essential and desired information. These main aspects were:

1. **Physical motion:** Patients with depression have a reduces drive and thus tend to move less, slower, and with reduced force, and vice versa for manic persons.
2. **Travel patterns:** Depending on profession and personal preferences, still most people have a pattern in their daily life during weekdays. Some places are dominant, like the home, the workplace, the grocery store, the gym maybe, and so on. In a normal state, these patterns will, with some divergences, be re-accruing. For both manic and depressive state, this pattern will change, e.g., become less frequent or more erratic.
3. **Social interaction:** How a person acts in the vicinity of other persons, and how a person interacts with other people is to some extent unique to everyone. However, as person specific social interaction might be, what every person has in common is reduced desire and reduced ability to interact with others (again in a person-specific extend) in a depressive phase and heightened in a manic phase.

It should be highlighted once again that the actual expression of these behavioral aspects is very person depended. During the study, one psychiatrist said:

If I have hundred bipolar patients, I will have to deal with a hundred unique forms of bipolar disorder!

This means that a particular value of a parameter can mean depression for one person and mania for another.

3.7.1 Acceleration Features

Before further processing, the raw 3-axial acceleration data had to be re-sampled to a fixed sampling frequency. This was necessary due to effects caused by the Android operating system. Since the orientation of acceleration is unknown, dependencies were removed by calculating the magnitude of the three data-streams. Any further calculation was performed on this magnitude signal.
As a next step, all parts of the magnitude stream that were around the gravitational force (9.81) were removed. This is feasible since the magnitude of a three-axis accelerometer is equal to the gravitation (9.81) when not moved, and thus all data-points in the magnitude stream close to the gravitation depicted that the phone is worn not on the body of the patient, but lying on a flat unmoved surface. After pre-processing was done, the following features were calculated and then used for classification on a daily basis:

- RMS
- RMS mean
- freq. centroid (fc)
- fc mean
- freq. fluctuation (ff)
- ff mean

3.7.2 Location Features

One pre-condition for the approval of the ethic's board was that the GPS traces recorded by the phone were to be anonymized. This way it should not be possible to map GPS coordinates onto meaningful locations and thus identify a particular person. For fulfilling this precondition, GPS coordinates were translated into a different coordinate system with the zero-point (0.0) assigned to the pair of coordinates that occurred most often. This zero-point was defined as "home". Since this way no semantic evaluation of the GPS traces was possible, only abstract features were calculated:

1. The number of locations visited (defined as a cloud of GPS points within 500m meters)
2. The number of how many hours per day the patient was outdoors (meaning during which hours GPS was visible at least once)
3. The average time a patient spends outdoors per hour
4. Distribution of being outdoors per day (averaging the subset of hours spent outside - e.g., in the morning or afternoon).
5. The variance of the times spent outdoors
6. The number of single stays outdoors (a stays outdoors = a consecutive number of GPS data points with no pause in between of 15 minutes or more. Timestamps of more than 15 minutes difference were marked a new stay.)
7. Ratio of time outside in 24 hours (the sum of the duration of all connected stays, divided by 24 hours.)
8. The distance traveled (sum of all distances traveled on any particular day.)

3.7.3 Sound and PhoneCall Features

To some extend the usage of a smart-phone by its nature can already be characterized as an intention for social interaction. For example, the traditional mobile phone activities as making a phone call or send text messages have all an apparent social component. A second option to leverage the phone is to analyze sound. Voice recognition itself is of course not an option due to privacy, but analyzing the frequency of voice, changes in pitch, or talking speed could provide valuable insight.

Phone Call Features:
Many Phone-Call-behavior features were extracted from the usage of the phone. These include the length of phone calls, whether phone-calls were incoming or outgoing and which caller ID numbers were involved (due to privacy, the numbers were anonymized, and only the last four digits were stored). Eventually, the following features were extracted:

- Number of phone calls
- Total length of calls (sum of call-length per day)
- Average length of phone calls
- Standard deviation of the length of phone calls
- Number of unique numbers

Sound Features
The sound features [3] were extracted by Muaremi [123] and [124]. Since these extracted sound features were used in this work to perform classification on them, for a better understanding the following paragraphs should summarize the work done by Muaremi. The following text (marked italic) was written by Muaremi and is taken from [123]: *We divide the sound features into speech features which describe the phone call interaction and voice features which are usually used to detect the emotions from the voice.*

Speech Features: *The aim is to understand the dyadic communication of the patient with the other person on the line. Starting from the voice activity detection (voiced speech vs. unvoiced speech), the speaking segments are created.*
Using these segments we can differentiate between turns, short turns, and non-speaking segments. Short turns or utterances are feedback words while someone else is talking, such as "okay", "hm", "right", etc. Non-speaking segments are either pauses or turns of the counterpart (see [125] for more details). The following speech features were then calculated on a daily basis:

[3] Sound Features have been published in * Muaremi. A. (2014), Wearable Sensing of Mental Health and Human Behavior (Doctoral dissertation). * Muaremi A. et al. "Assessing bipolar episodes using speech cues derived from phone calls". Pervasive Computing Paradigms for Mental Health, Springer International Publishing, 2014, 103-114

- *Average speaking length and speaking turn duration*
- *Average number of speaker turns and short turns/utterances*
- *Standard deviation of speaking turn duration*
- *Speaker turns per length in minutes and short turns/utterances per length in minutes*
- *% of speaking from the total conversation*

Voice Features: *The open-source "openSmile" toolbox [126] was used to extract the acoustic features. For each frame of the speech signal, (frame length: 25 ms. step size: 10 ms) different low-level descriptors are calculated: rms frame energy, mel-frequency cepstral coefficients (MFCC) 1-12, pitch frequency F_0, the harmonic-to-noise ratio (HNR), zero-crossing rate (ZCR). Then, functional like mean, standard deviation, extreme values, kurtosis and more were applied on all frames for each descriptor. The resulting feature vector was reduced by using the filter feature selection method. Finally, we end up with the following voice features:*

- *kurtosis energy*
- *mean 2^{nd} and mean 3^{rd} MFCC*
- *mean 4^{th} delta MFCC*
- *max ZCR and mean HNR*
- *std and range F_0*

3.8 Recognizing the State of Bipolar Disorder Patients

After evaluating the do-ability and collecting a large amount of data, the ultimate goal of this chapter was finally pursued. The analysis in the following chapter is structured step-wise. First, not unlike in the first chapter, the recognition of the state is done with a classification for each sensor modality individually. Second, sensor modalities will be fused to provide an overarching classification.

3.8.1 Single Modality Classification

Based on the possible state-scheme resulting from the diagnosis and assessments during the data recording, up to seven classes would, in theory, be possible. See also §3.6. These possible classes are:

- severely depressive
- depressive
- slight depressive
- normal
- slight manic
- manic
- severely manic

The analysis was calculated using the Weka [90] Data Mining Software. For the classification, first, a linear discriminant analysis [89] was applied onto the features to reach the first reduction in dimension. According to conventional practice in supervised learning, a randomly performed 66/33 training/test-data percentage split was applied onto the data-set.
To ensure that the classes were represented equally in both sets, the test-set was re-sampled As a classifier, the Naïve Bayes was utilized to train with the training-set and estimate the classes of the test set. Other possible classifiers were tested, including the k-nearest neighbor, the j48 search tree, and the conjunctive rule learner, but they achieved similar results as the Naïve Bayes. Thus all of them could have been used. Since the Naïve Bayes provides a probability distribution as output, which was supposed to be used for analyses in the next step anyway, as a matter of efficiency, only the Bayesian classifier was used eventually. All steps of the classification were repeated 500 times in cross-validation, each time randomly splitting the entire data-set into the test/training set. This was done to eliminate artifacts and "unlucky" or specifically "lucky" selections. At this stage, the classification was performed on a per patient basis, meaning training and testing on the same patient. At a later stage, the person-independence was evaluated (meaning, training with one patient and testing on another), but showed that with the given amount of data is not reliably achievable.

	Acceleration			**Location**		
Patients	total	Recall	Precision	total	Recall	Precision
p0101	75%	75.7%	76.3%	77%	87.0%	87.0%
p0102	76%	59.0%	61.7%	82%	72.3%	72.3%
p0201	68%	69.3%	70.3%	77%	81.3%	81.3%
p0302	66%	55.6%	57.3%	92%	81.8%	95.5%
p0502	72%	65.7%	68.1%	85%	83.5%	76.2%
p0602	66%	65.9%	67.9%	71%	68.6%	69.3%
p0702	73%	50.4%	50.5%	77%	100%	77.4%
p0802	77%	66.4%	70.2%	89%	76.9%	85.3%
p0902	70%	67.1%	68.3%	85%	85.6%	84.1%
p1002	71%	53.9%	57.1%	79%	79.4%	80.1%
Average	**72%**	**62.9%**	**64.8%**	**82%**	**81.7%**	**80.9%**

Table 3.5: Precision and recall of state recognition of Acceleration and Location

3.8.2 Single Modality Results:

The accuracy of the classification for the individual patients ranges between 66% and 92%, which, given the real-life application, is quite reasonable. The average for all patients shows reasonably good recognition of 81% for location, 75% for acceleration, 70% for sound analysis, and 66% for phone usage. The best results for the classifier are achieved by location and acceleration only. Especially promising are the results of location. It does not only work best in general, but also no patient has a worse accuracy than 71% (p0602) and the best individual accuracy reaching 92% (p0302).
In the acceleration analysis, even seven out of the ten participants have an accuracy of 70% or higher. Location

and acceleration also provide good recall of above 80%, with a precision of above 60%. See Table §3.5. In the Appendix A §7.3, in the section of this chapter §7.3, Table §2 provides a very detailed analysis of recall an precision of location and acceleration for every single patient and the distribution of results over the single classes per patients. Please refer to this table in the Appendix for a detailed look.

This analysis also shows that for all recognition results a low number of data instances for these specific classes is available. Thus lousy performing classes are not necessarily performing bad but may lack sufficient amount of data to provide reliable results. In general, phone usage and sound do not work as well as location or acceleration, but especially phone usage works worst in recognition. With single plain classification only, phone-call behavior provides recognition rates of only 66% accuracy.
The voice-features (sound) work better regarding single modality classification (70%), but still, they do not come close to the results of location for example. Note, eventhough, at the beginning of the study, the patients were assisted in transferring contacts and all relevant information to the study-phone, four participants did not use the study-phone to make phone calls. Therefore, for these patients, phone usage and sound could not be analyzed.

Of course, it has been pointed out earlier that, in the given context, classification rates are not required to be perfect to be useful. Still, location analysis shows that accuracy of around 80% is possible. Nevertheless, the low classification results of phone usage and sound do not mean that these modalities are useless, but as a single classification modality, they do not suffice.

	Phone			Sound		
Patients	total	Recall	Precision	total	Recall	Precision
p0102	62%	52.5%	53.2%	68%	60.8%	62.0%
p0201	75%	64.4%	70.1%	66%	51.2%	50.0%
p0302	71%	62.0%	63.6%	74%	64.5%	52.0%
p0602	36%	33.9%	50.0%	76%	68.5%	78.7%
p0702	72%	89.7%	78.8%	-	-	-
p0902	68%	63.7%	65.4%	71%	68.5%	68.5%
p1002	65%	59.3%	58.8%	65%	54.0%	61.2%
Average	**64%**	**60.8%**	**62.9%**	**70%**	**61.3%**	**62.0%**

Table 3.6: Precision, and recall of state recognition of Phone usage and Sound. (Note, p0101, p0502,and p0802 did not use the phone for calls)

3.8.3 Sensor Fusion

One of the limitations of the single modality classification for some patients and some sensor modalities in some classes is the limited amount of available data. This is precisely true for the location data, as some patients repeatedly switched off location tracking on their phones. Furthermore, some of the modalities did not perform well in a single modality analysis. In this regard, a next step towards optimizing the results is to fuse the sensory modalities. As for the single modality classification the Naive Bayes classifier was used, a list of probabilities for all possible classes, for each day and each modality (for all four modalities acceleration, location, phone usage, and sound) was available. These modalities should now be combined, both to improve accuracy and also to widen the range of days considered (different modalities would probably provide data for different potentially overlapping sets of days) [4].
In supervised learning, it is a fact that a sufficient amount of training data is essential. Therefore, it makes sense to trust a classifier applied on a larger data-set more than one applied on a smaller data-set. Thus, the fusion algorithm includes penalties for the size of respective the data-set. The fusing algorithm works as follows:

- For every day with only one modality available, the most probable class of the respective class probability list is chosen.
- For any other day with more than one data modality available, the class estimates were fused using this algorithm:
  - For each class, the ratio of available training data of all available modalities was compared to all training data.
  - A further penalization considers the amount of training data to down-weight modalities with small numbers. For this the coefficients (of the previous step) are input into a sigmoid weighting function.
  - The weighted coefficients are then multiplied with the estimated class probabilities for each modality.
  - In the last step, all resulting vectors of class estimates are summed, and the highest rated class is chosen as the winner.

Performing the fusion resulted in a final classification for each day data was available. These steps were performed in different combinations. Since location and acceleration worked best in a single analysis, these two were fused first. Afterward, as phone usage and sound were extracted in parallel, these two were also fused in a second step. Finally, the fusion was calculated over all four sensor modalities.

Acceleration and Location Fusion

As has been stated before, the recognition on location data alone works best with an average of over 80%, in overall accuracy and precision and recall. Acceleration alone performs less than location, yet yields an overall accuracy of around 72% with precision and recall above 60%. Still, as has also been pointed out already, acceleration provides distinctly more data than location (on average ACC has 47 data points vs. LOC has 33 data points).

[4] The work in this subsection was joined work with my colleague Gernot Bahle. In hindsight it is not entirely possible to separate again who has done which part of the work. Mr. Bahle has particularly supported my work in this subsection by implementing the sensor-fusion algorithm and running some analysis.

This means that the results of the acceleration data are more trustworthy than of the location data. Hence, if acceleration and location are fused, the average number of data points increases to 49.
This is reflected in the results. Table §3.7 provides the corresponding numbers. Even though the accuracy results of the fused data is with 76% (and precision and recall of above 70%) not as high as for location only, it is still considerably better than for acceleration only. Since the results of the fusion base on a larger data-set than either one of the single modality results, the fusion is in any case also more trustworthy than any of the single modalities classification.

Phone Usage and Sound Fusion

Similar to the result of the single modality classification, the result of the fusion of sound and phone usage is not overly satisfying. Even a fusion of both modalities does not really enhance the results. Indeed, the fusion accuracy is better than the accuracy of the phone usage alone, but not better than sound. To this extent, the effect is similar to the location acceleration fusion, where also the fusion is better than one of the modalities but not the other. Still, results below 70% are not a satisfying result. One tiny positive effect though is that the precision is better for the fusion than in any of the single modality classification and recall is not worse. Nevertheless, for phone usage and sound, not event the fusion improves the results enough to be useful.

All Sensor Fusion

In the previous sections, the single sensor modality partially showed good results for location and acceleration and not so good results for phone usage and sound. The fusion of two sensor modalities could make acceleration and location results more reliable, yet sound and phone classification was not improved. Results have shown that location and acceleration work far better than sound or phone usage. The same is valid for fusing location and acceleration compared to fusing sound and phone usage. Even though the results of location and acceleration are reasonably good, location and acceleration cover only the aspects of movement and activity, but they do not include social behavior. As has been elaborated before, in psychiatric treatment all three aspects are crucial. They are not equally relevant for all patients, as the expression of the state varies within patients, but none of them should be excluded.

Phone usage and Sound, the available sensor streams for the social behavior, do not perform well in single classification or when fused together. This does not necessarily mean that these sensor streams are not useful. In psychology, the social behavior alone also does not necessarily express a mental condition. Social behavior about other aspects though can provide a picture of the actual mental state. The same might be true for the sensor streams. Thus, the question remains, whether phone usage and sound would be able to contribute to the whole picture. Therefore, phone usage and sound were not discarded but added to an all-in fusion of all sensory modalities. Table §3.7 summarizes the result of the all-in fusion. For a better comparison, this table also includes a summary of the single modality results and the two sensor modality fusions.
Even-though it is small still, the table reveals that the all-in fusion accuracy is better than any of the two modality fusions. The all-in fusion even is slightly better than the location and acceleration fusion. The overall accuracy of the all-in fusion is 1.3 percentage points better than Loc&Acc fusion (which is the better performing fusion of the two-modality fusions), and the precision is 0.7 percentage points better than for Location and Acceleration fusion.

Both are considerably better than their equivalents in the Phone&Sound fusion. Specifically recall is also improved in comparison to both two-modality fusions. Regarding the all-in fusion, this actually means that the results are more stable. This indicates the correctness of the assumption that the combination of all three disease relevant aspects social behavior, movement, and activity, would optimize the state detection and thus provide better results than either single sensors classification or two-modality fusions.

(av. # instances)	**Recall** (std)	**Precision** (std)	**total**(std)
ACC (48)	66.7% (7.9)	67.8% (7.8)	71.7% (3.8)
LOC (33)	81.7% (8.6)	80.9% (7.6)	81.7% (6.5)
FUSION A + L (49)	**70.6%** (8.9)	**75.3%** (8.6)	**75.1%** (6.1)
PHONE (43)	60.8% (16)	62.9% (1)	64.2% (13)
SOUND (43)	61.3% (7.3)	62.0% (1.1)	69.8% (4.5)
FUSION P + S (43)	**60.9%** (4.3)	**68.2%** (7.4)	**68.0%** (5.5)
All Sensor FUSION (49)	**73.8%** (11.3)	**76.0%** (4.9)	**76.4%** (4.1)

Table 3.7: Comparison of recognition of different modalities and fusion of different modalities and of all-in fusion

3.9 Detecting Changes in the Mental State

In the previous section, it was demonstrated how to determine states of bipolar patients based on sensor data. Practically speaking though, determining when the state of a bipolar patient is changing would be much more relevant than only determining the state. To recall, bipolar disorder is defined by frequently changing state episodes of mania and depression. Phases without episodes (normal states) pause the episodes. The goal of any treatment is to keep these normal phases as long as possible and episodes, if inevitable, as mild and short as possible.
One factor that makes treatment so tricky and intensive is the fact that the earlier a change towards an episode is recognized, and thus the earlier medication can be applied or adapted, the better the episode can be handled.

If an episode has already manifested, the treatment is long-term and associated with heavy medication. Unfortunately, determining the onset of an episode is difficult even for an experienced patient. In this line of thought, the detection of state changes out of sensor data has a higher relevance than the actual recognition of a specific state. A few considerations regarding state change detection should be taken to heart, before going further:

1. Considering the application of state change detection, it means to promote a timely visit to the doctor as a key functionality. The patient should be enabled to understand that a change is starting and thus go to the doctor. The application cannot intend to diagnose the patient's state, as from an ethical point of view, the diagnosis will be made by the doctor only.
2. If possible, the state change detection should not require a specific state to start with. It might limit the usability if a patient has to be in a defined state before being able to use the state change detection. In this regard, it should not be necessary to understand in which state a patient is in at the moment for a change detection to work. On the contrary, it should be working from any given start-state[5].

Keeping these considerations above in mind, the following relatively simple and almost intuitive method was developed to determine the onset of change:

1. First, built a model of the given state out of the available sensor data recorded with the phone.
2. Then, once this model is stable (app. after 1-2 weeks), compare each new data point to the model.
3. Data-points within a certain threshold will be classified as "same state" and be used to enhance the model.
4. All other points falling outside of the defined threshold are classified as "change".

3.9.1 Single Modality Change Detection

Following the procedure of the state classification, the state change detection method first should be evaluated on single sensor modalities. Since in the state recognition, location and acceleration worked distinctively better than phone-usage or sound, this evaluation is only applied for location and acceleration data. The evaluation process is as follows:

1. **Building the baseline model:** Since this evaluation is not done online but with an entire data-set available and labeled, different classes could be tested as a baseline. In order to build the respective model, 66% of the data (from evenly distributed classes) was used as a training set to build a multivariate Gaussian distribution model. The remaining 33% of data was then used to test.
2. **Determine distance and appropriate thresholds:** To measure the distance of data points to the model the Mahalanobis distance was chosen. To train the model, the distances of each training points were calculated and normalized. If the normalized distance of a sample was within a given threshold it was classified as "no change", otherwise it was marked as "change". Different threshold values were tested (see the effects of the different thresholds in figure §3.10)
3. **Model evaluation:** All data that had not been used for training was used to test. This was done again by normalizing each test-sample's distance to the model distribution and then evaluating whether the value was within or outside of the thresholds. Based on these estimates precision and recall were calculated.
4. **Optimize:** To eliminate outliers, the evaluation - steps (1)-(3) - was repeated in 1000x cross validation with different randomly picked test/training splits.

	ACC only		LOC only	
Patient	Recall	Precision	Recall	Precision
P0101	78.0%	89.1%	83.8%	62.6%
P0102	70.1%	72.3%	93.7%	77.3%
P0201	90.0%	75.6%	97.5%	76.6%
P0302	60.5%	72.8%	100.0%	68.5%
P0502	83.8%	88.9%	72.4%	65.0%
P0602	98.5%	76.0%	100.0%	74.5%
P0702	77.9%	70.8%	91.4%	73.7%
P0802	66.9%	72.8%	94.1%	72.0%
P0902	74.5%	78.4%	100.0%	74.5%
P1002	76.4%	74.7%	100.0%	73.1%
Average	**77.7%**	**77.1%**	**93.3%**	**71.8%**
depressive	83.8%	80.4%	99.2%	84.0%
sever depr.	71.5%	70.0%	82.9%	58.6%
medium depr.	78.2%	75.7%	100.0%	76.9%
slightly depr.	98.5%	86.5%	100.0%	85.7%
normal	69.2%	73.2%	90.1%	65.8%
slightly manic	73.3%	86.8%	83.8%	49.3%
medium manic	100.0%	94.4%	-	-
Average	**82.1%**	**81.0%**	**92.7%**	**70.1%**

Table 3.8: Results of change detection for the single modalities acceleration and location for each patient (top) and summarized for each class (bottom)

The upper part of Table §3.8 provides the results of the state change detection for the two single modalities acceleration and location. The second part divides the result of each sensor modality into precision and recall over each possible class. For the acceleration, the results of precision and recall per patient are around 77%. Summarized over the classes, precision and recall of the acceleration based state change detection are even above

[5] Similar to the Sensor Fusion, the work in this subsection was joined work with my colleague Gernot Bahle. Due to time constraints in the underlying project, Mr. Bahle has supported my work in various different aspect, which in hindsight cannot be separated entirely. But I want to state in particular that Mr. Bahle has implemented the change detection and fusion algorithms and ran the evaluation of the change detection with the pre-processed and classified sensor data.

80%. Within location evaluation recall for both (per patient and classes) is at a very high above 90%. Precision results are, again for both, around at 70%. This means for the location-based change detection, more than 90% of changes were detected correctly, but also quite a number of false positives were determined.

3.9.2 Fusion of Sensor Modalities

The results of the single modality change detection evaluation are far from being bad. Still, since it has been highlighted that the detection of change has a critical value in real-life treatment, even better results would be desired Thus, not unlike in the state recognition, the next step towards improving the results is by fusing the sensor modalities.
Again, the fusion intends both, to improve the accuracy of the change detection and to extend the number of suitable sensor data-sets the change detection is based on. The state change detection does not use a method that was already implemented in a data mining library, but a unique probability density functions (PDFs) for the default state (start-state) was built and fitted to the specific requirements of the goal. Thus, it is now also possible to leverage this PDF to evaluate various strategies for fusing the modalities:

Different Fusion Variants

First of all, before the actual fusion could be performed, the change detection of each single sensor modality was evaluated according to the algorithm introduced above individually. Subsequently, the results of the individual modalities were fused. Note, like in the fusion for the state recognition, if data was available for only one modality, this modality was used. When data from two or more modalities was at hand, three different fusion modalities were applied and tested:

1. **Logical AND fusion:** only when all currently utilized modalities detected a change, a change was assumed.
2. **Logical OR fusion:** if any of the modalities detected a change, a change was assumed.
3. **Weighted fusion:** using the normalized distance for all modalities, they were summed up according to the sigmoid weighting function form the state recognition. Thresholds were also combined accordingly, and the fused distance was evaluated for either being smaller or larger than the new threshold.

Results of Fusion Modalities

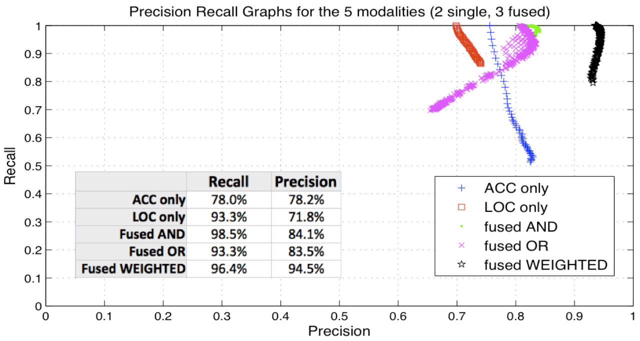
The precision/recall-graphs in Figure §3.10 provide a first impression of the results of the fusion. These graphs display the different fusion approaches and variants in sweeping over various thresholds. It is obvious that the weighted fusion approach is by far the best, reaching in its best combination a recall of above 95% and also a precision of far beyond 90%. For more detail, Table §3.9 provides precision and recall values of the different fusion approaches. The values in Table §3.9 are calculated using the optimal average threshold point.

	Fused AND	**Fused OR**	**Fus. WEIGHTED**
Patient	Recall / Precision	Recall / Precision	Recall / Precision
P0101	96.1% / 82.1%	81.9% / 79.6%	91.1% / 93.4%
P0102	96.8% / 90.5%	94.2% / 90.6%	86.2% / 96.8%
P0201	98.3% / 79.8%	94.2% / 78.7%	97.3% / 92.9%
P0302	100.0% / 82.0%	100.0% / 81.9%	100.0% / 93.8%
P0502	98.8% / 91.5%	95.7% / 91.3%	97.8% / 97.6%
P0602	100.0% / 69.7%	99.4% / 70.3%	100.0% / 87.4%
P0702	97.9% / 89.7%	94.6% / 88.1%	96.8% / 97.1%
P0802	98.0% / 83.9%	86.4% / 83.0%	95.6% / 95.2%
P0902	100.0% / 89.1%	100.0% / 89.0%	100.0% / 97.1%
P1002	100.0% / 76.3%	90.2% / 76.9%	100.0% / 91.2%
Average	**98.6% / 83.5%**	**93.7% /82.9%**	**96.5% / 94.2%**

Table 3.9: Results of the different fusion modalities (fusing acceleration and location) per patient (upper table) and over different classes (lower table)

Again the weighted fusion approach performs best. Moreover, even though inferior to the weighted fusion, in comparison to the single modality evaluation the results of both logic fusion approaches are better. Here, specifically the precision is explicitly better, all fair above 80%, than in the single modality evaluation, and also the recall is improved to more than 90% on average. Still, a closer look at single patients reveals that some individual results are clearly below 80%. Not so regarding the weighted fusion. Here not even one individual result drops below 85% (precision and recall). On the contrary, in terms of recall except for P0102, the weighted fusion of all is above 90% with even four out of ten patients reaching 100% recall. Meaning almost no change is missed.

The results for precision are equally high. Only patient P0602 has a precision of below 90% (87.7%). All other patients have a precision of more than 90%. Averaging over all, patients the average precision for the weighted fusion is at a high 94.5% and the average recall even at a high 96.4%. The fact that primarily the self-built approach performed best is not surprising, as optimal weighting is an essential aspect of successful classifier fusion.



	Recall	Precision
ACC only	78.0%	78.2%
LOC only	93.3%	71.8%
Fused AND	98.5%	84.1%
Fused OR	93.3%	83.5%
Fused WEIGHTED	96.4%	94.5%

Table 3.10: Prec./Recall graphs for ACC and LOC only, fused AND, fused OR, and fused WEIGHTED modality.

3.10 Challenges and Issues of the Real-Life Deployment in Mental Care

When deploying a real-life study in healthcare and especially when doing so in a psychiatric (hospital) environment, a number of challenges have to be faced. Mostly these challenges are of technical but even more prominently of human nature. The following part summarizes the most challenging problems that were encountered and introduces which strategies could be used to overcome these problems.[1]

3.10.1 Ethics Board Approval

Specifically, in deploying technology in health care, the ethics board (concerning including patients) or the works council (concerning including clinic personnel) plays and often underestimated but crucial role. Typically, the ethics board is comprised of professionals from health-care and natural science and furthermore includes a few members of other scientific areas. Therefore, when it comes to deploying a technology-based study in health-care, it is essential to present all information in a way that professionals from other disciplines can understand. Otherwise, if the proposal is rejected or even if there is the request to re-apply in a modified form, it can postpone the deployment.

A possible way to limit the risks of being rejected is to involve people from relevant other professions, in the process of creating the ethics board proposal, and to take all concerns, even though they might seem unimportant to a technician, very seriously. In our study, the ethics board application and hearing was done by a joint group of psychiatric professionals and computer scientists. The application was approved by the ethics board instantly, but not without restrictions: The most important and crucial of them was the constraint that all sensor readings had to be anonymized in a way it was not possible to defer information about the particular person. The privacy of the participant had to be guaranteed. While this was not an issue for the acceleration data, GPS coordinates had to be transferred into a neutral coordination system before processing them.

Furthermore, in order to perform frequency-analysis on voice during phone calls, algorithms had to be developed which scrambled the sound the way that it was not possible to restore, yet would still keep the frequency of the voice intact. This method has been introduced by Muaremi et al. [124]. This way the speech of a person is not understood any longer, while at the same time the performance of the acoustic analysis of the speech is not degraded.

3.10.2 Technical Challenges:

Next to the constraints imposed by the ethics board a number of technical aspects had to be dealt with:

Data Transmission: The original set-up of data transmission was designed to transmit the collected data automatically. All data would have been transmitted to a secure server belonging to the psychiatric hospital facilities via a secure connection. However, even though the infrastructure was already set up, this plans had to be discarded before the study started as it turned out that most of the possible participants neither owned an appropriate wireless Internet connection at home (at the time of the study) nor could full 3G Network and DSL coverage be guaranteed. Hence, the set-up was changed to storing all sensor readings on an internal SD card and transmitting them by hand every 2-3 weeks during the check-up appointment of the patients at the psychiatric hospital facilities.

Constraints by the Smart-phone OS: Another technical issue, at least at the time the study was conducted, was to find an appropriate android operating system. One of the significant advantages of the Android system is that it is not limited to one specific smart-phone brand but is available for various cell-phone types from various producers. This aspect, unfortunately, is one of the significant limitations that had to be dealing with, because the Android OS was partially adapted for different producers. The main issue that was encountered was to find an Android-based smart-phone with an OS that allowed to access the sensors in the background, even when the display was turned off. Not all android based operating system in the market permitted this in 2012. Later updates of the Android operating systems included this feature per default an thereby this issue was eliminated for later use.

Performance and Battery Life: First real test-deployments of the running smart-phone application revealed that the smart-phones had a tendency to get rather hot. In the beginning, the reason for this was not apparent, as this never happened during tests in the laboratory. Eventually, it turned out that, as some of the participants were living in rural areas where most of the time no WiFi was available, the app triggered the WiFi port, by default of the OS, to increase the scan-frequency. Moreover, next to increasing the temperature it decreased the battery-life tremendously.

As in numerous other technical applications, the central critical part in using a smart-phone for data recording is the phone's battery life. Constant operating of all sensors in a high resolution reduces the battery life to a couple of hours (even with using an extended battery pack) what would make this application unusable in real-life. To overcome the issues of performance and battery life, the design of the system was optimized. The acceleration sensor was used to trigger most other sensors. This

[1] In part, entire paragraphs describing the particular challenges and issues faced during the study have been taken from the following publication. All texts taken from this papers have been written solely by the author of this thesis:
Grünerbl A. et al. (2014) [54] and Mayora O., Grünerbl A. et al. (2014) [127] and (2016) [128], please refer to respective entries in the literature list or beginning of this chapter

is feasible for example, as an unmoved cell-phone will not change its position. Therefore GPS/WiFi sensing can be reduced to a minimum while the cell-phone is not moved. Furthermore, as long as a person stays inside of a building GPS is of little use and indoors WiFi might be used to navigate. Therefore, the usage of GPS, which itself is highly power-consuming, was turned off while WiFi was available (and vice versa)!

3.10.3 Practical Issues:

Humans dealing with Technology: Conducting a real-life study means having humans, more precisely people who have a mental illness, and were not explicitly trained in using technology, to deal with a sophisticated sensor system. This will raise issues of human nature. Using a cell-phone based sensor platform could eliminate the majority of these issues, e.g., dealing with a sensor-system that people were not familiar with. Cell-phones have grown into our society in the last 20 years and are accepted as part of our lives. Nearly everybody, including older people, owns a cell phone and therefore people usually are not afraid to get in touch with them. Nevertheless, a particular fondness of test-subjects and patients about modern technology turned out to be a precondition for a successful deployment of this study. We came across this issue with the first patient during the do-ability evaluation, who, even though was very eager to participate in the study, eventually was overwhelmed by the various unfamiliar functionalities the smart-phone provided and therefore dropped out.

Privacy: To our general surprise, most patients were not particularly concerned about privacy as long as a sensitive and anonymized treatment of their data was guaranteed. Bi-polar patient, especially when they start to realize and accept their disorder, are aware that they need help because they do not want to experience extreme episodes anymore. Therefore, a lot of them were willing to try new ways. It was primarily so if those ways provided an outlook to help them to reduce the number of anti-depressants or mood stabilizing medicines, since these medications usually go with side effects. Thus, when they were asked to participate in a study that might help them in the future to deal with their disorder, a sufficient number of patients was willing to participate. In this regard, of course with the help of the psychiatrist who had to establish the contact with the patients, the recruitment of the patients was less difficult than expected.

Benefits: While conducting this study, it turned out that an appropriate beneficiary/compensation system for the participants enhances the prospects of success. Precisely so in studies where the pool of possible test subjects is limited. In the studies conducted in this chapter, the practice of letting the test subject keep the smart-phone afterward proved to be an additional motivation for some of the participants.

3.11 Discussion and Conclusion

This chapter introduced, step by step, work starting at the assumption that patients with affective mood disorders express the state of their disorder in a way that is possible to grasp with common location and activity sensors. Thus, it was assumed that smart-phone internal sensors would be able to record particular aspects of the disorder. These data-traces would allow recognizing the mental state of these patients, and further detecting the onset of state change. The initial do-ability evaluation confirmed the fundamental correctness of this assumptions by showing linear correlations between various sensor-based features and the patient's self-assessed state within and above the 90% confidence interval.

Nevertheless, this do-ability evaluation also showed that there are no absolute features and values fitting all patients. Meaning that the correlation analysis showed that some of the features extracted from the sensor data work for some patients but not for others. For some patients many different features correlated, for others only a few did. This though, generally reflects the patterns of bipolar disorder. In this disorder there are basic direction like "depressive people move less" but the exact expression of how much less, or what "less" actually means is, to some extent, unique for each patient. Moreover, different patients have different focus points. Some patients express their state mainly through movement, other through activities or social behavior, again others through a combination of different aspects. Some psychiatrists say that there are as many expressions of the disorder as there are patients.

However, the analysis of the sensor data shows that, for every patient, features in all relevant aspects exist, which correlate with the patient's state. This alone is an exciting result, as it confirms that objective sensor data from a smart-phone can be used to determine the current state of a patient. This opens up various further options like a working state recognition. Hence, proceeding from this point, the next step certainly was to understand how methods for state recognition could look like. More importantly though, this state recognition and afterwards state change detection had to be developed to be robust enough to work even when not every feature used would be able to contribute.

Successfully, the developed methods for state recognition can provide accuracy of up to 70-80 % in single sensor classification and sensor fusion. The detection of state changes goes even beyond precision and recall of 95%. Thus meaning that even features, which alone do not sufficiently reflect the patient's state, can contribute to a stable detection of change onset.
All results presented in this chapter, however, must be

seen in the light of a noisy ground-truth and the fact that a patient's behavior cannot be expected to be entirely consistent on a daily basis. Even a severely ill person can have a good day, and a highly manic person might not always be on 150. Also, the change detection alone (which performs very well, with a precision/recall of around 95%) is enough in most cases since the definitive diagnosis has to be done by the doctor and for several ethnic and liability reasons has always been intended to be done by the doctor and not by a "smart-phone". Thus, a correct recognition of state (over seven possible classes) of 70-80% can be presumed to be a reasonably strong result.

Indeed, when discussing the value of the results presented in this chapter, it has to be considered that for some patients the amount of available labeled data and the number of state-classes was small. So, it is difficult to say whether in a larger-scale trial the same 95% accuracy for change detection would be reached for these patients. However, as outlined before, above 95% accuracy is not even required, and a significantly smaller accuracy would be sufficient for practical applications. Once again, the diagnosis is made by the doctor, and occasional and false positive alarms would at most result in an unneeded appointment with a psychiatrist. Thus false positives could even be helpful to remind the patient of being alert about their state but are less harmful than false negatives.

On the other hand, as described in section §3.5, the use case of a state change recognition possibly triggering a reaction after the persistent occurrence of change does not necessarily require unusually high accuracy. More significant than a few percentage points either way regarding performance is the fact that the data was collected under conditions that correspond precisely to the way a system would be used in real life! The most valuable achievement of this work is that the introduced system has been derived from and validated by a large, real-world data-set (more than 800 days of sensor recordings) that was recorded during the every-day lives of real patients. Further, these real patients mainly lived in a rural environment, meaning they were not necessarily tech-savvy or were owning fully developed technical infrastructures. The patients were given off-the-shelf devices with no other supervision than a visit to a doctor every three weeks. Additionally, the results with the limitations of the data recording mean that the system can handle irregular availability of data while still providing sufficient results. In my opinion, this is a significant improvement over artificial lab settings and qualitative studies.

Furthermore, the real-world application of the state change detection algorithm is almost plug-and-play. It only requires data of the current state to work, not data of all possible states as a state recognition would require. Therefore, it would be possible to utilize the change detection without excessive and time consuming training and labeling phases. Generally speaking, the system could aid psychologists from the patient's first visit onwards. To the best of my knowledge such a system like the introduced change detection, which is able to detect early changes in the state of a bipolar disorder patient and moreover, a system that works under the constraints of every-day life and does not require long periods of training and calibration, does not yet (2015) exist. For this reason, I believe that the work presented here could become a potent tool in supporting the treatment of bipolar disorder. In this context I consider the results of this chapter to be inspiring and relevant for potential future applications in mental care.

Evaluating Cognitive State Detection: Smart phone based Objective Sensing compared with Standards of Subjective Self-Assessment

Daily self-assessments using standard psychological questionnaires are an established technique. Currently, they are the standard for determining the progression of affective mood disorders in psychiatric care. However, in addition to the effort it requires, which causes adherence issues, self-assessments have the widely recognized drawback of being subjective! Subjectivity, in this context, means that self-assessments are easily biased. For example, the situation or emotions of the patient at the point they are performing a self-assessment will be reflected in their answers. Furthermore, many patients lack sufficient self-awareness to be able to report on their conditions correctly. Therefore, self-assessments often are inaccurate. Some psychiatrists go as far as to predict that many patients have deferral in self-awareness by up to 7 days. Thus, in reality, only experienced and self-aware patients can use self-assessment as a useful tool.

While many experts suggested that an automated sensor-based system, being objective by its nature, would be able to overcome the problem of subjectivity, to date this claim could not be supported, as there is too little evidence available. Providing the required evidence is the objective of this chapter.
Thus following part will compare automatic activity and location-based state recognition (the outcome of the previous chapter) and patient's self-assessments, directly to psychological state assessments (actual diagnosis). This comparison will be made with the same 12-week real-life data-set that had been introduced in the previous chapter.
Both the automatic sensor evaluation state curves and the self-assessment state-progression curves were evaluated against state progression curves generated from an objective psychiatric examination (psychiatrist diagnosis, ground truth). The comparison of the curves was calculated by using the dynamic time warping (DTW) technique and the DTW similarity metric as a quantitative quality indication. The results of this analysis highlight the three different aspects of this topic:

1. The doctor's perception that the patients' self-assessments are often biased and shifted for several days can be confirmed.
2. A comparison of (such a shifted) self-assessment and the sensor based automatic state recognition shows that still both, sensor data and self-assessment, correlate and thus primarily express the same information.
3. The DTW analysis comparing the similarity of curves, thus indicating which of both modalities comes closer to the actual diagnosis, explicitly favors the sensor based state recognition.

By taking the cumulative distance between the respective curves into account, it is demonstrated that the error in the sensor based state recognition, is considerably smaller (by an average of 62 percentage points) than the error in the self-assessment.

The author of this thesis has published the work, contents, pictures, most tables and also partially text of this chapter in following publications listed in this footnote (for more details about these publications, please refer to the entry [129], in the publication list). Any text taken from these papers and used in this thesis has been written solely by the author of this thesis. The work in this chapter has been published mainly in the following publications:

- Gruenerbl A. et al. "Sensor vs. Human: Comparing Sensor Based State Monitoring with Questionnaire Based Self-Assessment in Bipolar Disorder Patients." In: Proceedings of the 18th International Symposium on Wearable Computers. IEEE International Symposium on Wearable Computers (ISWC-2014), September 13-17, Seattle, Washington, USA, ACM, 9/2014. [129]
- Gruenerbl A. et al. "Assessing Delayed Self-Perception in Bipolar Disorder Patients." (to be submitted at) International journal of bipolar disorders, Springer Heidelberg, 2017
- Gruenerbl A. et al. "Comparison of Sensor and Self-Assessment Based State Recognition in Bipolar Disorder patients" (preliminary title)., (to be submitted at) In Journal of Affective Disorders, 2017

4.1 Motivation:

Modern medicine can draw upon a variety of technological applications. Different imaging modalities like X-ray, CAT scans, or MRI already are standard tools for the diagnosis of numerous diseases. Even further, many illnesses can be healed today because they can be detected early enough due to medical image analysis. On a kind of macro level, surgeons nowadays can use robots and endoscopic cameras to reduce the invasive-ness of treatments. Via tele-systems, it is possible to connect with specialists all over the world, even during surgery. Thus it can be stated that technology has changed and enhanced physical medicine during the last 30 years to a vast extent.
In contrast, until now, psychiatric care can barely rely on technology, neither for diagnosis nor treatment. In bipolar disorders treatment, doctors find it still to be a challenge to provide the optimal medication and treatment to their patients. One of the main issues here is the fact that the effectiveness of medication treatment highly depends on the time when it is administered. Ideally, the medication should start (or be adapted) as early as possible, at the point, a change towards an episode starts to manifest. However, to determine this point is rather difficult. For one, this is because, with 100 bipolar patients, the doctors have to deal with 100 unique expressions of the disorder. Thus, psychiatrists and therapists need to develop an understanding and a kind of "6th sense" to diagnose a mental disorder correctly. Besides experience and observation, the primary source of information, and so far, "standardized" diagnosis tool is and are various forms of self-ratings, self-rating-tests, and reports of relatives.

Unfortunately, self-ratings are prone to bias and are dependent on the subjects' mental state and therefore, have been criticized in the past. For example, in a depressive state, a patient has only a limited recollection of the manic days a week ago and vice versa. Austin et al. [130] highlight different phenomena in the response sets of self-reports with their correlation to different personality and cognitive disorders. These are acquiescence (or yea-saying - meaning people tend to agree with everything), socially desirable responding (responding the way the respondent believes the person asking would want to) and extreme responding (meaning either yes or no without reflecting options in-between). Moreover, psychologists have warned that human memory is fallible (Schacter, [131]), and people often "remember" events that never happened. Thus the reliability of self-reported data has to be considered tenuous. Therefore, to interpret the state of a patient correctly, and thus to get the diagnosis and the treatment right, requires a great deal of experience from the psychiatrist.

For the second primary source of diagnosis, besides interpreting the self-assessment, psychiatrist mainly try to capture the way the patients behave during the limited time at an appointment when the psychiatrist, in fact, sees the patient. Also, in various personal discussions, psychiatrists repeatedly stated that in their experience, many patients have a deferred awareness of the change of their unique mental condition. Some psychiatrists even got as far as to estimate a deferral to recognize a change in their state of approximately one week.
Still, so far, this is only an observation that has never been evaluated professionally so far. Ten years ago, a study by Elgie and Morselli [132] addressed the question whether self-reports could be valid at all, in the light of known problems with attention, concentration, and memory, reported by bipolar patients. Its results suggested that most bipolar patients demonstrate outward signs of cognitive impairment, but they are unable to report them accurately. At least by using available self-report inventories, the patients were unable to express impairments sufficiently. Even though this study backs the observation of deferred perception of emotion to some extent, despite a thorough evaluation no research could be found that actually conducted an extensive evaluation measuring if and how much self-perception of bipolar patients can be delayed.

Given these circumstances and considering the achievements of pervasive computing in various fields, including health-care, activity recognition recommends itself as an application for mental care. However, due to difficulties in performing daily monitoring and due to skepticism within the mental care field, which limits support from mental care professionals, very little work has been done yet.

4.2 Related Work:

In the area of mental health-care, the majority of systems deployed to date focus on supporting self-monitoring. Bopp et al. [102] and Yun et al.[103] describe systems that require patient feedback through frequent questionnaires or text messages.

Other systems, like those introduced by Burns et al.[118], Likamwa et al. [119], and the Optimism platform [121] similarly rely on self-reporting, in these case implemented on smartphones. They certainly are suitable for aiding the patients in logging self-reported mood, activities, and quality of sleep to monitor affective mood disorders. However, these systems often require constant interaction and feedback from the patient. Regarding automatic recognition of the mental state, much less work exists, in particular, work involving real-world studies and off the shelf devices like smartphones.

Massey et al. [116] describe an experimental analysis of a mobile health system for mood disorders where they introduce a list of possible sensors for mood detection, yet focus on technical aspects like the line of sight and reception rate, optimal coverage and optimal placement of on-body sensors.

In [122], Frost et al. use a self-developed application to record subjective and objective data from patients who have bipolar disorder. Even though their main focus lies on self-reported information, in passing, they also utilize coarse objective sensor data (acceleration fragments and phone call statistics) to try to estimate future shifts of a patient's mental state. These predictions are compared to forecasts derived from the self-reporting data.

4.3 Objectives and Contributions

The work in the previous Chapter (Chapter §3) has introduced a sensor based and thus objective way to evaluate the state of patients suffering from affective disorders with reasonable accuracy. In psychiatric care, self-assessments, the standards currently used to evaluate the state of disorder have the known drawback of being inaccurate. Since there are no better alternatives yet, self-assessments are used. With the promising results introduced in Chapter §3, the question naturally arises, whether this relatively objective method might be able to compete with the standard method used in psychiatry,(subjective) self-assessment, or maybe would outperform the self-assessment? In this regard, the primary objective of this chapter is to compare results of the -based activity recognition application to the patient's self-assessment.

After providing a quick recap of available data-sets (see section §4.4), the chapter starts with analyzing a simple smart-phone-based daily self-assessment. Its weaknesses and limitations will be depicted, and its inaccuracy evaluated. See section §4.5. In the course of this analysis, also the deferral of change awareness and self-perception of the patients, as mentioned by a psychiatrist, is evaluated objectively, which has not been achieved in psychiatry before. Results of this comparison reveal an actual deferral of state change awareness of more than four days in two third of the patients.

Subsequently, this work evaluates how smart-phone-based objective sensor data qualify better to reflect the patient's actual state than the current standard. Both the sensor based classification of state curves and the patient's self-assessment curves are evaluated with the dynamic time warping technique towards the ground-truth of frequent examinations. This evaluation shows more than 60% less error for the sensor-based method than the self-assessment. See section §4.6. After analyzing the feedback provided by doctors and patients in section §4.7 this chapter closes with a discussion (see section §4.8)

4.4 Collection of Different Data Sources

The work in this chapter, as has been pointed out, is an extension of the previous chapter (Chapter §3). Hence, the sensor data-set used for this analysis is the same and has been introduced. The following section thus only provides a summary of the available sensor/data sources and only goes into more detail if it is necessary for comprehension. For more information, please refer to respective sections in the previous chapter.

4.4.1 Sensor Data

The objective sensor data-sets were recorded with the patients' smart-phones and included three-axial acceleration, GPS-traces, phone-call behavior, and (scrambled) sound (during phone calls). In total data was recorded from 10 patients, including a total of more than 800 days of sensor recordings (average of 600 days of traces per sensing modality), documenting a total of 17 state changes across all patients (e.g., depressive to normal, etc.). See also the previous chapter for more detail.

4.4.2 Available Ground-Truth

The available ground-truth has been introduced in part in the previous chapter. Please refer for more detail.

Psychological and Psychiatric Assessment: Specially trained psychologists performed psychological state examinations (psychological standard scale tests performed with the patient) to validate the sensor data and to provide ground-truth. These scale tests included two foreign-rated and two self-rated tests designed for depression and mania and were carried out every three weeks. Please note that more frequent examinations could have resulted in a so-called "learning effect," meaning the patients might remember questions, criteria, and what they replied last time. Thus too many scale test would have biased outcomes. See also subsection §3.6.4 in the previous chapter.
The examinations resulted in a grade for each measurement on a 7-point scale between -3 for severe depression and +3 for severe mania with intermediate steps of depression, slight depression, normal (0) slight mania and mania. If necessary to express the state correctly, half grades were possible. To provide additional ground-truth for the time in-between the scale tests, specially trained psychologists talked to the patients over the phone and recorded their impressions.

Simple Self-Assessment: During the study, the patients were asked to fill in a simple daily questionnaire. Note that quite some work is being done currently in developing -based patient-self-assessment tools. As stated in [122], for example, it is essential to keep a self-assessment simple. Otherwise, the patients will quickly fatigue and stop to comply. Any other, the best interval of self-assessment entries have yet to be found. Asking the patient to give feedback every hour or several times a day is not feasible, the patient will start to forget to provide feedback or will get overwhelmed or stressed out soon. On the other hand, asking a patient to report weekly will very likely lead to mix-ups, and a blurred recollection of things and most importantly will likely miss many incidents. Therefore, in this study, it was decided to opt for a daily self-assessment, which would pop-up on the screen in the evening, at a time where people most likely would sit in front of the TV or spending time with other relaxing actives, but early enough so it would not wake the patients up.

The questionnaire was designed to take no more than 5-10 minutes per day. Next, to a few questions about time spent at specific locations (at home, outdoors, at work, and more) and about daily activities performed, the self-assessment mainly included (similar to [122]) self-ratings of mood, individual level of activity and subjective physical and mental state. After consultations with the participating psychiatrists, the self-rating was not done on the diagnostic Likert scale [133] (which some patients may have had problems with), but on a “how do you perceive your state” scale of 1 (very bad), through bad, ok, good to 5 (very good). This 1-5 scale is similar to an inverse school-grade scale (Austria) and therefore was easy to understand by the patients.

Furthermore, most people have a good perception of feeling inadequate or very bad, ok, good or very good, on the contrary to using a 7-point scale, where the patients would have been required to rate whether they felt slightly manic or manic or very depressed or slightly depressed. Such a specific self-rating would require experienced patients who can understand how it feels like to be "slightly manic." For later analysis, though, this procedure urged to develop a mechanism to bring the different scales together to ensure the comparability. Nevertheless, it helped the patients to rate themselves, and therefore, this procedure was used.

4.5 Quantifying the Limitations of Self-Assessment

Despite the outcome of the study by Elgie and Moreselli 2007 and the various sources questioning the reliability of self-report (as mentioned before), in the area of mental health care the majority of systems deployed to date still focus on supporting self-monitoring. Systems that require patient feedback through questionnaires or text messages are described in Bopp et al. [102] or Yun et al.[103]. Other systems, like those of Burns et al. [118] or LiKamWa et al. [119], for example, are suitable for aiding the patients in logging self-reported mood, activities, and quality of sleep to monitor affective disorders on their smartphones.

The main objective of the first analysis in this chapter was to evaluate whether the deferral in state change awareness, as mentioned above, is objectively measurable. This evaluation hypothesizes that the delay exceeds more than two days and thus actually might have a negative influence on the patient’s self-assessment.

4.5.1 Delayed Self-Perception

As mentioned in the introduction, psychiatrists report that patients’ perceptions of their condition may be delayed by several days. In private conversations, psychiatrists have mentioned that, in their experience, they believe that this shift can take up to seven days. The evaluation of this observation was initially not a primary objective of this work. However, the self-assessment analysis found that the experimental set-up (with frequent scale scores and self-assessments over several months for multiple patients) would easily allow us to take a first action in confirming this observation.

The first step in determining whether self-perception of the study patients actually was delayed, was to find a measure for calculating the similarity between the self-assessment and its so-called ground-truth (the actual diagnosed state based on the clinical score tests). Since both were present in numeric and curve form, the dynamic time warping (DTW) technique presented the optimal method. DTW is a well-known methodology to find the optimal alignment between two time-dependent data sequences, particularly when they vary in speed or are shifted in time. Briefly summarized the DTW finds the optimal alignment between two curves and afterward calculates the cumulated distance (plane) between the optimally aligned curves. Müller 2007 [134] provides a good overview of this technique. Using the DTW technique, the similarity between the self-assessment and the ground-truth was calculated.

In a second step, the similarity between the two sequences was re-evaluated, but this time in order to find the most optimal alignment between them by shifting the self-assessment along the ground-truth. This means the self-assessment was shifted day by day in respect to the ground-truth (imitating a delayed self-perception) from one to seven days and after each shift, the similarity to the ground-truth was calculated.

4.5.2 Results of Delayed Self-Perception

The results of this evaluation is summarized in Table §4.1. This table shows the cumulative distance (of the DTW) between the diagnosed state curves (ground truth) and the shifted self-assessment (shifted by 1-7 days) for all nine patients. For most patients, the cumulative distance is lowest and therefore, the similarity best, with a shift of approximately five days (median 5, average 4.3, mode 6!).

The result-table shows that the shift varies quite a bit in-between patients. For 2/3 (six out of these nine) patients, the cumulative distance is lowest with a shift of more than four days. Only 1/3 has a good self-perception, meaning the similarity is best with a shift of fewer than three days. However, for all patients, a shift of four days is definitely not best at all! Thus, for these circumstances, neither the mean nor the median is the best modality to express this result, but the most frequent value (mode). Looking at table §4.1 it becomes clear quickly that only 1/3 have a deferral of less than the average (4.3) and 2/3 show a deferral of more than the average. Moreover, only two of the patients have a deferral that comes close to the average, but 2/3 of the patients have a deferral close to the median of 6.

shift	1	2	3	4	5	6	7	Best
P0201	2.3	1.4	4.65	2.98	1.28	3.24	1.52	**5**
P0102	0.03	0.31	1.98	1.5	2.04	0	0	**1**
P0302	0.59	0.17	2.43	1.21	3.78	0.9	2.11	**2**
P0502	0.03	1.21	5.43	0.99	3.11	2.5	2.43	**1**
P0602	5.11	6.47	3.92	4.45	5.0	0.45	5.71	**6**
P0702	0.61	2.81	1.09	1.2	0.43	0.09	0.1	**6**
P0802	3.88	8.97	4.17	6.14	2.93	2.6	3.22	**6**
P0902	0.77	0.96	0.68	3.63	0.1	0.37	0.21	**5**
P1002	2.81	7.45	10.52	1.37	7.62	5.75	0.6	**7**
Average								4.3
Median								5
Mode								6

Table 4.1: DTW: cumulative distance between psychiatric scores and self-assessment (shifted by 1-7), distance is minimal with a shift of (mean) 4 days

These results support the psychiatrist's subjective observation that a shift in change-perception is measurable, and its deferral exceeds more than two days, and thus the hypothesis of this evaluation can be accepted. Nevertheless, this analysis also shows the uniqueness of the perception delay for each person. Therefore, regarding "enhancing the quality of self-reports," it would not be feasible to apply merely a 4-day shift onto the self-assessment. Moreover, systems trying to provide accurate self-assessment would either need a much more detailed and sophisticated (and therefore time-consuming) self-assessment or would need to spend significant effort in evaluating the individual shift of each patient's perception and a likely change of it within different states and over time. In summary, for most patients, the perception of state and hence the self-assessment is uniquely delayed and biased. Therefore, the results of the analysis above strongly suggest using other more reliable measures for monitoring the state of bipolar patients.

4.5.3 Discussion

The presented analysis was a first attempt to objectively evaluate the deferral in the perception of patients with bipolar disorder, generally affecting their ability to self-report, as has been reported by psychiatrists. The results of this analysis support this observation. Two third of the test subjects display a shift in their self-perception of five or more days. Only one-third of the test subject has an optimal alignment of self-reporting and self-assessment within two or fewer days.

Indeed, as the underlying study initially was not designed for this analysis, this work shows some limitations. Even though the study lasted for 12 weeks per study participant, the sample itself is small (9 test subjects) and therefore statistically not entirely representative. Furthermore, the study sample includes a high female rate (8 out of 9). Therefore, to strengthen the reported outcomes, this evaluation should be done with a more extensive set of test subjects. Besides, the current analysis provides no information on whether the shift is constant for each patient or might change with time or changes in the patient's condition. It is a reasonable hypothesis that patients in a neutral state might have a better perception of their condition than in a manic state. This aspect is also worth further investigation.

Nevertheless, with 66% of the test subjects showing a substantial shift in their self-perception, this work provides a first objective evaluation of the deferral of self-perception for bipolar patients. Thus this work calls out to psychiatry and psychotherapy to rethink the way the state of bipolar patients are assessed and work on more efficient and objective methods. Despite calling for better ways to assess the patient's state, the results of this work also come with a chance for therapists. More precisely, the knowledge about the likelihood of delays in the self-perception will help therapists and psychiatrists to understand their patients better and more efficiently interpret current self-assessment based measures for determining state and state changes and hence provide timely treatment.

4.6 Established Methods or New Sensors-Assessments?

Sensor-based analysis as such is, by definition, objective as they measure what is there to be measured. Of course, there are ways to influence the measured results, e.g., regarding a smart-phone as a sensor platform, sensors can be switched off, or the smart-phone could be forgotten (intentional or unintentional). Nevertheless, sensors are not influenced or biased by "daily well-being" or mood, and thus, sensor readings are deemed to be more reliable than subjective self-assessment. So the fundamental question is whether this can be confirmed in a real application like the state recognition of actual bipolar patients by a . Can smart-phone sensor based state recognition actually draw a better and more reliable picture of the patient's state and progression than the self-assessment? To answer this question is the objective of this section.

The envisioned use of state recognition is different:

Most of the time, the psychiatrist will not see the patient for weeks, except during acute treatment. Specifically, during a normal state phase, but also while going through an episode, the state recognition could provide frequent and valuable overviews to doctor about the patient's progress or state, and therefore offer the possibility to react, even in-between treatment appointments. Moreover, these overviews could, in retrospect, enable the psychiatrist as well as the patients to analyze the patient's behavior and progression to gain better insight into the patient's specific case.

The main advantages here are that -based sensor data is available daily without the requirement for interaction with the patients and the data. Thus the features and results are based on objective data, which is not shifted and not biased. Still, the question remains: can smart-phone sensor based state recognition actually draw a better and more reliable picture of the patient's state and progression than the self-assessment? The following part intends to provide an answer.

4.6.1 Sensor based State Recognition

Using the smart-phone sensor based data-set described earlier, and in the previous chapter, it was possible to successfully implement detection of patients' state changes and recognition of the mental state. (see also chapter §3) To briefly recall: With the acceleration and location features recorded by the patients' smartphones, supervised learning standard pattern recognition techniques were applied to the data to identify which state a patient had been in at a particular point in time. A randomly performed standard 33/66 percentage split was used to divide data into training and test samples. The actual classes for the classification were defined according to the diagnosis provided by psychologists (depressive, normal, manic with different degrees - up to 7 classes possible). As a classifier, the Naive Bayes included in Weka was used to estimate classes for the test-set. Other classifiers (e.g., KNN) were tested but achieved very similar results. The entire process was repeated 500 times in a cross-validation approach with random test/-training splits to eliminate artifacts. With this approach, it was possible to determine the state of the patient (up to 7 different degrees of mania or depression) with an accuracy of 70-80%.

More importantly, though, it was possible to confirm that sensor data can provide enough information to detect the necessary changes of state. Instead of a classifier that incorporates a model for each state relevant to the patient, the state change detection approach only has to build a model of one single "default state". All data-points falling outside this model are classified as a "change". In order to determine the border (threshold) between "in the model" and "outside the model," a set of values was tested, finally resulting in a precision/recall of more than 97%. The main application of a state change detection, obviously, is detecting the changes in behavior. The utilization of a sensor system to recognize/diagnose the patient's particular state, on the contrary, might be controversial. Patients are human beings, and therefore, they will not entirely behave according to a defined scheme, which could cause miss-diagnostics very easily.

The envisioned use of state recognition is different: Typically the psychiatrist will not meet the patient in person for weeks, precisely as long as a patient is in a neutral phase. Of course, this would be different during acute treatment. Regardless, during a healthy state, but also while being in an episode, the state recognition system would provide frequent overviews of the patient's behavior and progress of state to the doctor. Thus, it would provide a possibility to act and react even in-between treatment appointments. Of course, these overviews could, in retrospect, also enable the psychiatrist as well as the patients to analyze the patient's specific form of behavior and progression to gain a better understanding of the patient's particular case. The advantages here are that sensor data is available daily without requiring much interaction, and more importantly, the data is objective, not shifted, and not biased.

4.6.2 Correlation of Sensor and Self-Rating

Going back to the do-ability evaluation in the previous chapter (see chapter §3) a linear regression analysis on a small initial data-set showed that various features extracted from the sensor data correlates with the patient's self-assessment. In order to confirm the same effect with the new and larger data-collection, this linear regression was performed again, this time with even more features derived from more sensor-modalities than was done in the initial calculations.

The reason why this correlation was repeated with the new data-set is two-fold. First of all, the main reason was not to confirm the results but to analyze the effect of different features. As stated before, patients do not always concur with their behavior. Thus not every patient expresses their state in the same way. This way, it should be evaluated which features actually allow a fair comparison to the self-assessment. The second reason for performing the linear regression and T-test with the new data set was to confirm that both data sources carry the same inherent information about the patient's condition. If very few features correlated with low confidence, the question would be whether both modalities are actually comparable.

The performed correlation (linear regression) between sensor-data and the self-assessment shows values fair above 0 and therefore indicate a correlation between sensor data and self-assessment. In order to verify the statistical significance of the correlation again, a T-test of the regression was performed and proofed the correlation to be within the 90% confidence interval. For most patients and most features, the correlation even lies with the 99% confidence interval. Table §4.2 shows the correlation results (correlation and t-value) for all patients and for a number of different features. Note that Table §4.2 only

	Location					Acceleration				
Features		distances traveled		% of stay outdoors			movement/no movement		frequency variance	
Patients	N	correlation	t-value	correlation	t-val.	N	correlation	t-value	correlation	t-value
p0201	28	0.061	0.312	0.505	**2.987**	28	0.496	**2.609**	0.554	**7.291**
p0102	24	0.737	**5.118**	0.887	**9.003**	25	0.428	**2.341**	0.248	**-2.268**
p0302	41	-0.339	**-2.247**	0.412	**2.820**	45	0.272	**2.006**	0.349	1.675
p0502	22	0.599	**3.349**	-0.187	-0.852	31	-0.732	**-7.198**	0.131	-0.129
p0602	22	0.602	**3.369**	-0.259	-1.199	14	-0.923	**-7.147**	-0.850	**-5.046**
p0702	14	-0.505	**-2.026**	-0.748	**-3.898**	21	-0.215	-0.692	0.764	**3.586**
p0802	26	-0.662	**-4.324**	-0.354	**-1.852**	47	0.249	**1.792**	0.260	**-4.917**
p0902	32	0.123	0.679	-0.545	**-3.558**	28	0.468	**2.778**	0.307	**2.830**
p1002	5	-0.821	**-2.492**	0.117	0.203	15	-0.126	-0.891	-0.805	**-6.337**
90% C	N >=14 ->|t| >= 1.76 (N=5 ->|t| >= 2.02)									
99% C	N >=14 ->|t| >= 2.98 (N=5 ->|t| >= 4.02)									

Table 4.2: Correlation and t-test results for different selected acceleration and location features to the self-assessment

includes a few representative features. These are the distance traveled, percentage of stay outdoors, the relation of movement to no movement, and frequency variance. More features and their correlation values are listed in Appendix A in the table in section §7.3.
The t-value indicates that not all features correlate with the self-assessment for all patients within the 90% confidence interval, yet for all patients, features exist that do. This analysis, with its results, is significant as it confirms that the sensor-based analysis and the patient's self-assessment do indeed include the same information and thus are comparable. Regarding the evaluation of which of the data sources (sensor based or self-assessment), that is more accurate, which comes in the next section, this information is essential.

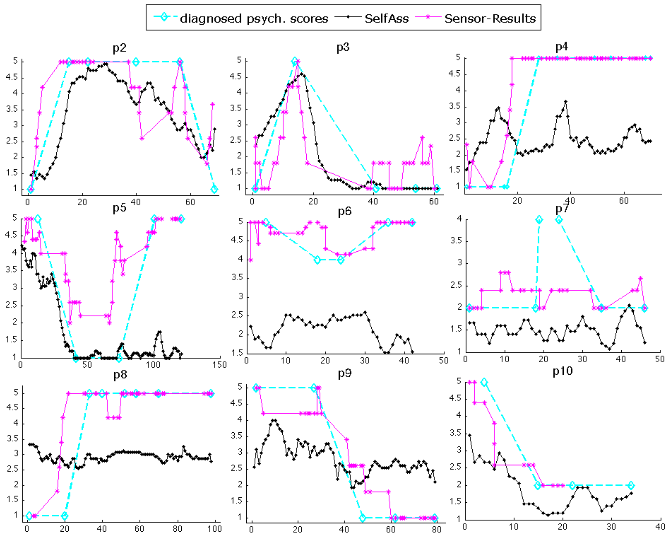


Table 4.3: Comparison of self-assessment (dots), recognition results (asterisks) and diagnosed psychological scores (dashed diamonds)

4.6.3 Sensors vs. Self-Assessment

As a first step towards evaluating which of the modes would be most accurate, self-assessment and sensor-based state recognition for each patient were compared visually with the ground truth (diagnosed psychological scores). See figure §4.3. Even without a trained eye, it is very clearly visible that the similarity between the sensor based recognition results (magenta asterisks) and the diagnosed state (cyan dashed diamonds) is much closer for most patients than the similarity of the self-assessment (black dots) and the diagnosed state! Specifically, for patients with weak self-perception (e.g., p4, p6, or p8), the sensor traces provide a much more accurate representation of the patient's state.

While analyzing the similarity between diagnosed psychological scores and either self-assessment or recognition results, one needs to keep in mind, that the diagnosed scores have been taken every three weeks with weekly intermediate telephone interviews (which, in turn, were aligned with the ground-truth scores scheme). Therefore, the ground-truth does not cover fluctuations by a few days. In light of this, neither self-assessment nor recognition results will match perfectly with the ground truth. Thus, the main interest in this evaluation lies in analyzing the similarity (shape and course) of respective curves and not in comparing day by day values.

For evaluating the similarity of the different curves to the ground-truth, once again the DTW technique [135] was used. This technique, particularly well fitting when time-depended data sequences vary in speed or are shifted in time. Note, in order to keep the comparison fair (self-assessment values are not necessarily in the same value range as the sensor values) all three curves were normalized between 1 and 0 before calculating the minimal distance between them. The resulting minimal cumulative distances for each patient, and each data source are listed in Table §4.4. It does not only include the minimal distance of the sensor data and the self-assessment but also a calculation of the minimal distance of the self-assessment shifted to the optimal alignment (shifted according to the evaluation of the previous section).

Except for p0102 (p3), the minimal distance of the sensor data to the ground-truth is considerably smaller (average of 62%) than the distance of the actual self-assessment to the ground-truth. Only for p0702 (p7), both curves are equally similar. Even in comparison to the optimally shifted self-assessment (shifted to reduce the deferral of self-perception, see also the previous section), the distance of the sensor data is still 57 percentage points smaller than of the shifted self-assessment. When excluding p0102 and p0702 (since for both the sensor-data and self-assessment perform equally good or

	minimal cumulative distance to ground-truth			difference of distance to ground-truth in number and %			
Patients	actual SA	SA shifted (by days)	Sensors	SA / Sensor	%	shifted SA / Sensor	%
p0201	0.33	0.24 (5)	0.05	-0.3	-83.9	-0.2	-78
p0102	0.29	0.29 (1)	0.62	0.3	109.5	0.3	112.7
p0302	8.69	8.63 (2)	0.29	-8.4	-96.6	-8.3	-96.6
p0502	21.6	21.0 (1)	0.6	-21	-97.2	-20.4	-97.1
p0602	6.27	3.28 (6)	0.54	-5.7	-91.4	-2.7	-83.6
p0702	5.1	3.98 (6)	5.09	0	-0.2	1.1	27.8
p0802	12.76	8.29 (6)	0.02	-12.7	-99.9	-8.3	-99.8
p0902	2.48	1.12 (5)	0.01	-2.5	-99.5	-1.1	-98.9
p1002	0.62	0.79 (7)	0.01	-0.6	-98.4	-0.8	-98.7
mean	**6.46**	**5.29**	**0.8**	**-5.66**	**-61.95**	**-4.49**	**-56.9**
w/o 0102+0702	7.5	6.2	0.2	-7.3	-95.3	-6	-932

Table 4.4: DTW Comparison of results: cumulative distance between diagnosed psychological scores and self-assessment (actualSA), shifted self-assessment (shiftedSA) and recognition results (sensors)

bad) the average cumulative distance of the sensor based recognition is more than 95 percentage points smaller than the average cumulative distances of the actual self-assessment. These results provide strong evidence that the sensor-based results are far more accurate and reliable than the self-assessment.

4.7 Participant's Perception

As has been pointed out in Chapter 2, a real-live deployment depends significantly on the participant's compliance and their understanding of why a study is deployed. Thus, during the study deployment, close contact with patients and psychiatrists, partially even relatives of the bipolar patients was established, and any of the participant's feedback was taken seriously.

4.7.1 Patient's Feedback:

Throughout the study most patients provided feedback. So we learned, that too much of detailed technical information was repelling for them because it confused them. Patients were mainly eager to know what the final goals of the system were and what their input would have to be. Most of them were not interested in what kind of sensors the application is logging. The few patients who were interested indicated their interest, and all the others were satisfied by the knowledge that sensors readings where logged. For most patients, it was much more important to be guided through the first days of using the new smart-phone and to get help in transferring all necessary data from their old cell-phone to the new one, than actually knowing what the system was doing. Therefore, after handing the to the patient and after transferring contacts to it, we offered them to visit the patients at the ward daily until they felt comfortable with the new device. Almost all patients have repeatedly used it.

Even though the patients could prohibit the usage of their data on specific days, no patient ever used this possibility. On the contrary, they did not see the need for such a feature. Now and then some journals had been forgotten to be filled in, yet the patients stated afterward, that they just forgot to deal with it and it was not about them not wanting to provide information.

4.7.2 Psychiatrist's Feedback:

The psychiatrists overall were more skeptical. This profession naturally seems to be very critical regarding using technology when it comes to their topics. Nevertheless, as one psychiatrist stated: "all other physicians have the possibility to look inside of their patients, and I would highly welcome the possibility to get a picture of my patient's mental state."

The potential of a system, as is described here, could offer was appreciated by most psychiatrists. Most of them found the topic fascinating, and to some degree, they stated that thay would most likely use a system that would assist them in gathering a picture of the patient's state. They encouraged us to provide them an understandable illustration of the patient's behavior (drawn from the sensor data collected). For example, they liked to get a plot of the paths of a patient (gathered by GPS/WiFi) or an overview of the patient's activity level over a couple of days or weeks. Mainly parameters that would picture changes in the behavior nicely would be of use for them. Next to this, psychiatrists are not interested in getting feedback or any suggestions from a technical system.

4.8 Discussion and Conclusion

The work presented in the previous chapter has introduced a smart-phone sensor-based system dedicated to facilitating the life of bipolar disorder patients and supporting their treatment. It incorporates sensor modalities covering (different) disease-relevant aspects of human behavior, but does not rely on self-assessment, thus having the potential of providing a less biased and more objective additional information source to health care professionals. This present chapter now had the intention to evaluate the results of this system in comparison to the actual current standards in psychiatric care, namely the self-assessment.

In the analysis, it was possible to prove, with statistical measures, that deferrals in the patients' self-perception (on average the deferral is 4 or more days) limit the accuracy of the self-assessment. Furthermore, a comparison of the self-assessment to the result of automatic state recognition was performed. This comparison demonstrated that sensor-based state recognition is not only an objective measurement but also substantially more accurate in displaying the patient's actual state. DTW showed that except for two patients, the cumulative distance between the recognized state and actual state is approximately 95 percentage points smaller than the cumulative distance between self-assessment and actual state.

A closer look at the patients, which undoubtedly provided less accuracy for the sensor data, still can reveal some interesting aspects but also some limitations of this work. Patient 0102 is the only patient whose self-assessment seems to be closer to the actual state then the recognition results. On the one hand, there are of course patients whose self-perception will be accurate. Such patients are those that have dealt with their illness for several years and have gathered much experience in understanding their minds' signals. On the other hand, the absolute values of the cumulative distances show that the difference between self-assessment and recognition results is relatively small. With a distance of 0.29 and 0.62, respectively (difference -0.33) in comparison to patient 0803 for example (a distance of 12.5), both sensor and self-assessment are very similar to the diagnosed state.

Moreover, a closer visual examination of the curves of p0102 exposes the limitation of only having a ground-truth value every three or slightly fewer weeks. When focusing on the peak around day 15, the curves show that the recognition detects the decline of the state a few days earlier than the self-assessment. It is hard to tell which one of the curves is the correct one, as the psychiatric score curve does not include a scores during these specific days. Concerning the evaluation of the shift in the patients' perception and the results of the similarity analysis in the previous chapter, the interpretation that the recognition is likely to be more accurate even for this patient than the self-assessment is valid. However, the available data does not allow to confirm this interpretation.

However, even though this work provides strong empirical evidence that sensor-based self-assessment is, on average, a more reliable and objective way of monitoring mental state and mood than patient's self-assessment, as was to be expected, some patients will likely be better in assessing their state than others.

At the end of this discussion, it should be highlighted again that the state recognition and self-assessment analysis is based on an extensive real-life data set. Thus, the results presented here are valid in a real-life setting, which emphasizes the value and relevance of these results. Therefore, I am confident that with these distinct results a basis can be set for further co-operation of pervasive computing and mental health care and furthermore establishing objective sensor-based monitoring support for the treatment of patients suffering from affective mood disorders.

Part II

Supporting People in Cognitively Stressfull Situations

Interlude

At this point, this thesis cuts across the topic. The previous three chapters analyzed the potential of location and activity recognition in terms of their ability to determine the well-being or mental state of individuals. In this context, two areas were identified that address issues that affect the lives of many people. Both are areas that would also benefit from sensor-assisted support in assessing cognitive states. They are dementia, an incurable cognitive disorder of older people, and on the other hand bipolar (affective) disorder, an equally incurable mood disorder that affects 2-10% of the population from adolescence or young adulthood onward.

In line with the objectives of this work, both scenarios were assessed progressively with varying degrees of complexity. Both began with an assumption and idea of how the specific requirements of each scenario could be met. Again, appropriate data collection studies were performed for both scenarios. An empirical evaluation of the collected data then helped to understand what kind of information, based on the original assumption, is actually contained in the data. Finally, a quantitative analysis of the algorithms developed to identify the mental states of the individuals involved in the different scenarios demonstrated the potential of these methods. In the second scenario, even a comparison with the current medical standards for the assessment of the course of the disorder could be carried out. The results of this comparison emphasize, in particular, the potential for the use of sensor-based detection methods in psychiatric practice.

In summary, chapters 2-4 have confirmed that, by using pervasive sensors, it is possible to collect enough relevant information to assess the cognitive state and well-being of a person without invasive or overly complex sensor systems. A next step could be, for example, to further increase the complexity or to go into detail, and thus, to perfect the already excellent results. However, this is more of a topic for potential business applications, and research-wise does not contain interesting scientific questions. Another idea might be to find other, even more increasingly complex scenarios, or to evaluate how far the detection of a condition can go, or how complex a scenario must be so that the condition detection no longer works! These questions are legitimate, but research-wise, there are much more interesting questions. For example, is it possible to actively support people in a stressful situation? Positively influencing behavior is one of the newer research fields.

In recent years moreover, in the field of pervasive computing, the question of how to analyze the behavior of a group of people has become increasingly relevant. So far, people were, in contrast, always analyzed individually. Particular questions are, "how can systems be trained to determine the cognitive status of a group", or "how can subtle interactions between people be recognized?" These questions include the analysis of collaboration between individuals as well as the question of how the behavior of a group of people, for example, in a high-pressure situation, can be assessed.
In this next part, in chapters 5 and 6, this work leaves the analysis of individual cognitive behavior and addresses precisely these questions.

Supporting Emergency Behavior: An Acceleration based Smart-Watch Instant Feedback CPR-Assistant

A health related emergency can happen to every person and many people will come across an emergency during their lives. Numerous articles in magazines, however, have addressed the issue that many people do not dare to perform cardiopulmonary resuscitation (CPR), the required emergency measurement in case of heart failure, in an emergency because they are afraid of doing harm. Expressly shocking are the figures in the German-speaking countries, where a First-Aid course is actually required to get a driver's license, and therefore many people should know what to do in an emergency. Yet, only 20% of the population would dare to resuscitate people with cardiac arrest!

An urgent emergency scenario is not only an ideal but also a relevant scenario. Positive support for people's behavior and self-confidence in emergency care can mean no less than saving lives. Therefore, the following chapter introduces an immediate feedback CPR support application based on smartwatch activity sensors. It is designed to increase confidence in performing resuscitation and provide the ability to acquire immediate skills for laypeople. Evaluations provide impressive numbers with laypeople being able to perform effective CPR for more than 50% of the time, which is an improvement of up to 160%.
This chapter will further assesses whether it is possible to positively support the behavior of people in an emergency and therefore addresses how people in a medical emergency scenario can be encouraged to do the right thing. Thus a particular side focus will be laid on evaluating whether such a system can help the users to be more confident in their actions.

In the following, based on the positive results of the first deployment of the CPR-watch with laypeople, the second part of this chapter goes even a step further. It will assess whether such an instant CPR feedback, which the smartwatch app can provide, also could influence the way medical professionals learn to perform CPR and to speed up the process of acquiring muscle memory for sufficient CPR. The effect of the support of the smartwatch app in a short training session was compared to standard human teacher CPR lessons. Improvements in the skills of 24% when using the watch a training-support versus 7.6% improvement during a human teaching lesson clearly display the potential of a smart-device for supporting the training of skills.

In summary, the following chapter asks certain questions in regards of sensor-devices supporting people in stressful situations. First, is an activity-based smart-watch app sufficient to support layperson confidence in performing resuscitation? Second, can such a smart-watch app help people gain the immediate skills they need in an urgent medical emergency? Furthermore, could such a smart-watch App even go as far as to help people learn to perform emergency procedures effectively?

The author of this thesis has published this work and most contents, all pictures, most tables and also partially text of this Chapter in the following publications. The author of this dissertation has written all the text in this chapter herself, specifically text-passages taken from these publications. For more details about these publications please also refer to the entries in the literature list:

- Gruenerbl A. et al. Smart-watch Life Saver: Smart-watch Interactive-feedback System for Improving Bystander CPR. In: Proceedings of the 2015 ACM International Symposium on Wearable Computers. IEEE International Symposium on Wearable Computers (ISWC), September 9-11, Osaka, Japan, Pages 19-26, ISWC '15, ISBN 978-1-4503-3578-2, ACM, 2015. [136]
- Gruenerbl A. et al. Training CPR with a Wearable Real-Time Feedback System. In: Proceedings of the 2018 ACM International Symposium on Wearable Computers. IEEE International Symposium on Wearable Computers (ISWC), Oktober 8-12, Singapore, ISWC'18, ACM, 2018. [137]

5.1 An Emergency Care Motivation

One of the leading causes of death in the western world is the "Out of Hospital Cardiac Arrest" (OHCA). In the United States every 90 seconds a person dies due to OHCA. This means more than 350,000 deaths per year. In Europe, approximately 40% of deaths in adults not older than 75 years are caused by OHCA. When suffering an OHCA, the chances of survival decrease by up to 7-10% per minute without measures to keep the brain oxygenated [138]. This means, eventually more than 95% of those suffering an OHCA die, as there are, for what reasons ever, no such measures performed. Measures to keep the brain oxygenated are called cardiopulmonary resuscitation (CPR), also known as chest compressions and mouth to mouth breaths, if possible in public spaces in combination with operating a publicly available automatic electronic defibrillator (AED). While operating and AED is easy and, in general, only means following instructions of the device, adequate and effective CPR by bystanders is essential for surviving an OHCA. Nevertheless, numbers regarding bystander CPR in real emergencies are quite shocking.

Throughout Europe, the number of people actually trusting themselves with performing First Aid and CPR varies quite a lot depending on the country. The average lies at 66%. In Denmark for example, as Wissenberg et al. [139] have ascertained, the rate of bystander CPR has increased from 21.1% in 2001 to 44.9% in 2010. Lay bystander resuscitation was attempted in total for 19,468 patients. Explicitly alarming though, are the numbers in the German-speaking countries, where according to the Red Cross and ADAC, only 15-20% of people would actually dare to perform CPR. People state that they would be insecure what to do and therefore were being afraid to do damage! Despite the fact that heart failure is already the worst case and no damage could make things worse, this reluctance to perform CPR is especially surprising, since like most EU countries German-speaking courtiers require obligatory First-Aid lessons for obtaining a driver's license and thus the majority of people has at least once in their life performed CPR on a training manikin.

A more detailed look by Grasner et al. [140] reveals that between 2004 and 2011 (n=11,788) in Germany bystander CPR was performed most often on young patients between 18 and 20 years of age (total 25%), and least often in those over 80 years (12%). Also, bystander CPR was performed significantly less often when OHCA happened in private homes, compared to OHCA in public areas! These observations are interesting. The actual reason has never been evaluated, but it seems to be feasible to conjecture that in private homes the shock of a relative or friend going into cardiac arrest could lead to a paralyzing behavior. On the other hand being in public and acting on an emergency of a stranger might be more comfortable and less emotionally stressful. Moreover, being in a public space might add to the expectancy to act by others (peer pressure), but also that the support by the peer (bystanders) might be encouraging.

Additionally, the availability of public AEDs (Automatic Electronic Defibrillator) that, in fact, can guide a person through First-Aid measurements might give a bystander the confidence and safety-net needed to dare to act (someone telling me what to do, so I have less responsibility). This theory seems to have been confirmed in another part of the world. In Japan, Sasaki et al. ([141] recorded the incidence of OHCA in Osaka. There, it was possible to measure that the increasing availability of AEDs in public places also increased the rate of bystander CPR. According to this study in 2004, the bystander CPR rate was at a total of 0%. By 2008, in four years, it had already improved up to 11%. In regards to the topic of this chapter, the fact of increasing bystander CPR with increasing availability of AEDs allows surmising that the availability of technical supporting devices can influence the willingness and possibly the self-confidence to act.

5.2 Related Work

The area of developing (whatever kind of) devices for supporting or assisting the act of performing CPR is still not very far stretched. Even-though emergency care is a very relevant topic and layperson CPR is an essential factor for surviving OHCAs, until today CPR supporting devices rather belong to the field of commercial products for paramedics (which are really not meant for bystanders) or in some cases to a kind of CPR training-tool for home-use, rather than actual scientific research intended to support and enhance bystander CPR. [1]

5.2.1 CPR Assistance Devices

A number of studies have shown that using CPR feedback devices can indeed help to enhance the quality of CPR. A few of these devices are commercially available. One of the most common is the CPR-meter (Laerdal, Philips, etc.), used in this work. Yeung et al. [142] provide a systematic review of the literature. Their findings, first and foremost, support the use of CPR feedback or prompt devices for improving skills during training. Studies by Buleon et al. [143] show significant improvements in CPR when people are using such CPR feedback

[1] Major parts of the text in this Related Work have been taken from the following paper by the author of this thesis. Any text taken from published papers has been written solely by the author of this thesis:
Gruenerbl et al., 2015 [136], and Gruenerbl et al., 2018 [137], please refer to respective entries in the literature list or beginning of this chapter.

devices. Other groups like Gonzales et al. [144] introduce alternative devices like photoelectric distance sensors that also manage to improve performance. All of these devices, however, are either expensive (several hundred Euros) and meant for laboratory and medical training environments, but not for being carried around in the hand bag of a by layperson during daily life, or are still in an experimental state.

5.2.2 CPR with Smart-Phone Apps

A look into Apple's Appstore in 2013 (time of the implementation of the CPR assistant) did not provide any iPhone-app assisting in CPR, by 2019 apps have arrived in the AppStore that are meant to guide people through emergencies. The Google Play Store (also 2013) listed some apps that were intended to assist in CPR, but these mainly gave information about First-Aid and how to perform CPR. Only "Metronome apps" provides live instruction (emitting sound in the proper frequency), yet lack a live feedback component.

On the research side, some papers dealing with smartphone-apps for CPR measurement are published. For instance, [145] Song et al. were using the trajectories derived by double integration of the acceleration of a smartphone for measuring compression depth. Their evaluation of the system shows only a minimal error range of 1.43 mm with a standard deviation of 1mm. Chan et al. [146] evaluate a CPR Feedback application for iPhones in a controlled study, with the results indicating the iPhone group reaching better compression depth than the control group (without iPhone app).

5.2.3 CPR support with Smart-Watches

The idea of using a smart-watch for assisting in CPR is not entirely new. The Philadelphia Business Journal reported in January 2015 on "Lifesaver", a smartwatch app, developed during the PennApps weekend hackathon [147]. To the best of my knowledge, however, this app has never been officially published in an App-Store, nor have any studies been performed with it. Also a look into the App Stores only provides a list of smartwatch extension apps to the existing support apps.

5.2.4 Assisting Devices in Teaching

An overview of various approaches of real-time feedback in motoring training can be found in [148]. This includes in particular augmented/virtual reality and different "classical" modalities such as screens and audio. In the wearable domain in specific tactile feedback (e.g., for music teaching) has been investigated [149]. Assisting devices in physics teaching have shown interesting results [150, 151] Recently, EMS muscle stimulation has also generated significant interest [152].

Concerning specifically CPR training, different approaches have been tried [142]. In the traditional approach, a professional instructor leads through the CPR steps, while in self-directed learning, an assisting device is used to teach the CPR procedure. Rasmussen et al. [153] show that participants with the assistance of a new dispatcher protocol performed CPR with a higher quality and higher motivation.

Some well-known self-learning methods for resuscitation are computer-based video training and application based methods conducted for smart-phones [154, 155, 156]. Alonso et al. [157] demonstrates improved success rates in CPR, using telematics support by an expert through head mounted displays. Wang et al. [158] introduce a feedback system using optical sensing that enhanced the chest compression quality in the paramedic's training.

In more recent years, newly developed augmented reality devices opened up research in regards of augmenting emergency training, as for example, the Microsoft HoloLens has been used for training of applying defibrillators [159].

5.3 Objectives and Contributions

Following the objectives of this thesis and considering the present situation in bystander CPR and the little research in this field, this chapter has set its goals to developing an easy to use smart-device application to support and promote bystander CPR for people with no or minimal knowledge about performing CPR. More concretely, this chapter aims to:

1. Determining an appropriate device, fitting into the daily life of people, serving as a platform for the CPR application, and developing a respective, easy to use, CPR support application for this device.
2. Evaluating the effect, this application has on the ability of laypeople to perform CPR and promote the user's confidence to perform CPR in emergency situations.
3. Evaluating the potential of the developed CPR application for settings that go beyond serving as a direct CPR assistant (e.g., for teaching CPR). Further, comparing the effect of the respective device beyond serving as direct CPR assistant to methods currently being used (e.g., in teaching CPR)

Following the objectives of this chapter, the specific contributions of this chapter are:

- Benefiting from the fact that a watch is not only worn most of the time but is also placed at the location where CPR is performed, at the wrist, a smart-watch is identified as optimal platform. In the following, an easy to use CPR feedback application (CPR-watch) for a smart-watch is developed. The applications is always at hand, in the right place, without requiring additional equipment. The CPR-watch application can allow untrained people to perform CPR according to the guidelines. (See Section §5.5)

- The CPR-watch application is then evaluated with 41 random untrained testers in three modalities. (See Section §5.6)
  The evaluation of the tests shows a distinct improvement in the effectiveness of the performed CPR when using the CPR-watch, compared to CPR without assistance. With watch assistance, around 50% of test subjects managed to stay within the recommended ranges at least for 50% of the time. Without support only just barely 20% of the participants were able to perform sufficient CPR at least for 50% of the time, even after receiving detailed explanations about effective CPR performance. (See Section §5.7)
- Additionally, the CPR watch application is evaluated for serving as an instant feedback training device for teaching CPR to nurse students. In a controlled randomized trial with 50 people, the effect of training with the feedback device is evaluated against the effect of standard human CPR teaching. The results confirm the impact of the CPR application as a training assistant, helping to improve CPR performance by more than 20%. (See Section §5.9)

5.4 Basics and Essentials of Cardiopulmonary Resuscitation

Cardiopulmonary Resuscitation (CPR) has been introduced first in 1960 [160], but parts of this concept already have been documented years earlier. The first documented chest compression in humans was performed by Dr. Friedrich Maass in 1891. In 1903, the first successful use of external chest compressions in human resuscitation was reported by Dr. George Crile, and in 1954, James Elam proved that expired air was sufficient to maintain adequate oxygenation.

Ever since the official introduction of CPR, its effectiveness and factors that are influencing the effectiveness of CPR have been researched scientifically. With gathering more evidence, the suggestions on CPR techniques, thus, have changed over the years. In 2005, both the European Resuscitation Council (ERC) and the American Heart Association (AHA) officially published the still stringent evidence-based guidelines for resuscitation. These guidelines suggest effective CPR as follows:

- perform compressions in a frequency of at least 100/min but not exceeding 120/min [161].
- perform compressions with a depth of at least two inches/5 cm [162] but not exceeding 6 cm (suggested by ERC, AHA does not specify a maximum depth)
- AHA and ERC recommend a 30:2 CPR rhythm (30 compressions, 2 breaths) - this suggestion is being changed to perform compressions without stop,specifically for OHCA and bystander CPR. For hospitals suggestions are still including breaths, but also allow to perform CPR with no breaths).
- In 2010 recommendations for compression depth was adapted from 40 mm to more than 50mm.

Even though there does not seem to exist evidence that deeper compression depth is related to damage from chest compression, the ERC recommends not to exceed 60 mm of compression depth even in large adults [163]. From the suggestions of the respective bodies, it can be inferred that two factors specifically influence the effectiveness of CPR. These are **compression depth** and **compression frequency**. Since publishing these suggestions, further research has confirmed the validity of these suggested values.

Idris [164] evaluated in 2012 that CPR is, in fact, most effective at a frequency of around 120 CPM (compressions per minute), but blood-flow and hence survival-rate decline with faster rates than 120 CPM. Idris also indicates that the probability of a return of spontaneous circulation is highest with 125 CPM but rapidly declines with higher compression rates. Furthermore, it also was demonstrated that with higher compression rates the compression depth suffers.
Further research indicates that compression depth of 40 mm or less depth results in fewer survivals than compressions of 50 mm and beyond. [165, 166]. Edelson [166] indicates that patients with OHCA who received CPR with a compression depth of 50 mm and deeper 30 seconds before operating an AED have a higher survival chance than those without. Stiell at al. [167] tried to evaluate the optimal compression depth and could confirm that survival rates increase with increased compression depths. Nevertheless, they could not provide clear evidence to back the 2010 adaptation of the compression depth recommendations. In 2014, Stiell at all. [168] determined the maximum survival within a compression depth interval of 40.3 to 55.3 mm (peak at 45.6 mm).

5.5 The smart-watch Instant Feedback Concept

Scanning literature, different options for possible smart-devices that promise to suffice as a platform for a CPR support application were considered. Nevertheless, these devices or sensor platforms have to meet a number of preconditions essential for such an application. These preconditions limit the number of possible devices. Following have been identified:

- The device has to be able to record the motion of performing CPR, meaning a device including a 3-axial accelerometer. Performing CPR has a very distinct signature and, performing it correctly, includes the entire body of the CPR giver, thus the location of the sensor is less important.

- The device has to be widely available for laypersons to be used; thus professional CPR devices like CPR-meters for paramedics had to be ruled out, since, as of now, they are not publicly distributed.
- The application has to be started easily, its usage must not disrupt the process of CPR and has to work on the fly without calibrations necessary.
- The device has to be able to provide instant feedback, thus needs a screen or the ability to produce sound.

With these preconditions, a number of possible devices seemed to work, still most of them had to be ruled out eventually:

- **Smart-Phone:** a smart-phone comes with all necessary functions. It can measure acceleration, can provide feedback (both visual and sound), is widely available, and apps can be started quickly. Nevertheless, using a smart-phone does raise the question of how to handle it or where to place it. Having the smart-phone in a pocket of one's clothes is easy to use, but strongly limits the options for feedback.
  Placed next to the patient hinders to measure the motion of CPR. Placed on the chest of the patient or in one hand of the CPR giver might either disrupt the process of performing CPR effectively or would require the smart-phone based app to be able to deal with a very uncertain situation. Thus, a smart-phone is not an optimal platform
- **Head-mounted devices (e.g., Google Glass):** as long as the devices would include the ability to measure acceleration and include a screen, such a device would work. Nevertheless, head-mounted devices are not widely available. Thus, Google Glass was downgraded in the list of ideal devices.
- **Smart-Watch:** Smart-watches are widely available. Most of them include acceleration sensors and of course they include a screen for providing feedback. Moreover, a smart-watch already is located exactly at the place where CPR is supposed to happen - the wrist of the CPR giver. Thus, a smart-watch seems to be the optimal device for incorporating a CPR support application.

Following these considerations and preconditions, an easy to understand and easy to use CPR feedback application was implemented by the research group of Embedded Intelligence at the DFKI [1]. For first tests, the application was installed on LG G Watch R with Android Wear OS, but any other Android-based smartwatch works as well provided it includes an accelerometer. The application has three main functionalities:

- **Provide Correct Compression Frequency:** At app start, the watch begins to vibrate with 110 bpm (beats per minute). 110 bpm was chosen, as it is the average frequency of the ideal compression rate of 100-120 CPM (compressions per minute).
  After the first tests showed that the vibration is a helpful feature to feel the rhythm when starting compressions, it also transpired that during performing CPR many persons stop feeling the vibrations. Thus adding to the vibration, a visual "metronome" was added in the form of a blinking screen (background of the application blinks black/blue in the respective frequency - see Figure §5.1 A/B). Due to the lack of a loudspeaker in older smart-watch versions, audio feedback was not implemented at this point.
- **Compression Depth Feedback:** The application provides color feedback in regards to the compression depth. In the center of the display a square is colored with either green (for correct compression depth of 50-60 mm), yellow if the compression depth goes beyond 60 mm, and red if the compression depth is less than 50 mm. See Figure §5.1 C-E.
  The compressions are detected by applying a peak detector to the accelerometer signal of the watch (see figure §5.2). By retrieving each peaks minimum and maximum and calculating their time difference, they are used to estimate the compression depth. The values and depth of compression were calibrated using a professional CPR-Meter.
- **Backwards Counting of Compressions:** Since both ERC and AHA still suggest a 30/2 compression/rescue breath alternation the application includes the ability to count compressions backward, starting by 30 and stopping once 30 effective consecutive compressions have been registered. As compressions smaller than 50 mm were considered not efficient, all compressions with red feedback are not counted.

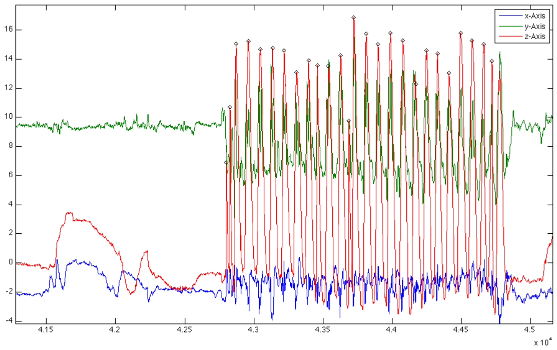


Figure 5.2: CPR shows very distinct peaks in the acceleration signal

[1] The first version of the smart-watch CPR app was implemented by my colleague Gerald Pirkl

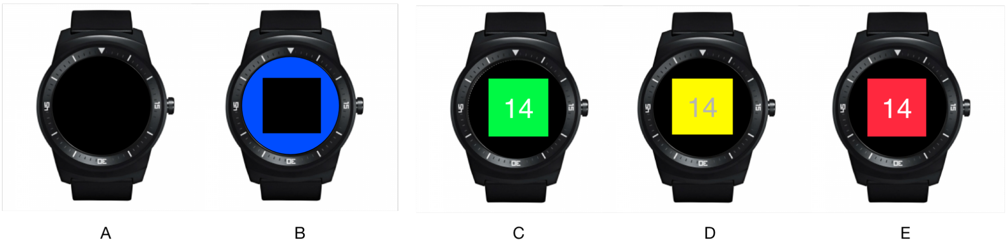


Figure 5.1: CPR smart-watch App: vibration and a blue/black (A, B) blinking with 110 cpm; color feedback about detected compression depth display center (green (C) - good, yellow (D) - beyond recommendation, red (E) - not sufficient)

5.6 CPR on the Fly - Gaining Instant Skills with CPR-Watch

The goal of using the CPR-watch was to evaluate whether this instant-feedback app would be able to help laypersons to perform CPR correctly. Nevertheless, to evaluate the effectiveness of the developed smart-watch feedback app, contrary to the work in the previous chapters, an actual real-life study was not applicable. A study intending to measure how people would perform in a real cardiac arrest emergency, would be challenging and tedious to set up, but also ethically hard to defend. Even paramedics do not train on real people but use specifically developed training manikin. Thus, for the evaluation of the smart-watch CPR-feedback application, a manikin, “Little Ann”(www.laerdal.com), like those nurses and paramedics training with, was used (see Figure §5.4).

5.6.1 Study Design

The CPR-watch was supposed to evaluate the effectiveness of the CPR with random people from the street. The only exclusion criterion was people physically not capable (either too young and small to perform CPR accurately or with medical conditions where performing exhausting tasks was not recommended), or persons with specific medical/resuscitation background, as nurses/-paramedics, police/firefighters and people with special and frequent First-Aid training (like First-Aid commissioners of companies).
In total 41 people were recruited. 24 of them male, 17 female and aged between 24-70 (average age 37, std 13). In the beginning, all study participants where asked when their last or only First-Aid course was. Most actually replied “during the courses to gain the driver’s license,” which was generally something between 5-35 years ago (average 16 years). Only 5 out of 41 had refreshed their First-Aid course at least once (2-25 years ago).

To record the actual performance, a professional CPR-meter (QCPR[1]) was used. CPR-meter are rather expensive devices that are designed for paramedic use during real CPR to provide instant feedback about the quality of the performed CPR (compression depth, frequency, ideal zones, release pressure, etc.), to the performer. During the data recording, the QCPR was one-way blinded. Meaning, the QCPR recorded the quality of the performed CPR, essentially the compression depth and the compression frequency, but the feedback display was covered not allowing the participants to see the feedback.

5.6.2 Study Implementation

The study was deployed in three steps. For each step every study participant was asked to perform CPR in a different modality (see also Figure §5.3):

1. **Step One: perform CPR without any additional information:** In the first step, each participant was asked to take a few minutes to think about their First-Aid course (in case they participated in one during their life) and how they remembered CPR would work, or (if they had never participated in a First-Aid course) think about what they have seen about CPR during movies, or how it was explained to them. Once the participant were ready, they were asked to perform CPR on the test manikin.

2. **Step Two: perform CPR with the assistance of the watch:** In the second step, the basics of CPR and its current regulations, and the functionality of the watch were explained to the participants. Afterward, the participants signaled they had understood and had made some try smart-watch “air CPR” attempts, they were again asked to perform the CPR, but this time with the assistance of the CPR-watch.

3. **Step Three: perform CPR with prior explanations:** First observations of the performance of the study participants showed an apparent effect of the watch and an improvement of the quality of CPR performed in the second step (with watch assistance). These effects though could have been influenced merely by the fact that in the second step, by usage of the watch, the basics of CPR were explained to the participants. Thus, it would be no surprise that the participants performed better in the second run. In order to understand what influence actually came from the CPR-watch and what just came from the explanation, the third round of data recordings were performed with the same participants, mostly a few days after the first two runs. In the third run, the participants performed CPR again without the assistance of the CPR-watch, but after a thorough introduction to CPR with explicit explanations .

Figure §5.4 shows a study participant while trying to recall how CPR works without any assistance. Figure

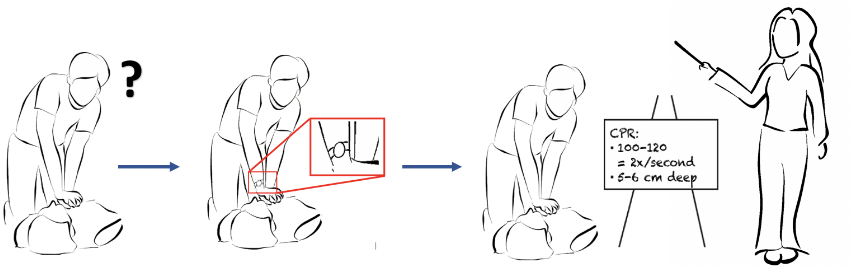


Figure 5.3: Study Implementation. Sketches by Hamraz Javaheri

§5.5 shows the CPR watch in action telling the user that the compression depth is not sufficient. In both pictures, the QCPR can (hardly) be seen peeking out beneath the hands of the CPR performer.

5.6.3 Data Set

Performing CPR is quite exhausting, one reason why professionals are tempted to switch frequently when performing CPR. Thus, to keep the exhaustion of the participants at a bearable level, but at the same have a kind of fluent CPR recording, within each run every participant performed five sets 30/2 (30 compressions, 2 rescue breaths - the rescue breaths were skipped actually, the participants were asked to make a short break inbetween the sets). So in total, every participant performed approximately 150 compressions in each run, 450 per participant in total. Overall, during the study, about 18.000 compressions were recorded by the QCPR. These are split into:

- 6000 compressions of CPR without any additional information (200 recordings of 30/2 CPR sets collected in the first run). Note, there is one data-set missing in the first step (performing CPR without any information), since one participant forgot to start the QCPR.
- More than 6200 compressions where recorded with the assistance of the CPR-watch (205 recordings of 30/2 CPR sets). The total numbers of actually performed compressions in this run was higher, since the CPR-watch app does not count ineffective compressions. Some of the participants needed a few compressions to get a feeling for the correct depths, and so performing more compressions than 30 per set.
- 5250 compressions with prior refreshing the information on how CPR should be performed according to the current standards and regulation (175 recordings of 30/2 CPR sets were collected). Unfortunately, not all of the 41 initial participants could be reached again. Thus only 35 persons were recorded.

Figure 5.4: Person trying to recall how CPR is done correctly.

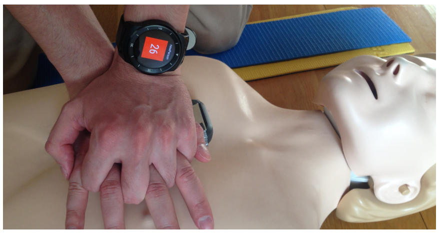


Figure 5.5: Person trying to recall how CPR is done correctly.

5.7 Results of Gaining Instant Skills in CPR

The evaluation of the recorded CPR sessions, provides clear results in favor of the CPR-assistance system. The study above shows that, when ordinary untrained people use the assistance of the CPR-watch, they perform considerably better in both essential factors (compression depth together with compression speed) than without device. Essentially, with the assistance of the watch most participants where able to gain instant skills and perform CPR effectively.

It is not explicitly surprising that most participants had problems to perform CPR correctly when simply being asked to recall from their First-Aid courses or from movies how CPR was supposed to work. The performance of several participants does increase visibly with refreshing the knowledge about the regulations and proper performances. This too, is not surprising. Nevertheless, the assistance of the watch significantly improves the performance and allows laypersons to perform CPR in a way it becomes actually effective. Detailed analyses will be provided in the following sections.

5.7.1 Person Based Analysis

Table §5.1 list how many persons deviate from the suggested intervals over of the five runs of each modality (w/o any information, with prior explanations, with the watch) and for each factor (frequency, depth). Regarding compression depth, more than 50% of the participants did not manage to get the depth right at all. With

Depth	N	run 1	run 2	run 3	run 4	run 5	mean
number (percentage) of Persons with deviation from ideal depth							
w/o information	40	25 (63%)	20 (50%)	19 (48%)	18 (45%)	21 (53%)	**21 (52%)**
prior explanation	35	18 (51%)	17 (49%)	16 (48%)	15 (43%)	16 (46%)	**16 (47%)**
with watch	41	16 (39%)	19 (46%)	16 (39%)	11 (27%)	11 (27%)	**15 (36%)**
average deviation (mm)							
w/o information	40	4.23	4.41	4.97	4.35	4.14	4.42
prior explanation	35	3.77	3.88	3.38	3.91	4.21	3.83
with watch	41	5.77	3.54	3.39	3.97	3.36	4.06

Frequency	N	run 1	run 2	run 3	run 4	run 5	mean
number (percentage) of Persons with deviation from ideal frequency							
w/o information	40	29 (73%)	28 (70%)	28 (70%)	25 (63%)	25 (63%)	**27 (68%)**
prior explanation	35	14 (40%)	14 (40%)	14 (40%)	15 (43%)	13 (37%)	**14 (40%)**
with watch	41	8 (20%)	9 (22%)	10 (24%)	8 (20%)	6 (15%)	**8 (20%)**
average deviation (cpm)							
w/o information	40	11.76	12.76	13.07	12.43	12.79	12.562
prior explanation	35	10.56	10.38	8.62	9.45	9.4	9.682
with watch	41	9.12	3.86	7.87	5.58	5.61	6.408

Table 5.1: Persons deviating from the suggested intervals in each of the five runs of each different modality (w/o any information. with prior explanations. with watch assistance) and for each factor (depth and frequency)

prior explanation still, 47% lack the feeling for the correct depth. With the assistance of the watch, only 36% of the test-persons fail to reach the correct depth. Looking at the actual value of deviation, it becomes clear that persons deviate, regardless of what modality, only around 4 mm. Here actually an effect transpires, showing that with practice the feeling for depth gets better, as the average deviation in mm is highest in the first performed modality and lowest in the last performed modality).

While the effects of the watch are decent concerning the compression depth, regarding the correct frequency, the effect is much more significant. Without any information, almost 70% of the test-persons do not manage to find the correct frequency, not even once.
After using the watch and with an additional explanation (like "remember 120 CPM, that means app. 2 compressions per second"), still, 40% do not manage to hit that frequency once. This is a definite increase in comparison to the first modality. Please note though, that for this improvement the experience of performing CPR with a correct frequency in the second modality could have had a positive impact on the performance in this third modality. Nevertheless, the watch-assistance explicitly enhances the ability of 48% of the test-subjects (in comparison to without information) to perform the compression frequency correctly. Only 20% of the test-persons are not able to find the correct frequency even with the assistance of the watch.

5.7.2 Learning Curve

Analyzing the five repetitions of each modality reveals some interesting aspects. Regarding the compression depth, neither CPR without information nor CPR with prior explanation show a clear trend with each further repetition but a slight hint that the test-persons might be getting tired at the end. See figure §5.6 (left). In both modalities, the least number of persons deviates completely form ideal depth at the fourth repetition, but at the fifth again more do. The trend for the modality of using the watch is different. Here, in the last two repetitions, distinctly fewer persons deviate than in the first three repetitions. Even though there are not enough repetitions to speak of a learning curve in using the watch, nevertheless, these trends hint at people getting better in their performance while using the watch.

Regarding the compression frequency, it seems as if some of the test-persons, during the modality without any information, got a natural feeling for the correct rhythm. In figure §5.6 (right), this trend shows. Nevertheless, this trend is not repeated in the modality with prior information. Also for the modality with the watch assistant, a trend is only vaguely noticeable.

5.7.3 Temporal Analysis

In analyzing the CPR performance of laypeople, obviously, it is not only interesting to evaluate how many people can benefit from a watch assistant. It is also interesting to know how much percentage of the time those that are able to perform CPR correctly occasionally, actually perform CPR correctly. Table §5.2 provides these results in detail:
In the first modality, CPR without having any further information, on average the test-subjects were able to keep the ideal frequency only 19.78 % of the time (std. 33.7). When using the watch assistant, the time performing in ideal frequency increases to 61.3% which is an increase of more than 200%! The ideal depth even without any information is better than the frequency results, with a correct depth in 48.7 % of the time (std. 25.8), but with less than 50% not sufficient enough. Using the smart-watch for assistance, results in an increasing ideal compression depth by more than 30-65% of the time.

The third modality, which is also the third time the test persons performed CPR and the second time they performed CPR without help but with extensive prior information about correct CPR, does only slightly improve the result. On average the test-subjects performed an ideal compression depth 45% of the time, and at an ideal frequency at 44% of the time. To recall, performance with-

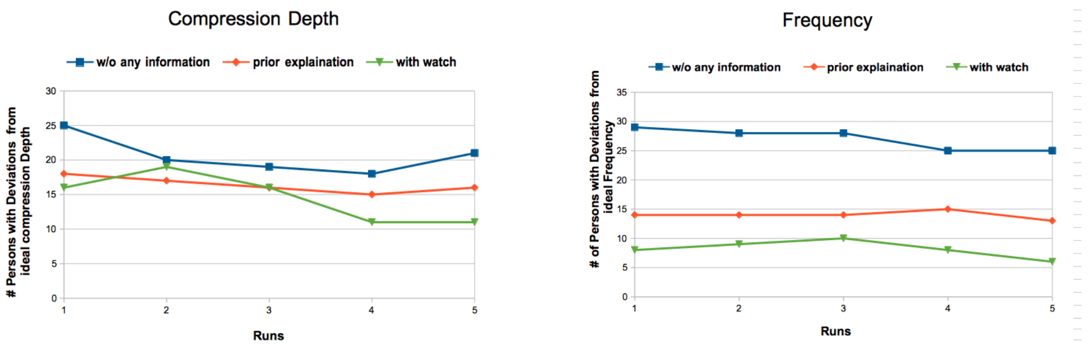


Figure 5.6: Number of persons deviating from the ideal depth (left) or ideal frequency (right) in course of the 5 repetitions. In compression depth "with watch" a learning curve seams tangible.

out explanation was 48% (depth) and 20% (frequency). Thus, for compression depth, the test-persons even performed slightly better without information. This effect could be due to remembering from First-Aid courses that "you have to push hard", while 5 cm do not seem to be so much depth. This analysis also shows that even extensive prior explanations and, not to forget, the experience of performing CPR in two previous sessions (with and without watch) cannot substitute the effect the watch-assistance can provide. The last three columns in Table §5.2 detail these improvements between the different modalities.

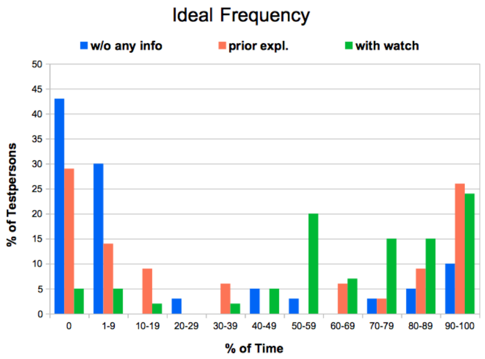


Figure 5.8: The percentage of how much time spent in ideal frequency (100-120 cpm) for the three modalities.

5.7.4 Ideal Depth or Ideal Frequency

All results indicate that correct compression depth is easier to achieve than correct frequency. Generally, it seems that people tend to compress too hard rather than compress too little. Even without help people have a basic understanding that chest compressions are "hard" to achieve and thus try to put some effort in compressing the chest effectively. Figure §5.7 (right) shows the compression depth the participants reach in the three different modalities. More than 50% of participants can reach a depth of 50-60 mm for more than 50% of the time even without receiving information about how deep to compress. Still, the usage of the CPR-watch even can improve this performance. With watch assistance, more than 70% of the test-subjects manage ideal depth for more than 50% of the time (with 50% of the test-subjects reaching effective depth for 75% of the time).

The effectiveness of the compression frequency seems to be more influenced by the watch. Thus it is worth digging into the frequency more deeply. Figure §5.7 (left) provides a detailed overview when CPR worked for how many persons. Without assistance, most test-subjects are either too slow or too fast. 75% of them are not able to stay at the ideal frequency of 100-120 CPM for more than 10% of the time. This changes explicitly with the assistance of the watch to approximately 80% of the test-persons being able to keep the rhythm and stay in the ideal frequency for more than 50% of the time. This is an improvement of 60 percentage points. More than 50% or the test-persons even manage to keep an ideal frequency for more than 75% of the time!

An even closer look at the best frequency performances (see Figure §5.8) for the three modalities actually reveals that without help people are either very bad in finding the correct frequency (more than 70%) or very good (app. 15%). Only a few persons hit the correct frequency sometimes. Additional verbal information explained before the third run shifts this to some extent, meaning that fewer people are terrible and some more are very good, but there is still no real middle field. In this regard, both modalities show a kind of inverted Gaus curve.

The watch on the other hand can help those specifically, who are bad at keeping a suitable frequency. In Figure §5.7 (left) the performance for those who were bad before, is moved beyond the center, which is a real improvement of more than 50%. Furthermore, it can be seen that more than 80% of all participants are able to maintain a correct frequency for at least (or more than) 50% of the time.

5.7.5 Effective CPR

Effective CPR does not only mean to get one of the main factors right, but both have to be performed correctly at the same time in order to be effective (effective performance). Thus, this effective performance should be analyzed for all participants for all of the different modalities. In table §5.2 the last two rows summarize the average of how many people manage to provide adequate performance (last row) and how much time they can keep the effective performance going (second last row).

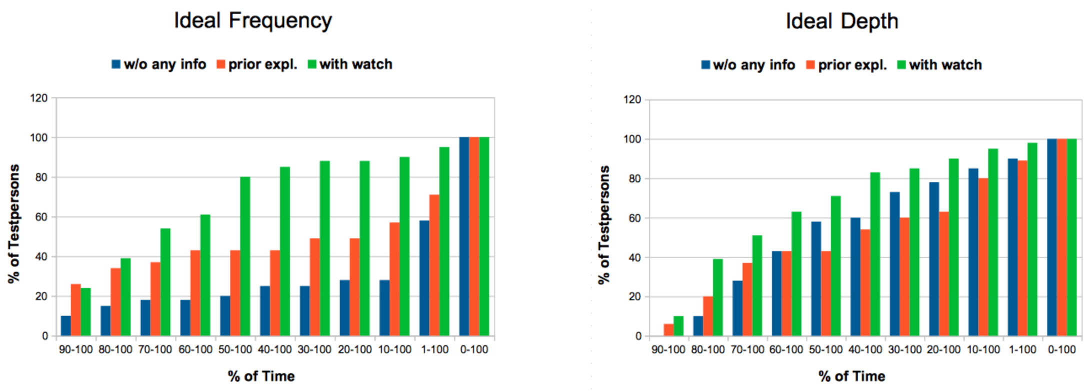


Figure 5.7: The percentage of time spent in compression depth (right) or frequency (left) for the three modalities.

	N = 40	N=35	N= 41		Improvement	
	w/o any info	**prior explanation**	**watch**	w/o anything -> watch	prior explanation -> watch	w/o anything -> prior explanation
average depth in numbers	60.49	61.66	59.76			
ideal depth 50-60mm	**48.31%**	**45.15%**	**65.01%**	**34.56%**	**43.98%**	**-6.54%**
too shallow < 50 mm	*21.99%*	*12.46%*	*17.16%*	*28.15%*	*-27.41%*	*-43.35%*
too deep > 60 mm	*32.56%*	*42.74%*	*20.32%*	*60.22%*	*110.34%*	*31.28%*
average frequency in numbers	102.12	107.05	104.40			
ideal frequency 100-120cpm	**19.78%**	**43.91%**	**61.31%**	**209.96%**	**39.63%**	**121.98%**
too slow < 100 cpm	*51.70%*	*31.62%*	*32.13%*	*60.91%*	*-1.58%*	*-38,83%*
too fast > 120 cpm	*29.92%*	*26.10%*	*9.32%*	*221.03%*	*180.08%*	*-12.76%*
ideal (depth+freq) 50-60mm + 100-120cpm	**20.14%**	**29%**	**52.14%**	**160.4%**	**80.8%**	**44.0%**
persons in ideal % out of total	57.5%	80%	95.1%	65.4%	18.9%	39.1%

Table 5.2: Total time participants were (in)correctly performing compression depth or frequency individually, CPR effectively, and how many participants were actually able to achieve this (columns 3-5). Improvement of total time participants were (in)correctly performing CPR in the three modalities, for compression depth and frequency individually and combined (columns 7-9).

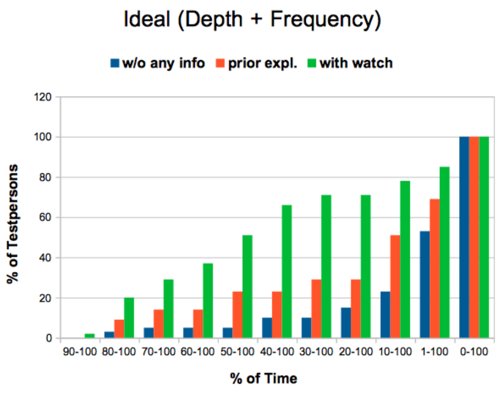


Figure 5.9: Percentage of correct CPR (frequency and depth) for the three modalities.

Without having any additional information available, which is basically like any emergency scenario, where it cannot be expected to have explanations at hand, let alone time to browse the Internet for “how to act in an emergency”, only an average of 57% of all test-subjects managed to perform CPR effectively at least for a short time. However, still, even those 57% who did, could keep it up for only 20% of the time or less! Interestingly though, after getting additional information (and two CPR sessions to gain experience) certainly more people (23.5 % more) manage to reach a sufficient performance. Still, they do not manage to keep it up for much longer (only 29% of the time) than during the first attempts (20%).

A glance at the data of using the watch assistant shows that it helps to increase the overall performance (see also Figure §5.9). With watch-assistance, less than 5% did not manage to find an effective rhythm and depth, and those 95% who did, could keep it going for more than 50% of the time. This is a very distinct improvement in regards to both of the other modalities.
A detailed look reveals some interesting further details. Without any additional information, more than 70% of all test-subjects were not even able to reach the effective range for at least 10% (see Figure §5.9). 48% of the participants were not even able to find the ideal range at all, not even for a few compressions. Moreover, only barely 5% were able to keep staying in the ideal effective range for at least 50% of the time!

Getting an introduction on how CPR has to be performed did improve the values slightly. Still, almost 50% of the test-subjects perform poorly (time in the ideal range for less than 10% of the time) and still, 30% did not manage at all. 14% were able to stay in the ideal range for most of the time, which is almost three times as many. Again, the assistance of the CPR watch has a much more significant impact. Only 15% (6 out of 41) of the test-subjects failed entirely in reaching the ideal range (compared to 48% and 30% respectively). Also, more than 50% of all test-persons could keep a sufficient performance going for more than 50% of the time. 29% even achieved effective performance for more than 75% of the time. In total an improvement of more than 45 percentage points (pp)!

	positive			neutral	negative		
Questions	*absolutely*	*yes*	**avg.**	*neutral*	*no*	*not at all*	**avg.**
Is the topic of the study (bystander CPR) relevant?	75.0%	25.0%	**100%**	0.0%	0.0%	0.0%	**0.0%**
Could a Live-Feedback System help saving lives?	32.1%	60.7%	**92.8%**	7.2%	0.0%	0.0%	**0.0%**
Could such a system help to reduce fear of doing damage?	35.7%	53.6%	**89.3%**	3.6%	7.1%	0.0%	**7.1%**
	very secure	*secure*		*neutral*	*insecure*	*very insecure*	
How secure were you about CPR before the study?	3.6%	32.1%	**35.7%**	25.0%	35.7%	3.6%	**39.3%**
Did the watch help you to feel more secure?	35.7%	53.6%	**89.%3**	7.1%	3.6%	0.0%	**3.6%**
Did the watch help you to perform CPR better?	46.4%	46.4%	**92.8%**	7.1%	0.0%	0.0%	**0.0%**
Did the watch irritate you while performing CPR	0.0%	3.6%	**3.6%**	7.1%	39.3%	50.0%	**89.3%**
Would you install this App if you had a smart-watch?	35.7%	39.3%	**75%**	10.7%	14.3%	0.0%	**14.3%**

Table 5.3: Participants' Feedback. The relevance of the topic is clear to all, and in most questions the replies are quite in unison as most favor using the watch for assistance.

5.8 Shaping Confidence

One of the explicitly stated goal of this chapter was to promote more self-confidence in acting in an a medical emergency. In this regard, since self-confidence is a subjective feeling, the effect of the CPR-assistant devices on the self-confidence of the participants could not be captured with the sensor readings. Thus, to evaluate how the CPR-watch would influence the confidence of the users, a questionnaire was handed to the study participants after they had finished the third run. [1]

In total 30 questionnaires were handed out. For different reasons it was not possible to hand the questionnaire to all 42 participants. Out of these 30, 28 were filled-in and returned. Which is a return rate 93%. The questionnaire included questions about the participants' (subjective) self-rating of their ability to perform CPR at the beginning of the study and how they improved. Other questions included were also about the participants' confidence in performing CPR without assistance and how they perceived the watch to influence it. See Table §5.3 for all questions.

100% of study-participants rated the topic of the study to be positive (either very important or important), with 75% presuming it to be very important. Also the vast majority was certain that a live-feedback system like the watch-app could help to save more lives (93% positive rating). Only 7% were neutral in this question, no one actually doubted the potential of the CPR-watch to support saving lives.

Only 7% doubted that a system like the CPR-watch would help to lift the fear of doing damage in performing CPR and promote more confidence in laypersons. 89% were confident that the watch would give them more confidence in doing CPR correctly.

Being asked about their self-perception of knowing how to perform CPR correctly, the replies were more divergent. 35% were actually quite sure how to perform CPR at the beginning of the study. 39% were not secure about CPR, and 25% had not idea if they were or were not sure (likely had never really thought about CPR in recent years). Despite the initial perception, almost 90% stated that the watch did give them more confidence in performing CPR. Only 4% did not believe the watch helped them to feel more secure.

All of the participants rated using the CPR-watch as CPR assistant positively (93%) or neutral (7%), none of the participants had a negative attitude towards the usage of such an app. A small minority (4%) of the test-subjects (namely those who generally do not wear watches on a daily basis) stated that the usage of the watch was kind of unfamiliar. However, the vast majority (89%) were explicitly positive about using a watch. In regards of the practical deployment of a smart-watch app as a CPR support, three third of all participants would immediately install such an app on their smart-watch (provided they would own a smart-watch).

The feedback of the study participants in general backs the initial assumption of this chapter that appropriate sensor based feedback devices would be able to influence cognitive well-being in a positive way.

[1] Parts of the text of this section have been taken from following publications of the author of this thesis. Any text-passages taken from these papers have been written solely by the author of this thesis:
Gruenerbl A. et al. (2015), [136], please refer to the respective entries in the literature list or at the beginning of this chapter

5.9 Support Effective Learning - Training CPR

After the positive outcome of the bystander evaluation, a second question to be answered was: since people can perform CPR correctly with the help of an assistance device, does this instant-feedback, such a device can offer, sticks? Thus, in a second study, it should be evaluated whether people would be able to train CPR effectively when receiving instant-feedback in comparison to traditional training classes.

This part of the chapter compares the effect of the traditional teaching of CPR (in the following called teaching) with the effect of training with different wearable instant feedback devices, like the CPR-watch introduced above (called training). A study with 50 test persons (23 nurse students, 27 novices) was conducted where both, the order of training or teaching first and the device (smart-watch or smart-glass) to be used was randomly selected. The results, indicating the clear superiority of the device training, are evaluated with an ANOVA analysis.

5.9.1 Background

A typical standard CPR classroom teaching session, like in First-Aid training lessons or in the way nurse students will have to attend repetitively over the course of their education, looks as follows: First, a teacher will provide the theory about CPR to a group of 10-15 students. This will include why and when to apply CPR, what CPR means, which effects CPR has, etc. After the theory part, the teacher will give some practical demonstrations with the help of dedicated training manikins.

After this, the students will, either individually or in small groups, (attempt to) perform CPR themselves on the training manikins. Meanwhile, the teacher will observe. In case necessary, the teacher will provide input to the students and give hints and feedback. Maybe, the training sessions even will be videotaped and analyzed afterward.

Another possible teaching scenario is that each student gets a time-slot to perform CPR while the rest of the group is watching. This way, the "instant" feedback the teacher will give during these attempts, can be a benefit to all students. In a way, this modality comes close to a kind instant feedback. Nevertheless, the feedback a human can provide always is vague. It can be in the manner of saying "you have to go deeper!" or "slow down, you are too fast!" However, without the help of a devices, a teacher will not be able to say "you are five compressions per minute slow!" or "you are now slightly beyond 60mm!" Devices that are sophisticated enough to give such detailed feedback, are not yet widely available.

5.9.2 CPR Assistant Devices

In the course of and subsequent to the first study, some potential future users of a smart-watch instant CPR feedback application were asked to test the smart-watch CPR app. Instantly it transpired that, though many people loved the watch and its natural and intuitive way of usage, some persons had problems to use it. A few people particularly had issues in grasping the different colors combined with the blinking display. Others stated to have problems feeling the vibration of the watch and could not connect to the blinking of the display.

Since feedback showed that the main issues with the watch-application were the visual aspects of it (different color, blinking), it was decided to add an alternative system, implemented on a smart-glass (Google Glass). Other feedback hinted that the backward counting feature was more distracting than helping and real frequency feedback (not only a metronome) would be desirable. Thus the watch-app was adjusted accordingly. Both devices are intuitive and straightforward to use. Short explanations about the meaning of color for the CPR-watch and meaning of numbers for the Glass suffice for a user to be able to use them effectively. The two applications for instant feedback were [2]:

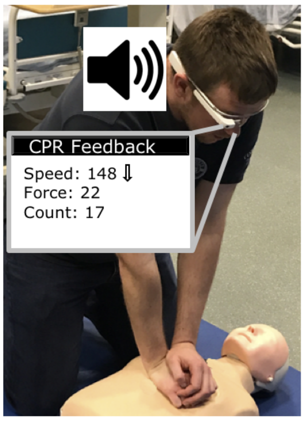


Figure 5.10: CPR Glass Assistant is providing instant feedback on compression depth (force) and speed while performing CPR

- **Smart-Glass Instant Feedback Application** The feedback application of the smart-glass combines audio output (metronome) and visual feedback (see also Figure §5.10). When the app on the Google Glass is started the Glass begins to click with 120 clicks per minute, which denotes the upper border of the recommended compression rate of 100-120 compressions per minute. While performing CPR, the display of the Glass provides the current compression frequency (based on 3-axial acceleration), and (if necessary) prompts the user to slow down or speed up with upwards or downwards pointing arrows next to the compression frequency. The current depth is also shown in the display.
- **smart-watch Instant Feedback Application** The instant feedback application combines, like the initial version (see above) haptic information and visual cues

[2] Both app, the watch app and the Glass app were initially implemented by Gerald Pirkl, but for this particular study were adapted by the author of this thesis

(see also Figure §5.11). When the app is started the watch begins to vibrate and blink (black/blue) with 110 CPM, which denote the average frequency of the recommended compression rate of 100-120 compressions per minute. While performing CPR, the watch displays the actual compression frequency in the center of the display. (This is the main change to the version used in the first study)

The compression depth feedback is provided in color like in the first version. The center square in the display is green for good compression depth (50-60 mm), turns yellow if the compression depth is beyond 60 mm and turns red if the compression depth is not deep enough.

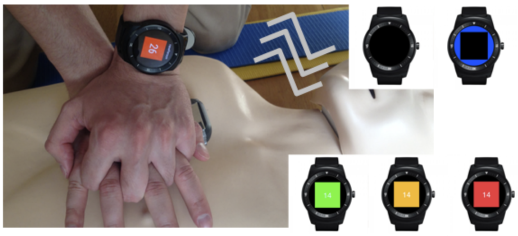


Figure 5.11: CPR Watch Assistant is providing frequency instructions and instant feedback on compression depth and speed while performing CPR!

5.9.3 Evaluation

The influence of training with instant feedback devices should be evaluated in a randomized, prospective simulation study. This study was designed to assess the CPR performance of nurse candidates and novices alike by comparing the effect of CPR teaching in the standard way versus the impact of training CPR with an instant feedback device.

Study Group: A total of 50 volunteers participated. 23 of them were chosen among first-year nurse students who already had a basic understanding of what CPR means, but had not yet trained CPR intensively. The nurse students were recruited at the Health Department at the University of Southampton. The other 27 participants were laypeople, who had absolutely no medical background or experience in CPR (except for possible First-Aid courses during gaining the driver's license, as obligatory in many European countries). They were mainly students recruited in the Technical University of KL and the German Research Center for Artificial Intelligence.

Study Implementation; In the course of the study, the participants were randomly distributed into two groups. The first group would get a CPR teaching session first and afterward would train with one of the devices. The second group would first train with one of the devices and would receive a thorough CPR lesson afterward. Figure §5.12 graphically explains the study procedure. For both groups, it was also randomly decided which participant would use which device. To obtain even and fair groups, randomization was promoted to distribute nurse students and not medicals as evenly as possible between both groups (teaching or training first) and between both devices. To capture the CPR performance of each participant and their improvement (or worsening) after teaching and training, baseline recordings (measurement points) were done in the beginning, after each session and at the end of study (see figure §5.12). The measurement points were obtained by using a standard CPR training manikin, which was equipped with a pressure sensor beneath the "chest skin" of the manikin's chest (not visible from the outside). The pressure recorded on the manikin was calibrated and aligned with one of the aforementioned professional CPR recording devices.

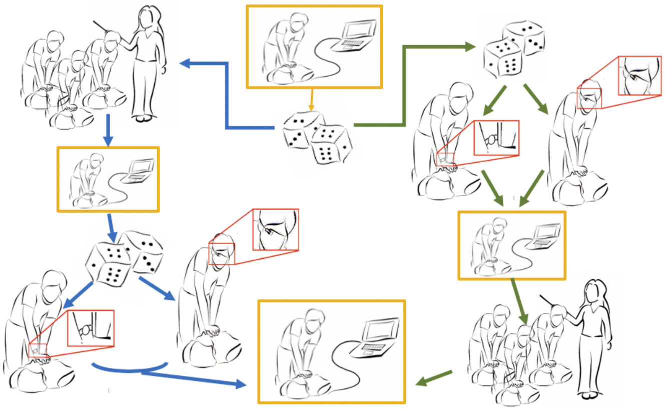


Figure 5.12: Study design: after first baseline recording it is randomly decided if teaching or training comes first. For training it is also randomly decided which device is used. After teaching or training another baseline is recorded (without assistance) and the modality is switched. At the end of the study, the third baseline is recorded. Drawing by Hamraz Javaheri

Data Set: During each data recording (each measurement point) every participant was asked to perform three cycles of 30 compressions with a short few seconds break in-between. The 30 compressions cycle rhythm was chosen since current regulations of performing CPR for nurses still teach to perform a 30/2 (30 compressions, two breaths) rhythm. Therefore, in total, we recorded a data-set with 90 compressions per person per measurement point (8100 compressions were recorded in total). Thus evaluations of teaching vs. training first, as well as Watch vs. Glass are based on 2700 compressions in each group. The evaluation of the general impact of training and teaching is based on a data-set of 5400 recorded compression in each group.

5.9.4 Results

The comparison of the overall effect (regardless of which was applied first) of a teaching lesson onto the performance of effective CPR (depth + frequency correct) versus device training is shown in table §5.4. Improvement of teaching is less than nine percentage points (pp), while the improvement of training is almost 25pp, which is clearly in favor of device training.

The only aspect, teaching overall has an improving im-

	before		after		improvement	p-value	f-value	f-crit
	average	stand. dev.	average	stand. dev.	average			
TEACHING			N = 50					95%
% **effective CPR**	**34.95**	34.3	**42.59**	32.7	**7.64**	**0.26**	**1.30**	
% correct depth	54.03	42.7	68.00	36.3	13.97	0.08	3.10	3.94
% correct speed	52.94	38.3	62.83	33.0	9.35	0.19	1.71	
TRAINING			N = 50					
% **effective CPR**	**30.43**	28.2	**55.37**	36.1	**24.94**	**0.00**	**14.85**	
% correct depth	58.30	39.5	74.90	36.2	16.60	0.03	4.80	3.94
% correct speed	47.81	34.3	71.37	33.9	23.57	0.00	11.93	
comparison of baseline	Teaching		Training		difference			95%
% effective CPR	33.94	34.5	30.43	28.2	-3.51	0.58	0.31	3.94

Table 5.4: A comparison of the overall effect of a teaching lesson and of a training session with any device on the CPR performance. Results show clear superiority of the training with feedback devices.

pact of more than 10pp it the compression depth. In all other aspects, specifically frequency, but also the effectiveness of CPR, the training has a far more significant impact (around 24pp). To evaluate the quality of the improvement, a single-factor ANOVA was performed on each before and after the data-set of teaching and training. The results of the ANOVA clearly show that the effect of training is significant, while the effect of teaching does not reach the 95% confidence.

Teaching First vs. Training First

However, the above analysis compares the overall effect of teaching versus train regardless of which learning method was applied first. In order to better understand whether the order of teaching/training has an impact, each group was also analyzed individually. See table §5.5. For this analysis, the results were split according to whether traditional teaching or training with the device was done first, as it influences the start condition. Table §5.5 also separates between the results of the nurse students and the novices.
It can be seen that device-training improves the performance of effective CPR significantly of approximately 20pp or even more (with an advantage when administered after the teaching lesson). A teaching lesson, on the other hand, can only influence the performance positively when it comes first. When teaching comes after the device-training, it has little to no impact.

A glance at the performance at the end of the day though, clearly shows that a teaching lesson that provides all relevant information, boosts the effect of the following training (total improvement of teaching + training is 44.5pp!), while after training teaching has no effect (total improvement of training + teaching is 24.4pp). This effect makes sense in the way that both devices cannot explain how to perform CPR correctly (e.g.,correct posture, where to apply the pressure, etc.) which is done in the teaching session. On the other hand, a teaching lesson after the user has gained some muscle memory in training prompts the trainee to over-think the new information instead of trusting the skill achieved.
Figure §5.13 additionally shows how many persons in particular improved their skills. This figure compares how many persons manage to improve or worsen their skills by at least 5% or more, with teaching versus device training. Again, this analysis is in favor of the devices, and confirms that more persons were able to improve their skills with the help of the device training. 30% improved with training, only 23% improved with teaching. 5% worsened during training, while 11% did with teaching.

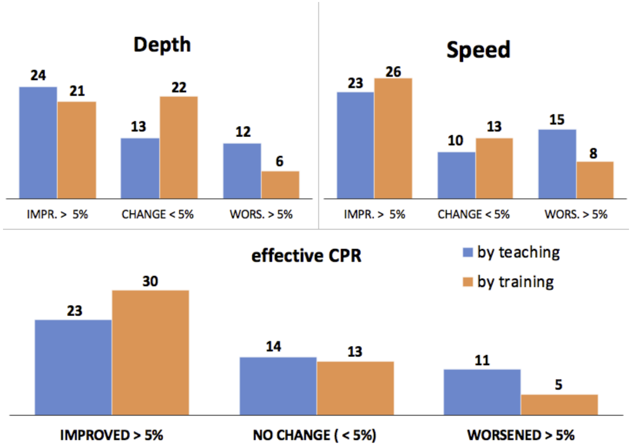


Figure 5.13: The effects of training and teaching on the improvement of performing CPR.

Watch or Glass

Finally, the question remains, which of the devices would have a more substantial impact on the CPR performance. Direct feedback from students during initial earlier tests, did not favor one of them (a reason why both devices were used).
Table §5.6) summarizes the most important aspects. The results of the evaluation also do not clearly favor one of the devices. In total numbers, the group using the Google-Glass sightly performs better after training (approximately 2.5pp), nevertheless is less effective in regards of compression depth. The watch group shows a higher improvement in terms of compression depth (10pp). Glass again is very slightly better in terms of frequency (1.5pp). Essentially the question Glass or Watch is a matter of preference. Nevertheless, during the study an interesting fact revealed in discussions after the study. Some test persons stated that they did not really like the device. However, evaluations of the performance highlighted that even people with a device they were not very

TEACHING first									
	nurses		novices		total				
effective CPR	avgerage %	stand. dev.	avgerage %	stand. dev.	avgerage %	stand. dev.	p-value	f-value	f-crit (95%C)
Base	32.54	29.20	19.47	22.59	23.14	25.55			
after Teaching	37.94	28.54	38.47	27.12	37.76	27.94			
improvement by teaching	**5.40**		**19.01**		**14.62**		**0.065**	**3.579**	4.043
after Training	63.95	42.70	68.24	26.42	65.21	30.36			
improvement by training	**26.01**		**29.77**		**27.45**		**0.002**	**10.625**	4.043
base - end of day									
imporovment	**31.42**		**48.78**		**42.07**		**0.000**	**27.673**	4.043
TRAINING First									
	nurses		novices		total				
effective CPR	avgerage %	stand. dev.	avgerage %	stand. dev.	avgerage %	stand. dev.	p-value	f-value	f-crit (95%C)
base	32.42	26.38	14.50	26.43	23.82	26.86			
after Training	51.75	35.60	38.11	41.48	45.20	37.58			
improvement by training	**19.33**		**23.61**		**21.38**		**0.028**	**5.144**	4.043
after Teaching	58.21	38.21	37.34	32.87	48.19	35.84			
improvement by teaching	**6.46**		**-0.77**		**2.99**		**0.783**	**0.077**	4.043
base - end of day									
imporovment	25.79		22.84		24.37		0.011	7.062	4.043

Table 5.5: Effects of teaching first (top) vs training first (lower half) on CPR performance, before and after teaching and/or training, and the overall improvement at end of day.

	before		after		improvement	p-value	f-value	f-crit
	average	stand. dev.	average	stand. dev.	average			
WATCH			N = 24					95%
% effective CPR	27.14	27.2	53.12	36.4	25.98	0.01	7.84	4.05
% correct depth	42.33	42.0	70.57	36.4	28.24	0.02	6.17	
% correct speed	40.86	30.1	64.32	34.5	23.46	0.02	6.30	
GLASS			N = 26					
% effective CPR	32.53	31.0	60.65	35.1	28.12	0.00	9.38	4.03
% correct depth	47.53	40.5	65.63	40.0	18.10	0.11	2.63	
% correct speed	58.44	35.4	83.44	25.4	25.00	0.01	8.57	

Table 5.6: Watch vs. Glass. The comparison of the effect between the Glass and the Watch group shows an advantage of the watch regarding compression depth.

comfortable with did improve in their ability to perform effective CPR!

5.9.5 Discussion of Results

It is not particularly surprising that one short teaching session of a, in its nature, rather complex motion as performing CPR correctly, fails to lead to immediate significant improvement, in effective CPR. In reality, nurse students and other medical and paramedical professionals will follow such a training session with repeated extensive practice and further teaching sessions, followed by more practice. On the other hand, even a short individual training session with a real-time wearable feedback device can brag with significant results. This is a very clear indication of the potential such a tool can bring to teaching and practicing of CPR.

In analyzing the results it becomes clear that there is a mismatch between the effect of traditional teaching on individual aspects like compression depth where teaching could help to improve to some extent, and on the overall effectiveness where teaching could not help very much. A possible explanation is that the motor coordination of the CPR task is complex and requires two different aspects to be aligned perfectly. It is much easier to manage to get one of both aspect right, e.g., maintaining the appropriate speed (particularly for persons with a good feeling for rhythm) OR the required depth. Either one of them individually can be reached via concentration and focusing on the particular motion. Coordinating both in the way they are supposed to be performed is much more complex and more difficult to do, specifically without someone or something telling us if they are right. Thus getting both right requires an understanding of how “getting both right” feels, which means this requires a kind of muscle memory or proper real-time feedback.

Nevertheless, the results, specifically the comparison of traditional teaching first or device-training first also shows that the training devices work optimally as an addition to human teaching. Trainees using the device for training after a teaching lesson improved their overall performance by almost 45% in comparison to trainees who trained first and later received teaching. Those only improved overall by 24% (basically due to device training). In all honesty, this again is not really surprising. This effect makes sense since both devices (Watch and Glass) lack the ability to explain how to perform CPR correctly (correct posture, where to apply the pressure, etc.). Without this information trainees could, in theory, do it wrong (e.g., apply compression to the abdomen) even though depth and frequency were correct, or when the body posture is not optimal, CPR is extensively more exhausting and when getting tired even the support of the devices has its limits.

Even though a single teaching lesson might not have a very lingering effect on the CPR performance, still it is able to provide the relevant information, which, in turn, will give the following training with assistance device an optimal foundation. Insofar, this effect, discussed above, naturally suggests to use both methods in correct order. Start with a human teaching lesson to provide the nurses-to-be with all relevant information and afterward let training with device take over. Another option that comes to mind when considering that the so far tested feedback devices are coming short in providing essential information, is to use a device that actually is able to do so. Since recent developments have brought a number of augmented reality device to the market (e.g., the Microsoft HoloLens) such devices could in fact combine both requirements, to provide information (e.g., in form of a virtual teacher) while at the same being able to provide instant feedback.

5.10 Conclusion and Outlook

This chapter mainly introduced a smart-watch based instant feedback system that should allow laypersons to perform heart chest compressions correctly without advanced training. The basic idea of the system is to provide instant and live feedback to the attempt to perform CPR and thus allow the user to adapt immediately and improve the performance within minutes to point where CPR is effective. In this line of thought, the system is meant to shape the skills in performing CPR of any person on the fly, and thus, boost confidence, while at the same allow to learn instantly.

The system itself is rather simple and is comprised mainly of a metronome and acceleration based estimation of compression. Further, the measuring of compression depth is solely based on the magnitude of the three-axial acceleration data and a simple algorithm that was calibrated in comparison to a professional CPR device. The actual correctness and precision of the CPR feedback was not evaluated in detail and since the feedback was only relying on acceleration, the smart-watch CPR app, as it was used, would not suffice to measure the exact compression depth.
Nevertheless, the aim of this chapter explicitly was not to measure how well yet another device works, but to evaluate whether a device (as coarse as the values might be) was able to support skill training and promote more confidence. Given the potential imprecision of the device, the results in this regard speak for themselves.
In every analyzed aspect, the assistance of the Watch (or Glass) lead to a far better performance than even a detailed instruction could. As was evaluated in the second study, the assistant devices even outmaneuvered a professional teaching session. Specifically Figure §5.9 and Table §5.4 display the impressive results of using the assistant devices. Figure §5.9 highlights the positive impact of the watch in comparison to performing CPR as might be recalled from courses years ago, or even to performing CPR after having received a detailed explanation on how to. Some of the study participants of the instant skill study (first study) had never had any contact with CPR other than what they saw in movies. Still, even these persons managed to improve significantly with the help of the Watch.

The results of the instant skill study can be summarized as: **more than 50% of the participants were able to provide effective CPR for more than 50% of the time** (in comparison to: less than 5% were able to provide effective CPR without any information and only 25% were able to turn detailed instructions into actually effective CPR). Concerning the instant feedback device training versus a standard CPR teaching lesson, even nurse students who have a basic understanding of CPR could improve significantly better by using a feedback device rather than in a teaching session. Tables §5.4 and §5.5 demonstrate these results. Summarized, in a **teaching lesson the candidates improved the effectiveness of their CPR performance on average around 8%** while in a **training session with one of the feedback devices, the candidates improved their effectiveness in CPR by around 25%!**

On the positive side, it also can be added that these results generally show that effective tools do not need to be highly complex or sophisticated, but oftentimes a simple (and maybe imprecise) system can suffice to provide adequate support. It also shows that such simple systems can have a tremendous impact on the way people perform or can perform. Moreover, as the feedback provided in the questionnaires has stated, systems like the CPR-watch app can influence the confidence of people. Practically, such a system would be worth to be brought to the actual user. Nevertheless, in reality for such systems, potential liability issues have to be taken seriously. Therefore, future efforts should be put in bringing this app into the App Stores.

5.10.1 Outlook

In terms of future research though, one of the main questions that remain is, whether personalized and more elaborate functionalities (e.g., feedback on the compression frequency, information about performance, and more) would strengthen the usability. Further questions like: "could such a system also be leveraged for health professionals? Could nurses benefit from a CPR-assistant?", alternatively, "How could the public benefit from such systems" should be examined in the future. The work described in this chapter only depicts the start of various possible applications, and I hope that this work has been able to lay the basis for a new generation of possible resuscitation support apps for both health-care professionals and laypeople.

Furthermore, in the future, research should also focus on long-term effects. The presented studies have not evaluated how long the learning effect actually lasts (both in teaching and in training)? Moreover, what are the optimal numbers of training sessions to have a long-lasting effect. Also, it might be worth looking into improving the feedback delivery. Since there has not been a clear winner between Watch and Glass (some people visually preferred the one or the other) better and individualized representation could probably combine the particular aspects of each device that users preferred in their favorite, into one device that fits the needs of the majority of people. Certainly, looking into possible HoloLens adaptations of the CPR-feedback assistants as has been suggested in the discussion should be a future endeavor.

During several studies and various discussions with health care professionals new ideas were developed. One of them is to combine the CPR-watch with other medically used devices. A particular devise, which is being used publicly and that would benefit from an interaction with the CPR-watch is the automatic public defibrillator (AED). AEDs are very sophisticated and can guide inexperienced laypersons through all necessary steps in an emergency. Nevertheless, even AEDs are lacking the ability to detect if CPR is performed, and if so, whether CPR is performed effectively, while the AED is charging. Charging commonly takes about 2 minutes in which effective chest-compressions are essential for the patients survival. If an AED was connected to a device like the CPR-watch, it could detect that CPR is performed and performed effectively and if necessary give appropriate instructions.

In the medical field, the CPR-feedback devices have received substantial praise. Many nurse-students, nurses, doctors, but also representatives of regulatory bodies (e.g., ERC members) have expressed their interest in the CPR-watch.

Evaluating Group Behavior: Detecting Collaboration in unscripted Ad-hoc Groups during Emergency Care Situations

So far, work in this thesis has focused on recognizing an individual's cognitive state or helping individuals in stressful situations. However, stress or cognitive health do not just depend on a single person or affect only individuals. Often, people in a group need to work together on cognitively strenuous activities or in an emotionally pressurized situation. An example may be an accident on the street, calling a random group of strangers to save the life of one or more persons. Studies [140] have shown, as mentioned in the previous chapters, that emergencies at home are the most often cause of death, as persons who face emergencies of their loved ones are often unable to act. At the same time, random people in public areas are most successful in implementing emergency measures. In this context, it is of particular interest to understand what constitutes such group dynamics. How do these random groups interact naturally and what can be done to help people in these groups reach the goal most effectively.

As in this dissertation, general activity and context recognition research has focused solely on identifying what individuals are doing and how they interact with their environment but not with each other (e.g., physical activity, device interaction, etc). Lara et al. [11] provide an overview of state of the art in detecting human activity with portable sensors. In addition, so far, the presence of multiple users could have a negative impact on activity detection of a single person. As Gordon et al. [169] elaborate, this was often viewed as an annoying disruption. The "Multiple occupancy problem" [170] is one of these examples. Now, as the work in this chapter aims to understand the interactions of people in a group, it has to go in a different direction. It will follow a recent trend in pervasive computing and seek to move from the classic "single-user single-system" view to a more comprehensive "system-ensemble-user-collective view".

Based on a real-world emergency care scenario, this chapter examines how multiple users' presence can be used to support rather than disrupt individual activity recognition through novel collaborative approaches. The main focus of this work is on multiple people who perform various, mainly physical, activities in smart environments. Moreover, not only do these people work in the same environment, but some of them may work together as a group, while others work individually or some interact within certain activities while others do not. The central assumption of this application is that there is no a-priori knowledge about which people belong to (a) collaborating group(s) and what the structure and role distribution within the group might look like. Also, the structure and role allocation can change dynamically and evolve with time. Besides, groups may have complex hierarchies, as is the case with naturally-growing groups, and it is possible for individuals to contribute to different groups at the same time. It is also possible that people intentionally try to avoid collaboration or interaction.

This chapter will present a new method for detecting collaboration in dynamic groups by mapping sequences of low-level atomic actions, performed by individual persons, onto collaboration patterns concerning high-level compound activities. The method is evaluated with observations extracted from video footage of real emergency care training sessions of soon-to-be nurses and reaches collaboration recognition of up to 95% accuracy.

The work, contents, all pictures, most tables and also partially text-passages of this Chapter have been published by the author of this thesis in the following publications. Any text in this chapter, specifically text from paragraphs taken from these publications has been written by the author of this thesis. For more details about these publications please also refer to the entries in the literature list:

- Gruenerbl A. et al. Detecting spontaneous collaboration in dynamic group activities from noisy individual activity data. In Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on. IEEE, 279-284. 2017 [171]
- Bahle G., Gruenerbl A. et al. From Individual Activity Recognition to Unscripted Collaboration Analysis. submitted at: IMWUT Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2019.

6.1 A Real Life Motivation

In the training of future emergency nurses it is highly essential to make sure the nurses are confident in their work, know the procedures defined explicitly for handling every situation, and most importantly can act accordingly. Therefore, training schools put a focus on real-life like simulated training, where students in all semesters are confronted with carrying out nursing sessions with simulated patients, or with sophisticated training dummies in small groups of 3-5 students. The huge challenge for trainers during these sessions is to monitor the ongoing scenario and understand which students are ready to proceed and who still needs additional training/courses/lectures.

Questions such as “is there someone in the group of the nurses taking the lead?”, “do the nurses interact efficiently, do they interact at all?”, “are they doing what they actually are supposed to?”, “is there someone not ready to handle this situation? e.g., by occupying themselves during emergency sequences with trivial tasks?” etc. have to be answered in order to provide adequate measures. Such training sessions currently are time consuming, and trainers spent hours over hours repeatedly watching video taps taken from the training sessions to evaluate them and answer the above questions. Specifically, during emergency sequences where procedures are fast, and everything happens at the same time, it is almost impossible for a single person to capture all details.

6.1.1 Recognizing Team Dynamics

The recognition of activities performed by a person has been a core research topic of ubiquitous computing over many years. Examples here range from industrial assembly and maintenance through elderly care to sports and wellness. As has been pointed out, most works on activity and context recognition concentrated on recognizing the activities of an individual and how they interact with their environment and handle objects.

In scenarios like the emergency care training though, monitoring of an individual, or individually monitoring every person will fail to grasp the entirety of what is going on. Since a single person monitoring the scenario is unable to judge and evaluate what is happening, also it is not enough to follow only one person with sensors to display the entire picture. Moreover, it will often limit the accuracy of any activity recognition, since, in many human activities, several actions require more than one person to be carried out. The list of possible examples is endless.

The emergency care scenario or a surgery scenario, of course, are examples, where nurses and doctors and other professionals will have to interact and work together to save the life of a patient. In nursing, a clear focus has been set in the early to late 2000s to highlight the importance of the “competence of collaboration” in nursing practice and education [172], [173]. A study has revealed that experienced nurses spend approximately 45% of the time in collaboration with other health-care professionals (40% with other nurses, 3.5% with doctors and 1.5% with pharmacists) and only work alone about 40% of their time [174]. The numbers for young nurses are even more leaning towards more collaboration with others and less working alone.

Other examples can be construction work, moving from one apartment to another, or generally assembly of things, where a number of people might be necessary to carry a large component from a truck/transporter to the construction site/apartment. Here it will be possible as well that different groups carry components in parallel.

The list goes on. Analyzing a variety of multi-player sports is another example. How well has the football team interacted during the match? Which players are able to play in the team, who is a lonely wolf not being able to adapt to the team and thus either slowing the entire team down or just being not helpful? Specifically, in team sports, it is well known that ensembles of brilliant individual players can lose against an average team, just because the average team works together and is a very motivated collective, while in the team of brilliant players everyone wants to “shine” individually.

Nevertheless, recognizing what is going on in such scenarios not only requires detecting which person is performing which activity at the moment, but it is even more critical to detect who is interacting and collaborating with whom. Essentially, this means that we need to be able to understand and analyze group dynamics. As in team sports so in all situations where teams have to act together, even well interacting teams can “fall apart” when dramatic incidents happen (receive an unexpected goal; an altogether healthy patient with a simple broken arm goes into cardiac arrest without any warning signs; etc.). Vice versa, a poorly performing team can be energized through appropriate coaching and/or positive events. These effects can be found in all situations where teams have to act in environments where quick and dynamic reactions to (often unforeseeable) events, stress, and high physical/and or cognitive loads are needed.

The answer to the question, how well a team performs depends on a variety of factors. One of them is how well and flexible the individuals in a team can foresee the needs and adapt to the actions of others and to synchronize individual activities. Meaning to perform precisely the activity that is needed right now. Furthermore, emotional factors such as motivation and mood, which can also develop and evolve in the group dynamically, have a significant influence on the performance of a team.

The long-term vision in this regard is to be able to build multi-agent systems that can leverage context and activity recognition to analyze and understand such group dynamics. However, just like traditional context-aware support applications can influence the activity of an individual, for example, by automatically recommending the next trains when being at a train station, the goal of these multi-agent systems will be to influence the factors that make a group act well together to reach a particular goal and thus improve the overall group performance.

6.2 Related Work

Plan and goal oriented recognition has a long history in computer science and has been existing for over 40 years. Thus in plan recognition a variety of different fields and applications have developed, ranging from simple recognition of single-agent plans (both very early and more recent work) and probabilistic plan recognition, to multi-agent, team-plan, and teamwork recognition. [1]

6.2.1 Plan Recognition

From early on Plan recognition has been a necessary component of many applications, such as software help-systems [176], story understanding [177], psychological modeling [178], and natural language dialog [179]. In general, all work in plan or goal recognition can be split into two aspects: determining agent's plans/intends/-goals for coordination and collaboration in Human Activities and steering (and partially understanding) Robot activities. In early plan recognition, only goal-oriented agents existed who's activities were consistent with its knowledge base, and which formed a single plan, e.g., [180]. During the last 10-15 years, more complex and less restricted methods were developed for plan-recognition. These include attempts for recognition with only partially available plans, probabilistic plan-recognition and also multi-agent or team plan-recognition:

6.2.2 Plan Recognition with Noisy Input or Partially Observed Team Traces

Some systems like the work by Sadilek and Kautz ([181], [182]) are capable of recognizing multi-agent plans from noisy observations. By using noisy real-world GPS data, they try to solve the problem of modeling and recognizing activities that involve multiple game-related individuals while playing a variety of roles. Their model though, incorporates explicit team rules and dynamics imposed by the games' geometry and a player's motion model with probability and logical functions.
Zhou et al. [183] in their 2011 publication focus on methods of plan recognition that allow team traces that were only partially observed. To solve the underlying multi-agent plan recognition problem, they introduce a method for building candidate occurrences followed by soft and hard constraints to encode the correctness property of the team plans and eliminate the false candidates. In 2012 then Zhou et al. [184] went further to allow online recognizing of plans from incomplete observations. Online in this case means that a plan library is not required beforehand. The work is based on the availability of a set of action models.

6.2.3 Probabilistic Plan Recognition:

Saria et all [185] have developed a theoretical framework for online probabilistic plan-recognition in cooperative multi-agent systems, which is based on a hierarchical dynamic Bayes Network. This work intends to provide a framework for analyzing interaction among multiple cooperating agents. In their approach, they even allow the specification of hierarchy-levels where individuals are coordinated. Below this specified level though, all plans are executed independently (without interaction). In this regard, interaction is determined by agents finishing their individual plans. The work of Saria et al. is an extension of Bui's [186] multi-agent plan recognition. Even though this approach can handle uncertainty and can be trained, it cannot deal with structured relational data represented in first-order predicate logic.

6.2.4 Multi-Agent Plan Recognition:

While the above-described systems all focus on plan recognition of a single agent, multi-agent plan recognition searches an explanation of observed team-activity traces. Within these, multi-agent plan recognitions aims to identify the team structures and behavior of agents within a team (or changing teams) [187] Multi-agent plan recognition is no new field. In the past 10-15 years, different approaches have introduced techniques to recognize team plans automatically. E.g., Banerjee et al. [188] formalized multi-agent plan recognition in a model, whose key feature is the use of partitioning as a hypothesis pruning mechanism in order to eliminate observations that cannot coexist. A pre-requirement for this method to work is that fully observed team traces and a library of full team plans are available.
Most prior multi-agent teamwork research requires explicit coordination protocols or communication protocols and Generalized Partial Global Planning (GPGP) [188]). Each of these protocols works well as long as all agents know and follow their protocol. Some work in multi-agent teams even requires their agents to work with their teammates in predefined ways such as locker-room agreements [189].

Interaction of Teams with Multi-Agent:

Works in the past introduced models of hierarchical relationships between agents that can recognize team plans and involve multiple agents. E.g. Intille and Bobick [190] rely on coordination constraints among football players to recognize team-tactics. Similarly to work in this chapter, they focus on the interactions between agents. However, the entire domain they are addressing relies on specific team-interaction activities and multi-agent specific plans. In contrast, our work has no specific team-plans available but instead tries to determine the interaction between multi-agents pursuing single-agent plans. Inferring team's states from a team member's routine communications. They provide an efficient probabilistic algo-

[1] Major parts of the Related Work and Text in the related work have been taken from following papers of the author of this thesis. Any text taken from these papers has been written solely by the author of this thesis:
mainly form Gruenerbl A. et al., 2017 [171], and partially from Bahle et al. 2018 [175], please refer to respective entries in the literature list or beginning of this chapter

rithm for plan-recognition designed explicitly for monitoring communications. The plan library used in this work includes information about the average duration of plan steps, which is used to calculate the likelihood of an agent terminating one step and selecting another without being observed to do so. Even though this work would be suitable for our purpose if we would be aiming for detecting communication-interaction.
Zilberbrand et al. [191] introduce the initial steps towards a method for tracking groups and changes in these groups (e.g., merging and splitting) by saving information on the typical plan that each group executes. This work, in contrary to most other work in multi-agent recognition, does not use previously available static social structure or rule-based information. Instead, they use the plan library to identify dynamically changing structures of the groups. For example, a group of passengers in the airport, in one situation like the security check form one group while afterward split into various groups and thus change organizational structure. In this regard, they try to identify agents that behave differently from other agents in the same group and further try to gain a better understanding of the agents by saving the history of the differently behaving agent. The primary method to pursue this goal is to use Dynamic Hierarchical Group Model that indicates the connection between agents. So generally spoken, this work implies an interaction between multi-agent groups by determining strange and group-unlike behavior. Even though our work also strictly speaking implies interaction, but our work addresses an entirely different kind of interaction.

6.2.5 Ad Hoc Team Plan Recognition:

Summarized, all of these works focused on specific social structures/team-game-rules, enabling agents to form teams based on a-priori agreements using specific plans. Therefore, to recognize team plans, the monitoring agent must first know which plans are ideal.
So far lesser explored is the approach to apply plan-recognition for ad-hoc teamwork. It has only arisen in the last years as a necessity in industrial or military settings ([192]). In contrary to common multi-agent teamwork, ad-hoc teamwork cannot take the availability for any protocols for granted. Ad-hoc teamwork is a setting in which teammates must work together reach a common goal without any prior agreement regarding how to work together without knowing each other.

Jones et al. perform an empirical study of dynamically formed teams of heterogeneous robots in a multi-robot treasure hunt domain [193]. They assume that all the robots know they are working as a team and that all the robots can communicate with one another, whereas in our work. This approach cannot be fully applied to the nurse scenario, as though we can assume that nurses can talk to each other, we also have to assume a subtle non-verbal communication that does not follow any communication protocol! Genter [194] presents a role-based approach for ad-hoc teamwork, in which each teammate has a specialized role that shows a specific behavior or performs a particular task, depending on their capabilities but also on the roles been assigned to other team members. This paper highlights the importance of agents being aware of the role they take over.

Bowling and McCracken explore the concept of "pick-up" teams in simulated robot soccer [195]. They introduce coordination techniques designed for a single agent that wants to join a previously unknown team of existing agents. They provide the single agent with a play-book from which it selects the player most similar to the current behaviors of its teammates. The agent then selects a role to perform in the presumed current play.

Ad-Hoc Autonomous Teams, Teamwork without Pre-Coordination

Until this decade, systems were designed to adjust and tune the agents' behaviors to enable them to interact well with one another. Since recently the field progresses toward settings that require on the fly interactions with other unknown agents. In most emergency cases different agents will come together, have no prior information about the other's abilities and qualifications, but still, they have to act and interact quickly. To illustrate it we use the same example used by Stone10 et al. [196] to challenge the community to develop ad-hoc autonomous agents, and which is also relevant to our work. Given a person collapses, be it on the street, anywhere in public, or in a hospital, people currently present are urged to react immediately and perform different tasks, like checking the persons' condition, call for emergency assistance, secure the area (if it happens in the street). In this scenario, the present agents have to interact quickly to figure out how is most capable of performing which task.
In regards of team recognition, agents in ad hoc team settings are not all programmed by the same people (concerning robots) or have not the same background (for humans), and may not all have the same communication protocols or world models or do not speak the same language. Moreover, the capabilities of the other (e.g., in our example "is one of the others a doctor?") may not be fully known to one another. Therefore, the ad-hoc team recognition algorithms described above are not applicable, as a-priori team strategies are not available. In this field only very little work has been done so far. As already mentioned, in [196] by Stone et al. call for developing theory and implement ad-hoc team agents that can interact without pre-coordination with teammates unknown to them and also with team-mate of an older school (robots that a not flexible to adapt to tasks). They specify potential ways to evaluate such approaches and also provide some theoretical analysis of parts of this problem and illustrate an empirical approach using robot-soccer. Stone et al. in [197] also consider agents that can adapt to an environment and the actions of the other agents in ad-hoc teamwork. They studied the optimal strategy to lead a teammate with limited memory and given finite action set and formulated a sequential

decision-making problem in ad hoc settings of 2-agent teams (A and B) in a k-armed bandit formulation.
STABR an algorithm introduced by Sukthankar et al. [198] is a Team Assignment and Behavior Recognition Model. It intends to recovering agent-to-team assignments where the mapping of agents into teams, changes over time. This paper addresses the problem of behavior recognition for teams with dynamic team-composition, yet it is based on matching agent positions to pre-specified geometric formation templates. So despite recognition a kind of interaction within multi-agent teams, this work has no real similarity to ours.

6.2.6 Detecting Interaction in Nurse Emergency Scenarios

To my best knowledge after a thorough literature review, so far no work has been done in detecting the interaction between nurses neither in emergency scenarios nor nurse training scenarios.

6.3 Objectives and Contribution:

The work described in this chapter, can only address the first steps on the way towards building systems that are able to provide context-aware support not just for individuals, but for groups of interacting individuals. Based on a real-world scenario, emergency training of nurse students, the primary goal of this work is to develop mechanisms to detect interactions and collaborations between the multiple agents (nurse students) in a random and ad-hoc team without pre-assigned roles or tasks and without prior knowledge about the actual activities that each agent will carry out. Note, of course there will be knowledge about the domain available and which activities might be or are supposed to be carried out. However, there is no a-priory knowledge what will actually happen in the scenario and which agent will carry out which task.

This chapter presents a method for mapping sequences of low-level actions performed by individuals onto collaboration patterns concerning high-level compound activities. It will do this by first defining the problem and analyzing the requirements (see section §6.4). In the following, a method will be introduced to turn existing "directives of action" into a library of plans, namely into an initial hierarchical domain model (see section §6.5).
For this first simple the domain model an algorithm for detecting collaboration under the use of the domain model will be introduced and evaluated with observations extracted from video recordings of a real training session (see section §6.5.3). Using video traces and thus, assuming the availability of "recognized actions" instead of using actions inferred from actual sensor readings, will allow to explore the impact of different recognition accuracy on the performance of the proposed method. At the same it will allow focusing on the core of this method, the detection of collaboration. The recognition of similar basic actions has been extensively studied, although certainly not entirely solved, and thus is not really a prominent problem of this chapter. Bahle et al. [199] have shown that it is possible to recognize nurse care activities from smart-phone sensor traces. Thus it is feasible in this work, to assume that such recognized activities are available.

The evaluation of the initial domain model reveals some requirements to enhance the model. These enhancements are done to form the 5 layer hierarchical logic goal oriented semantic tree plan model (see section §6.7). Equally, the collaboration detection algorithm is adapted to serve the requirements of the enhanced model and is tested and evaluated within the nurse emergency domain, again by using data derived from video footage.

Since, the method was explicitly developed with the nurse emergency case in mind, scale-ability and generalize-ablity might be an issue for the proposed method to be useful. Therefore, the model was transferred to two other domains (see Section §6.8). With these additional domains the collaboration algorithms were evaluated as well.

6.4 Problem Definition and why Common Techniques do not Apply:

Building such a group "aware system" will require the ability to identify group level activities and collaboration patterns between various agents. Thus, the first step in this chapter deals with the question of how to go from the recognition of individual actions ("classical activity recognition") to identifying collaboration patterns in groups of people. The method proposed in this chapter, in general, follows approaches from plan recognition. Compound activities are represented in a logical tree based plan library that incorporates the lowest level of "basic" atomic actions as leafs. A method for detecting collaboration is introduced as, within the boundaries given by the logical constraints, when "leaf actions of a specific semantic activity have been performed by a number of different agents". If the respective leaf actions were all performed by a single person, then there was no collaboration. If different leaf-actions were performed by different agents, then those agents are considered to have collaborated on this particular compound activity.

Figure §6.1 intends to make the idea behind this approach clearer. With the number of possible atomic actions (A1, .., An), the possible different compound activities known (CA1, .., CAm), and provided it is possible to recognize (with reasonable probability) which agent performs which atomic action at what time, then the question is to determine which of the compound activities have been executed collaboratively and by which (sub) groups of people. In this regard, we are looking for a method to map individual actions, executed by individual users, to specific action nodes as instances of the compound activity models.

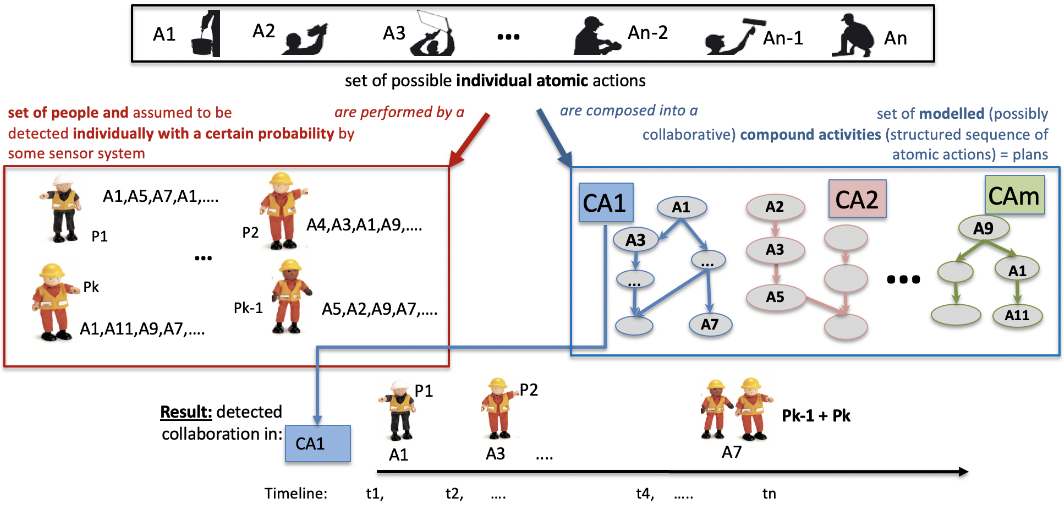

A1
A2
A3
...
An-2
An-1
An
set of possible **individual atomic** actions
set of people and assumed to be detected **individually with a certain probability** by some sensor system
are performed by a
are composed into a
set of **modelled** (possibly collaborative) **compound activities** (structured sequence of atomic actions) = plans
A1,A5,A7,A1,....
P1
P2
A4,A3,A1,A9,....
...
Pk
Pk-1
A5,A2,A9,A7,....
A1,A11,A9,A7,....
CA1
A1
A3
...
...
A7
A2
A3
A5
CA2
...
CAm
A9
A1
A11
Result: detected collaboration in:
CA1
P1
A1
P2
A3
....
Pk-1 + Pk
A7
Timeline:
t1,
t2,
....
t4,
.....
tn

Figure 6.1: Illustration of the problem addressed in this chapter: The starting points are n types of possible relevant individual atomic actions (A_1- A_n black box). Out of these atomic actions m different compound activities CA_1-CA_m are defined (blue box). Then, there are k agents (red box) performing a sequence of atomic actions, some of them likely together, in an ad-hoc manner, to complete some or all of the compound activities, some of them alone.

Previous research on collaboration recognition in ubiquitous computing has generally dealt, for example, with the analysis of a meeting [200] or with the analysis of explicit social interaction [201]. In contrast to these researches, the work in this chapter focus primarily on collaboration in physical activities. Taking the nurse emergency scenario into account, activities of this domain can be divided into different hierarchical structures and can be characterized by the following properties:

1. They are hierarchically compound, meaning that a number of simple actions together form a specific semantic activity, and a number of semantic activities together are necessary to reach a specific goal. For example on the lowest level of "assessing the circulation of a patient" are the primitive actions "pick up blood pressure cuff" and "attach the cuff to arm of patient", which both are part of "measuring blood pressure" which is part of "assessing the vital signs".
2. Despite the hierarchical structure, which by itself imposes boundary conditions, the scenario requires a lot of freedom in the way how the overall activities can be performed. The agents in the scenario may jump between different higher level activities, by their preference or as is required by the circumstances. In the nurse training scenario, for example, a change in the state of a patient (e.g., a sudden cardiac arrest) will cause the group to instantly change their course of action, no matter what each person was about to be doing at that time.
3. A lot of the activities can be performed either individually or collaboratively by a different number of people. For example, one nurse can measure blood pressure on her own, or one nurse can apply the pressure cuff, and another nurse will take over and measure the blood pressure.
4. There is also no predefined role-structure. The collaboration will evolve and changes dynamically and probably differently in each separate training scenario. For example, when the patient goes into cardiac arrest, there is no plan saying that nurse A is responsible for pressing the alarm-button and nurse B has to start chest compression. Such activities will be carried out by the person that, in the specific case feels responsible, is closest (to the alarm button), has their hands free (to perform chest compression). Thus the ability of the group to self-organize is a crucial training goal.

As elaborated in Related Work, to my knowledge, no existing work in related areas such as traditional ubiquitous activity recognition, plan recognition or multi-agent systems addresses this problem for the settings outlined above. Moreover, a main and not yet solved issue in Artificial Intelligence (AI) is how to lead a machine to learn and to understand what is going on in the real world? How can we teach a computer semantics and thus, bridge the semantic gap ([202]).
The generally accepted approach is to build a model of the real world and feed it to the computer. In combining observed probabilistic low-level activities (e.g., derived from sensors and common activity recognition methods) with the model, the machine should be able to derive high-level meaning. So far, AI has developed a variety of algorithms and concepts in both, deriving models of

the real world called "planning" and automatically recognizing activities on different levels, e.g., "plan or goal recognition." Still, all these approaches have their limits, specifically when it comes to recognizing and understanding quite complex and possibly random scenarios like the nurse training.

6.4.1 The Nurse Emergency Scenario

In the nurse training scenario, we deal with a, in real-life, well defined and established scenario. It is, in general, not necessary to define or derive a plan, like in "hostile plan recognition" or in "assessing team sports," because the plans or plan like constructs (algorithms, quality management, directives for nursing) already exist. For example, the process of how a nurse, step by step should figure out what is wrong with a patient or how to react to a dysfunction in the patient's body, already has been established in health care.

Furthermore, these directives, like "the A2E" algorithm, are well described in the literature ([203], [161]) and taught to nurse-students all around the globe. Therefore, these directives are well known (or in regards of nurse students should be well known) to its actors.

The A2E algorithm (A2E or also known as ABDCE Algorithm stands for - Airways, Breathing, Circulation, Disability, and Exposure) applies to every human emergency. It includes general instructions on which bodily functions to assess during the examination in which order and what actions to take (treatments to apply) for each particular dysfunction, specifically in the case of emergency. Mainly, it is defined for single actors (a single nurse has to be able to examine a patient, and in the emergency, a single nurse has to be able to "keep a patient alive" until help arrives). Nevertheless, in reality (concretely in the nurse-training scenario), the examination is commonly performed by a set of 3-5 nurses. These are not only very likely randomly put together, but also have no pre-defined roles or assigned tasks.

6.4.2 Data Collection

The nurse emergency scenarios, we were able to video tape, are part of the nurse's higher education at the nursing department of the University of Southampton in the UK. A group of 3-5 nurses are presented with a highly sophisticated patient dummy that simulates a patient that, for example has "just been admitted" to the emergency ward, or "just came back from surgery". The dummy, called "SimMan," has a heartbeat, can breathe, and talk/react (remotely controlled), vital signs can be measured, injection administrated, and CPR and various other procedures can be applied. Besides remotely controlled talking and reacting, all "bodily functions" of SimMan can be modified at any time. Meaning, SimMan can go into cardiac arrest, breathing can stop, and essentially SimMan can "die".

The task assigned to the nurses (nurse students) is to diagnose the problem (e.g., why SimMan "feels pain in his chest"), and initiate a solution (namely by following the A2E algorithm). In most cases, the "patient's state" will change (deteriorate) during the procedure and the group has to react to these changes. A vital aspect of the simulated training is that the agents (nurse-students) require the ability to work together and interact, but as an ad-hoc group they get not specific instructions, there are no predefined roles and, most often, the team has never worked together before.

The scenario is interesting in itself, as it contains a mixture of strict structure (first assess Airways, then Breathing, then Circulation, etc.) paired with very dynamic improvisation. Some individual procedures have to be followed by the book. For example, CPR (cardiopulmonary resuscitation) has to be administered precisely, otherwise it would not be effective. Alternatively, if a patient is unconscious, it is essential to make sure the airways are free, the patient is breathing, and the heart is beating, before the state of the skin can be examined. On the other hand, the nurses are free (have to be free) to collaborate in any way they want. Thus CPR may be performed alone (alternating between the chest compressions and breathing steps, or even skipping breaths for a certain amount of time), but with two or more people around, the nurses will alternate in doing specific actions (taking turns in chest compressions, since applying them is physically exhausting). Furthermore, specifically in emergency situations, protocols might be thrown overboard and the task at hand will be done regardless of what regulations say.

In total four training sessions, as described above, were recorded on videotape. Each of these sessions included three randomly chosen nurse trainees in their second year. None of the trainees had information about details of the scenario (except that they had to care for a "patient" and make sure the "patient" gets better) also, no roles were assigned beforehand. Each training session lasted between 20 and 25 minutes. On average each dataset included approximately 800 unique data-points. Figure §6.2 shows a scene of one of these sessions.

6.4.3 Challenges of the Nurse Scenario

It is a common understanding that the A2E algorithm should be followed, but despite thorough training, there is no guarantee that the activities performed by the nurses, will strictly follow this protocol. Even though it is designed to help the nurse (agent) to reach a goal, namely, figure out what is wrong with a patient and treat all eventual illnesses and handle emergencies, the agents might use experience or intuition to reach the goal in different ways and thus not following the algorithm to its letter. For example, the nurses might take shortcuts like, when the patient speaks clearly and normal, and is conscientious, there is no need to assess the Airways or Disability. Moreover, agents might mix plans, perform plans in parallel at the same time, interrupt plans, skip steps or repeat parts of the plan.

Is this scenario carried out during a nurse-training session, it might even go further. The "plan" can be "frozen" at any time for a trainer to explain something. Alterna-

Figure 6.2: Nurse Trainees collaborating in attending a patient-dummy in an emergency

tively, nurse-students might discuss what to do in the middle of a plan or might ask the teachers around for help (which is not part of the general plan).

Regarding plan recognition, this scenario reveals the limitations of commonly established algorithms, as neither plan recognition techniques for single agents nor universal plan recognition for multi-agents can be fully applied. Even though "the plan" is designed for a single agent, it has to be assumed that it is carried out by multiple agents. Therefore, pruning-mechanisms for eliminating false observations, (e.g., in regards to the time) do not apply. For example, in a multi-agent scenario, the single-agent-time-assumption does not fit that "a nurse cannot attach equipment before she has fetched it in the storage room," as it is possible that another nurse has brought it and placed it next to our nurse. Most multi-agent approaches will not work either, as we have no team plans available (as required for most multi-agent approaches, e.g., in [190]) nor are agents assigned explicitly to only one team at a time (as required in [204]) moreover, the plan library does not include any team rules (as in [188]). Furthermore, there is no guarantee that any plan or any part of a plan will be performed only once or at least only once).

In recent years works about "ad-hoc," multi-agent plan recognition have been published. Unfortunately, these works, generally require the knowledge of roles of team members within the team, or respectively that each team member knows its role within the team even if the team members do not know each other. Specifically, Genter et al. [194] highlights the importance of role knowledge for ad-hoc multi-agent plan recognition to work.

To sum up: a nurse-(training)-examination-emergency-session is a particular scenario which, in regards to plan recognition, includes some constraints and limitations on the one hand, but has to assume a lot of freedom and free actions on the other:

- single agent plans carried out by multiple agents
- no knowledge about the team, no specific team rules, no pre-defined roles within the team, agents can be part of various teams at the same time
- plans might not be followed strictly: skipped steps are possible, actions can be performed parallelly, special events might not be captured in the plan

6.5 From a "Directive of Actions" to the Plan Library

The background of health-care comes with the big advantage that many processes are well defined, as the above mentioned A2E algorithm shows. It was designed for nurses or paramedics to be followed step by step in order to understand what is wrong with a patient and react to it as efficiently and effectively as humanly possible. These processes are well known to all professional players in the field of a hospital. Despite the availability of these processes, specifically in examination/emergency treatment, modeling of these as workflows has been limited to particular activities or very coarse models. As an example, basic flow-models of performing resuscitation exist and are used to teach the underlying process, but the entire flow of how to deal with a not-well patient has never been modeled in detail in a professional modeling notation. In the practice of nurse-teaching such "flows" often are provided descriptively (e.g., see [203]). The goal of this section is to provide a way to capture the entire A2E Algorithms in detail.

6.5.1 Modeling the Process Flow

So far, to my best knowledge the full hospital examination/emergency flow has never been modeled. Therefore, the first step in developing an appropriate model for the nurse emergency training was to put the scenario as it should be performed down into a work-flow model. Even-though this scenario exist between well-defined borders, it is still quite complicated. In order to keep the model perceivable, it was broken down into different hierarchical steps of detail. The flow models are modeled in a simplified BPMN (Figures §6.4 and §6.3).

The flow model in Figure §6.4 already shows that every branch of the A2E, namely Airways, Breathing, Circulation, Disability, Exposure includes two main blocks, "check" and "react" (which is either treatment, emergency treatment or nothing). All five branches including both blocks were modeled in further detail, see the flow-model for Circulation in Figure §6.3) as an example.

6.5.2 Flow-Model of the A2E Algorithm.

The BPMN provides the main advantage that both, the location activities take place in and the dependencies between activities can be modeled. Location in this model is expressed in the form of swimming lanes. Dependencies can be followed through logic gateways.

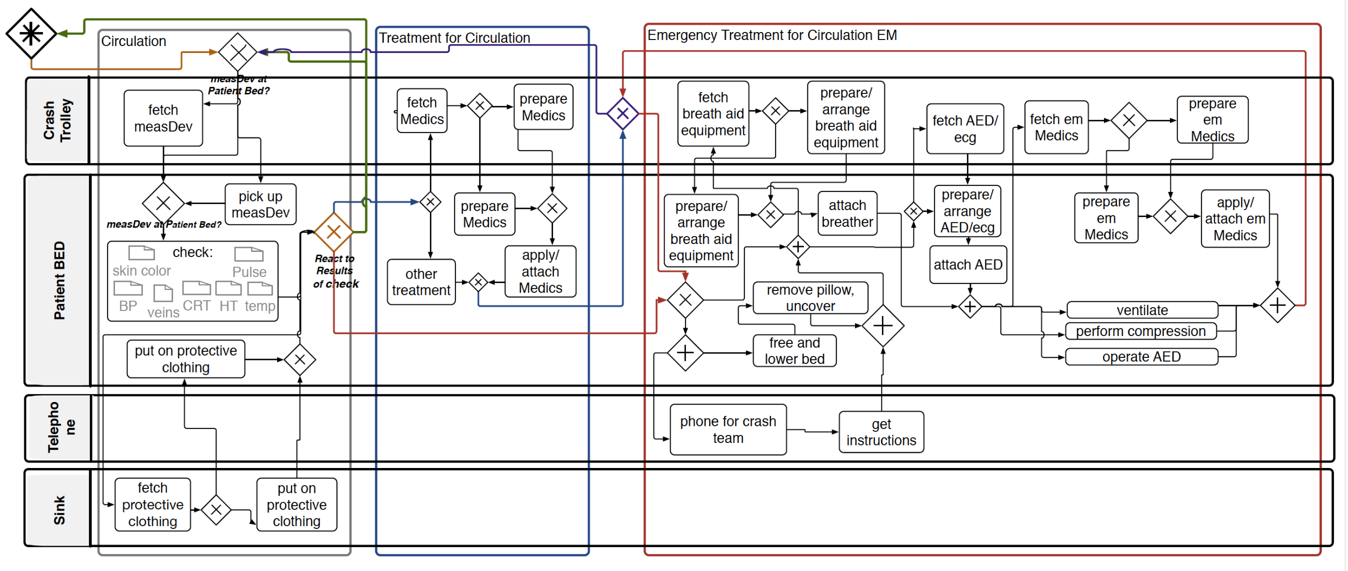


Figure 6.3: The section "Circulation" modeled in detail.

Graph based Recognition of Interaction?

In order to determine interaction the first idea that comes to mind seeing this flow model is to use simple graph techniques. For example, follow each nurse's path in the graph using real-world traces. Then determine, how many nurses are necessary for the graph to be covered thoroughly.

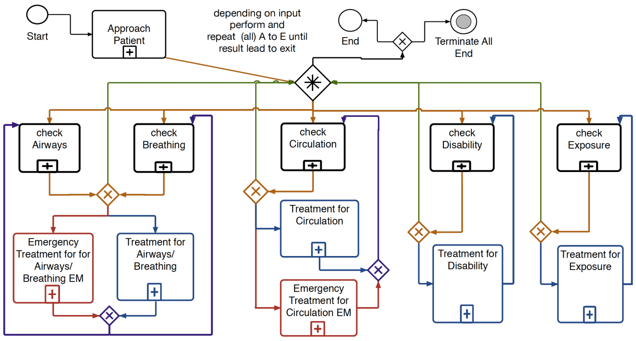

Start
Approach Patient
depending on input perform and repeat (all) A to E until result lead to exit
End
Terminate All End
check Airways
check Breathing
check Circulation
check Disability
check Exposure
Emergency Treatment for for Airways/ Breathing EM
Treatment for Airways/ Breathing
Treatment for Circulation
Emergency Treatment for Circulation EM
Treatment for Disability
Treatment for Exposure

Figure 6.4: A high level view on the A2E algorithm.

Very High-level interaction: If the path of one nurse does not cover the entire process, it is very likely that traces of other nurses will be found in this graph. Therefore, we can define: Find the traces of all nurses necessary to cover the graph. It is feasible then to assume that these nurses interacted in dealing with the patient.

High-level interaction: Find all nurses necessary to cover parts of the graph. Then define: for all nurse traces necessary to cover a part of the graph we can assume that these nurses interacted in a specific high-level part.

Low-level interaction: Find all nurse-traces in a specific task or sequence of tasks, and then we can define: for all nurse traces in this specific task we can assume that these nurses interacted on this low level.

The significant disadvantage of this model notation can be found in the lack of order. The availability of dependencies in a graph, self-imply the lack of clear order. To be more precise, in this graphs at any given point in time and without any other information, it is impossible to tell which nodes had been visited before or which path was taken to reach the present node. Therefore, this kind of model (graph) requires explicitly storing of each node. In decision trees on the contrary, by its nature, at any given point in time, at any possible node in the tree, the path to reach this node is unique. In this regard, the challenge to improve the model was to turn it into a decision tree based model on the one hand but keeping the benefits of including dependencies.

6.5.3 The 3-Layer-Model (3l-HGOST)

Based on the available flow graphs a first simple tree model was derived. See Figure §6.6 as an example. This model consists of three basic hierarchical layers:

World Context Layer (WCL): The first layer incorporates, implements, and structures general high-level human knowledge or high-level regulations about a domain. More specifically, for the nurse emergency scenario, it is based on the A2E algorithm and consists of the WCL-nodes "Airways", "Breathing", "Circulation", "Disability", "Exposure." See Figure §6.5.

Semantic Activity Layer (SAL): The SAL is a child-level to the WCL. It details the higher-level WCL nodes into concrete semantic activities (SAs). The SAL describes each specific SA to be done within each WCL node. Every SA is a child to one WCL parent. (See green nodes in Figure §6.6)

Atomic Action Layer (AAL): Each semantic activity (SA) is comprised of a sequence of ordered basic atomic actions (AA). Therefore, a sequence of AAs details the higher-level SA. AAs are leaves of the tree and do not have any children. Furthermore, SAs can either be start actions, stop actions, or sequence actions. Regarding the Nurse-Emergency-Scenario, most SAs like "measuring blood pressure" are comprised of "fetch" (if the measurement device is not available at the bedside), "pick up

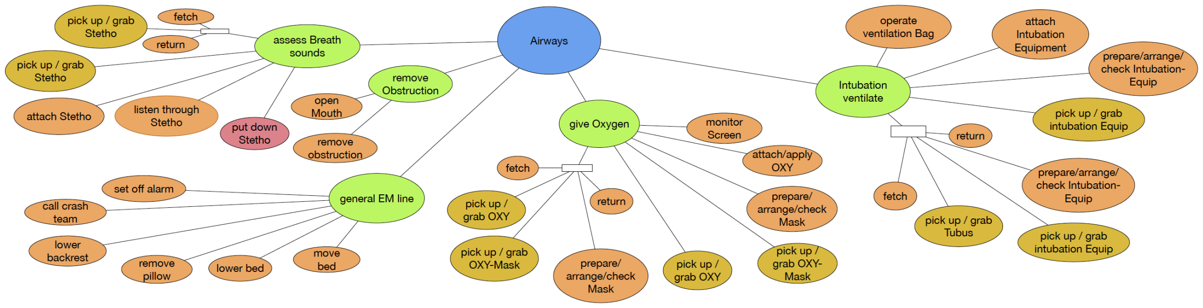


Figure 6.6: A part of the Model specifying the WCL Node "Airways" (blue) and its SAL (green) and AAL children (orange)

a device," maybe "arrange/prepare/check a device," "attach a device," "measure," and maybe "monitor screen."

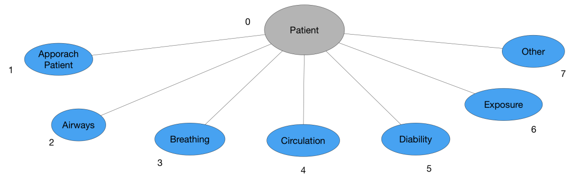


Figure 6.5: 7 Nodes of the the A-to-E Algorithm in the World Context Level.

A2E Model Details

The 3l-GHOST model of the A-to-E Algorithm incorporates 7 unique WCL Nodes (Approach, Airways, Breathing, Circulation, Disability, Exposure, Other). Each of these has 1 - 29 children (SAL Nodes) out of a pool of 46 unique compound SAs. Note that any given SA can be child to different WCL nodes. E.g. 'give oxygen' can be part of either "Airways" or "Breathing", 'Intubation' can be an emergency treatment of either "Airways" or "Breathing" or "Circulation" or "Disability". Each SAL node has 1 - 10 children (AA leaves) which are out of a pool of 73 unique AA (atomic action). Again, note that any AA can be child to several SAL nodes. E.g. 'fetch' is a possible AA for 36 different SAs,

WCL (nodes)	# SAL (nodes) per WCL node	# AAL (leaves) per SAL node
Approach/Other	1 / 7	2
Airways/Breathing	15 / 14	2 – 10
Circulation	29	1 – 10
Disability/Exposure	21 / 16	1 – 10
unique total	46	73

Table 6.1: Number of Nodes in each Level

6.6 Detecting Collaboration with the 3lGHOST Model

In general, in this Chapter, collaboration is considered if an instance of a compound activity exists, in which two or more people have executed at least one basic atomic action. Interaction or collaboration can happen at each of the higher levels (WCL and SAL). For example, two nurses can both perform a task which in some way belongs to Circulation (e.g., one counts pulse, the other attaches the IV) - in this case, both interact on the WCL of "Circulation." However, going one degree of granularity down in the tree to the SAL, these two nurses do not collaborate, because one nurse works on the SAL "check patient" while the other nurse works on the SAL "treat patient." In this sense, interaction/collaboration can happen on each of the layers in the tree model, and thus collaboration has to be defined.

6.6.1 Definition of Interaction

People interact when they are working together to fulfill a task. Regarding the nurse scenario, interactions can happen on different levels (e.g., "assess Circulation" highest level, "check blood pressure" meta-level, "inflate cuff" atomic level. Interaction can happen in different ways:

- **parallel:** a sequence of atomic actions of the same SA (e.g., measure Pulse -> "grab wrist/touch neck/touch leg," "count pulse" (atomic level)) can be performed by several nurses at the same time.
- **sequential:** several nurses can perform the atomic action of a SA in alternating manner (e.g., one performs compressions, another one operates the venti. bag)
- **together:** several nurses can perform different atomic actions of the same SA at the same time (e.g., one nurse can attach the ventilation hoses at the oxygen plug while the other nurse attaches the mask)
- **with time difference:** one nurse can start a SA, but perform only some of the corresponding atomic actions and another nurse could finish the missing actions with a time difference (e.g., one nurse could go to the storage to pick up a measurement device and bring it to the patient's bed. However, instead of using it the nurse could be called off to help with another activity. Later another nurse picks up where the first nurse left and finishes.)

Figure §6.7 is an illustration of some key problems involved in mapping sequences of individual atomic actions onto team executions of compound activities: Both figures (left and right) start with the same sequence of recognized individual atomic actions P_1 to P_k and the same plans (models of compound activities) CA_1-CA_m.

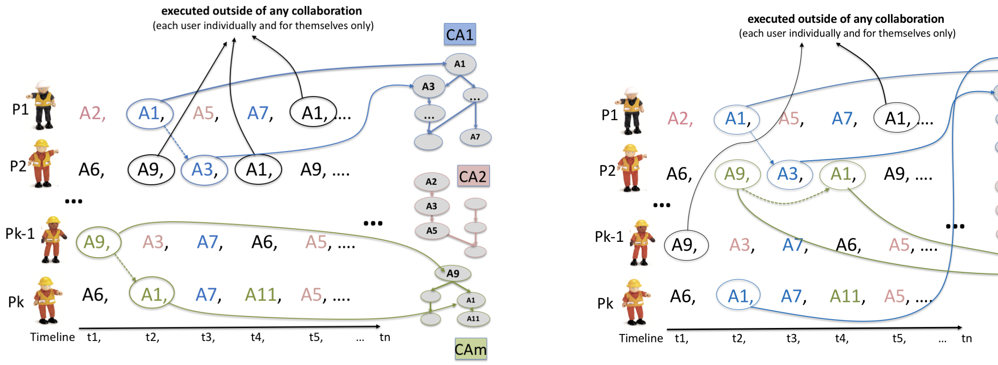

executed outside of any collaboration
(each user individually and for themselves only)
CA1
CA2
CAm
P1
P2
Pk-1
Pk
A2, A1, A5, A7, A1, ...
A6, A9, A3, A1, A9, ...
A9, A3, A7, A6, A5, ...
A6, A1, A7, A11, A5, ...
Timeline t1, t2, t3, t4, t5, ... tn
A1
A3
A7
A2
A5
A9
A11
executed outside of any collaboration
(each user individually and for themselves only)
CA1
CA2
CAm
P1
P2
Pk-1
Pk
A2, A1, A5, A7, A1, ...
A6, A9, A3, A1, A9, ...
A9, A3, A7, A6, A5, ...
A6, A1, A7, A11, A5, ...
Timeline t1, t2, t3, t4, t5, ... tn

Figure 6.7: An illustration of some key problems of mapping sequences of individual atomic actions onto team executions of compound activities.

In this example, we center our attention on the individual atomic actions A_1 and A_9. A_1 represents a standard action like screw driving in an assembly scenario and is performed four times by three of the different agents. Moreover, it is part of two of the compound activities CA_1 and CA_m. The atomic action A_9 is performed only by two users and is part of CA_m only. Both the pictures in Figure §6.7 (left and right) are examples of valid assignments of actions A_1 and A_9 (of which more exist). In the left picture the sequential constraint between A_1 and A_9 imply that A_9 performed by P_{k-1} is the only option for CA_m. In the right picture though, the possibility of two agents contributing to the same individual action within a CA is displayed (e.g., when one starts putting in a screw then leaves and the second one finishes the job).

6.6.2 High-Level Collaboration

The first step in determining collaboration is to traverse "fill" the tree with atomic activity data of various agents participating in the scenario. Once the tree is deployed with data of several agents, collaboration can be detected.

Traverse the Tree

At each point i in time T, for the current newly recognized atomic action AA_i:

1. **assign probabilities:** find all possible locations PL of AA_i in the tree (leaves = atomic action) and calculate the probability for each of the PL_z:

$$pAA_{i|PL_z} = 100/\#PL \qquad (6.1)$$

2. **backtrack:** evaluate the probability of the siblings SB of AA_i ($SB_x|AA_i$) in each PL (=probability of siblings if available, gained in the iterations $T_{i-10} - T_{i-1}$) weight $pAA_i|PL_z$ according to the probability of its siblings $pSB_{1..x}|pAA_i|PL_z$.

3. **backtrack siblings:** adapt probability of sibling accordingly

4. **upwards probability:** calculate probability of higher level SA based on probability of its atomic actions

$$pSA_i = (pAA_1|_{SA_i} + pAA_2|_{SA_i} + ... + pAA_x|_{SA_i})/x \quad (6.2)$$

5. backtrack on higher levels

Each of the steps described above is conducted for each data point (a data point = primitive atomic action recorded in 5 seconds or fewer intervals)

Detect Collaboration

In order to detect collaboration, the tree has to be traversed for each participating agent individually. Afterward, the tree is traversed again with joint data of two or more nurses. Collaboration is detected when both or all nurses contribute to one SA (semantic activity), which means that one nurse does some atomic actions (comprising SA) and the rest is done by the other nurse(s). Collaboration is calculated as follows:

1. **traverse** tree for each nurse **individually**

2. **traverse** tree with **joint data** of two or more nurses (N_{1+2},N_{1+3},N_{2+3}, N_{1+2+3})

3. **detect collaboration as**: for each semantic activity SA at time i (SA_i): if the joint probability of both (or all three) nurses is greater than the probability of the individual nurse.

6.6.3 Evaluation

Drawing on more than ten years of activity recognition research and its progress in reliably detecting various activities with sensors, we can safely assume, and actually have shown before [199] that sufficiently correct recognition of nurse activities is possible, even with unobtrusively worn smart-phones.

Thus, for the evaluation of our model and the algorithm for determining collaboration, we use label-data, derived from video footage of real training-sessions (see Figure §6.2). This data set includes a total of 975 atomic actions performed by 3 agents and includes a total of 530 atomic level collaborations. In terms of collaboration, each collaboration was calculated for all combinations of agents N_{1+2},N_{1+3},N_{2+3}, N_{1+2+3} and the results were validated with the ground-truth (actual collaboration extracted from the video footage). Table §6.2 shows the precision and recall of this evaluation for the different combinations of collaboration and for each activity. Note

	Collaboration N1,2,3	Collaboration N1,2	Collaboration N1,3	Collaboration N2,3	Mean
	prec / rec	prec / rec	prec / rec	prec / rec	prec / rec
Approach	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 0.91	1.00 / 0.98
Breathing	0.82 / 1.00	0.72 / 1.00	0.82 / 1.00	0.95 / 0.96	0.83 / 0.99
Circulation	0.98 / 0.92	0.99 / 0.95	0.94 / 0.94	1.00 / 0.92	0.98 / 0.93
Disability		1.00 / 0.90			1.00 / 0.90
Exposure	1.00 / 1.00	0.95 / 1.00	1.00 / 1.00	1.00 / 1.00	0.99 / 1.00
Other	0.98 / 0.99	0.99 / 0.99	0.99 / 1.00	1.00 / 0.99	0.99 / 0.99
Mean	0.96 / 0.98	0.94 / 0.97	0.95 / 0.99	0.99 / 0.96	0.96 / 0.96

Table 6.2: Precision and Recall of Detecting Collaboration (Nurse 1 + 2 + 3, Nurse 1 + 2, Nurse 1 + 3, Nurse 2 + 3)

that empty spaces for "Disability" mean that no collaboration happened for this activity except between nurse 1 and nurse 2. The evaluation result of the model is relatively high, with an average precision and recall of above 90% for most activities and combinations.

Evaluation with Recognition Errors

Using label data implies a 100% correct recognition rate of atomic activities, which is not realistic even for highly accurate system. Thus, this must be assumed to be one of the reasons for the very high accuracy in the evaluation. Therefore, to provide a more realistic evaluation, uncertainty was added to the label data. This was done by randomly picking an atomic activity (out of the 73 possible AA) instead of the correct label, XX% of the time. In order to validate the stability of the model, for detecting collaboration, the algorithm was evaluated with 5%, 10%, 15%, 20%, 35%, 50%, 60% and 75% added recognition error.

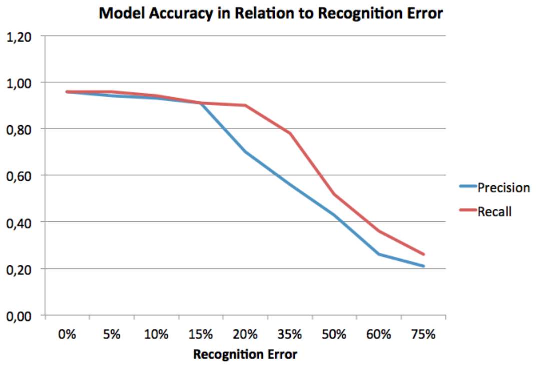


Figure 6.8: Precision and recall of detecting collaboration in relation to recognition error.

Figure §6.8 provides an overview of the accuracy of this evaluation. It clearly shows the stability of the model even with added recognition error of up to 20% (with a precision of 70% and recall of 90%). With more than 35 % of recognition error the model, of course, starts to worsen (recall still at 78%) and breaks down with 50% and more recognition error, as was expected. These results show that the model can handle a fair amount of recognition errors and still performs well.

6.6.4 Discussion

To summarize, the 3-layer model is a hierarchical way to structure information describing a domain. It groups single basic or atomic actions into semantic compound activities. The semantic activities in the respective layer are further grouped into a very high-level meaningful (world) context.

The results of detecting collaboration on the high level of WCL nodes with this 3-layer model works very well, as has been demonstrated above. The 3-layer model though is rather coarse and does not include much structure and only limited dependencies. In other words, SAs like "measure pulse" and "perform CPR" have a time-orderly dependency to each other. First, by checking the pulse, it has to be determined that the heart has stopped and CPR is necessary! Therefore, "measure pulse" has to come before "perform CPR". This clearly calls for the introduction of time-wise dependencies into the model.

On the other hand, "measure pulse" is an examination while "perform CPR" is a reaction or treatment to the result of the examination. The general goal of the 3-layer model was to provide a way to determine collaboration on a very high level - on the level of the world context.

Real life applications, however, will likely require detection of collaboration on a lower level (e.g., whether two nurses perform CPR together, or whether specific heavy pieces are being carried by at least two persons, etc). In this regard, the natural availability of more detailed grouping - like in the measure pulse/CPR example above - SAs can, for example, be grouped into "examination SAs" and "reaction/treatment SAs". These groups also indicate a time-wise dependency, as (to repeat the above) the examination task(s) will always come before the reaction/treatment tasks.

Errorrate	0%		5%		10%		15%		20%		35%		50%		75%		Inst
	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	
Approach	1.00	0.98	1.00	0.96	1.00	0.94	1.00	0.92	0.70	0.90	0.65	0.79	0.51	0.39	0.10	0.10	6
Breathing	0.83	0.99	0.79	0.98	0.75	0.98	0.71	0.96	0.55	0.95	0.36	0.85	0.23	0.58	0.03	0.26	5
Circul.	0.98	0.93	0.96	0.92	0.93	0.90	0.90	0.86	0.71	0.86	0.59	0.75	0.58	0.62	0.35	0.51	83
Disability	1.00	0.90	0.93	0.90	0.89	0.89	0.87	0.91	0.80	0.86	0.50	0.77	0.09	0.29	0.02	0.09	3
Exposure	0.99	1.00	0.99	0.98	0.98	0.95	0.98	0.93	0.76	0.91	0.64	0.77	0.42	0.50	0.08	0.18	5
Other	0.99	0.99	0.98	0.97	0.97	0.94	0.96	0.90	0.80	0.87	0.71	0.73	0.60	0.55	0.38	0.28	60
Mean	0.96	0.96	0.94	0.96	0.93	0.94	0.91	0.91	0.70	0.90	0.56	0.78	0.43	0.52	0.21	0.26	

Table 6.3: Accuracy (Precision and Recall) of Detecting Collaboration with increasing added recognition error 0 - 75%, and average number of instances available for each activity.

6.7 The Extended 5-layer HLGSTP Model

To accommodate these main considerations/requirements from the discussion above, two more layers were introduced into the model. First of all, to provide a method for including the required dependencies between the layers and also between different nodes, a logic meta layer was defined (see details below).
Furthermore, with the introduction of the Semantic Order Layer (SOL), a layer between the WCL and SAL was created which groups SAL-Nodes within a WCL area into semantically ordered groups. In this way, the original 3-layer model (3l-GHOST) was enhanced and extended to become the 5-Layer-Hierarchical Logic Goal Oriented Semantic-Tree-Plan (5l-HLGSTP):

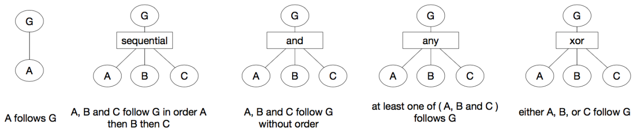

G
A
A follows G
G
sequential
A
B
C
A, B and C follow G in order A then B then C
G
and
A
B
C
A, B and C follow G without order
G
any
A
B
C
at least one of (A, B and C) follows G
G
xor
A
B
C
either A, B, or C follow G

Figure 6.9: The LCL logic connection layer provides different logic connections between leaves.

Logic Connection Layer (LCL): The LCL is a vertical meta layer (spanning all four other layers) connecting both, all of the above layers with each other and also every single node with logic operators. These include the logic AND, the logic XOR, ANY (meaning, at least one or several with no specific order) and Sequential (meaning, AND but with specific order). Furthermore, this level includes information about activities that might not happen "maybe" or might be repeated "maybe repeat". See Figure §6.9

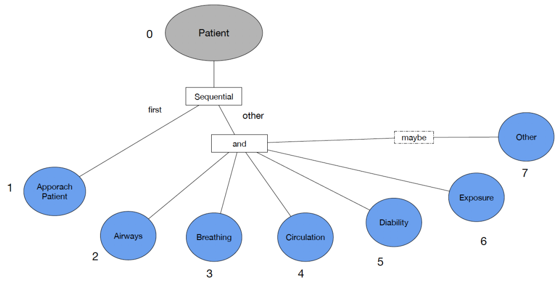


Figure 6.10: The WCL world context layer, structures high level human knowledge/regulations about a domain

World Context Layer (WCL): The highest level layer is basically the same as in the 3l-HGOST and represents and structures general high-level human knowledge or high-level regulations about a domain. Mainly, the WCL of the 5l-HLGSTP differentiates from the 3l-HGOST as the WCL-Nodes are structured by logic meta nodes (from the LCL). Compare Figure §6.5 of the 3l-HGOST and Figure §6.10. For the nurse emergency scenario, this layer also includes the same A2E classes as nodes.

In the 5l-HLGSTP model, the WCL are logically structured to explain that initially, the nurse approaches the patient and checks the patient's general medical state. Afterward, all the other classes "Airways", "Breathing", "Circulation", "Disability" and finally "Exposure" have to be visited, ideally but not necessarily in the order A to E. This is the general order of the assessments, yet depending on the first one (approaching the patient and getting a brief picture of the patient's state) the order of the other assessments may change. E.g., if the patient is conscious and speaks normally and does not seem to be short of breath, the priority to check the airways is reduced.

Semantic Order Layer (SOL): The SOL is a child layer of the WCL and provides a semantic order to its children, the semantic activities (SA). In itself, the SOL is still rather high-level but divides the WCL into more structure and detail. The SOL in the nurse emergency scenario, for example, divides all the possible nodes (assessments) of the WCL into either "check" or "react". Whereas "react" could either be "do nothing or finish doing the current action", "a treatment", "an emergency treatment" or "repeat the check action". See Figure §6.11.

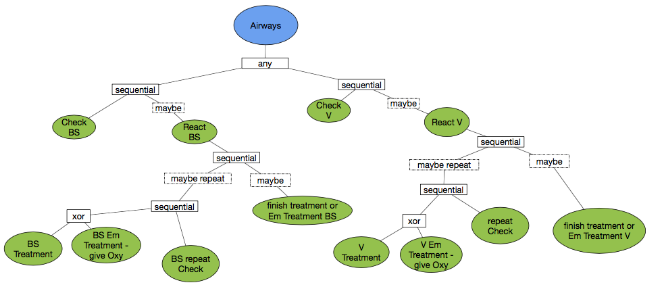


Figure 6.11: The SOL semantic order layer provides a coarse semantic and time dependence structure (dark green) for a WCL node (blue)

Semantic Activity Layer (SAL): The SAL is a child layer of the SOL. It details the higher-level SOL node into concrete Semantic Activities (SA). Therefore, e.g. in the nursing scenario, all "assessment-SAs" are children of "SOL-check", explicit treatment-SAs are children of "SOL-Treatment", explicit emergency-treatment-SAs are children of the "SOL-Emergency-Treatment", see Figure §6.12.

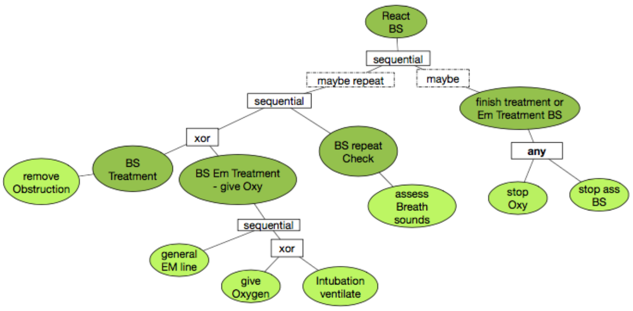


Figure 6.12: The SAL semantic (compound) activity layer is a collection of semantic compound activities (light green) belonging to its SOL parent (dark green)

Atomic Action Layer (AAL): Each Semantic Activity (SA) is comprised of a sequence of ordered single atomic actions (SAA). Therefore, a sequence of AAs comprises the higher-level SA. AAs are leaves of the tree and do not have any children. Furthermore, AAs can either be start actions, stop actions, or sequence actions. In terms of the Nurse-Emergency-Scenario, most SAs, like "measuring blood pressure" are comprised of "fetch" (if the measurement device is not available at the bedside) or "pick up device", maybe "arrange/prepare/check device", "attach a device", "measure" and maybe "monitor screen". See Figure §6.13.

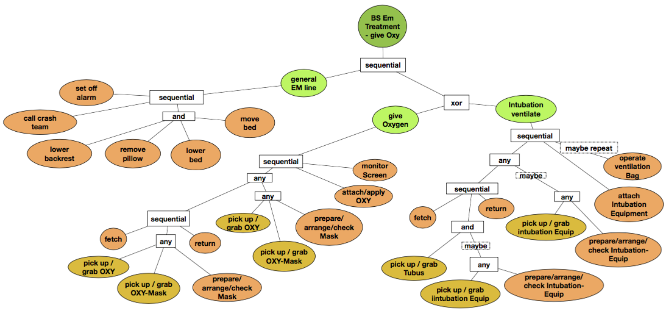


Figure 6.13: The AAL atomic action layer specifies the single atomic actions (orange), which together form a semantic compound activity (light green)

Regarding plan recognition, this model could be summarized as "logical tree-based collection of semantic plans." It includes four semantic layers hierarchically incorporating plans, organized in various structured semantic levels (layer 1-3). Atomic action sequences comprise each of the plans (SA nodes) from layer 4. Additionally, logic operators are connecting each node of every plan within the plan, and each plan to its semantic structure (layer 5). To get a better understanding how the five layers are connected, Figure §6.14 explanatory shows an excerpt of the nurse emergency care 5lHLGSTP model with some selected nodes and leaves. The complete 5l-HLGSTP model of the emergency training domain, taken down into various parts, can be found in the Appendix §7.3.

6.7.1 The Emergency Training 5l-HLGSTP

The 5l-HLGSTP model of the A2E algorithm, like the 3l-HGOST consists of seven unique WCL Nodes (Approach, Airways, Breathing, Circulation, Disability, Exposure, Other). Each of these WCL nodes has 2-8 children (SOL nodes) structuring the lower levels. These SOL nodes are generally either of the kind "check" or "react". Every SOL node has 1 - 29 children (SAL nodes) out of a pool of 46 unique compound SA (semantic activities).

Note that any given SA can be a child of different WCL-SOL nodes. E.g. "give oxygen" can be part of a SOL-treatment of either "Airways" or "Breathing", "Intubation" can be a SOL emergency treatment of either "Airways" or "Breathing" or "Circulation" or "Disability". Each SAL Node has 1 - 10 children (AA leaves) which are out of a pool of 73 unique AA (basic atomic action). Again, note that any AA can be a child to several different SAL nodes. E.g., "fetch" is a possible AA for 36 different SAs. A summary of nodes in the different levels is summarized in Table §6.5.

6.7.2 Detecting Low-Level Collaboration

To evaluate the 5l-HLGSTP emergency model, a custom framework for parsing the tree(s) and feeding atomic activity information into the tree for detecting collaboration was developed and implemented by Gernot Bahle and published in [175]. In course of this thesis, less interest lies on the theoretical details of the detection algorithm, but on the practical prove of concept. Thus, in the following, the collaboration detection algorithm will be summarized only briefly, but the interested reader can get more details from Bahle et al. [175].

The collaboration algorithms has similarities with the collaboration detection algorithm of the 3-layer model above, but is more sophisticated. Besides, dealing with more depth in the model, it also includes handling the logic connections between knots and time constraints. Nevertheless, the basic detection of collaboration, like in the 3-layer model, is defined as: if different actors fulfill various atomic actions of a particular semantic compound activity, or different compound activities belonging to the same semantic order knot, these actors collaborate on the respective layer.

Collaboration Detection Results: Table §6.4 summarizes the results of recognizing collaborations between the nurses of 2 person collaboration and 3 person collaboration. Note that recognizing collaboration does not merely mean, collaboration is detected in general, but also collaboration is correctly detected in the correct class (within a total of 7 WCL level classes). It averages the different collaboration combinations of two and three per-

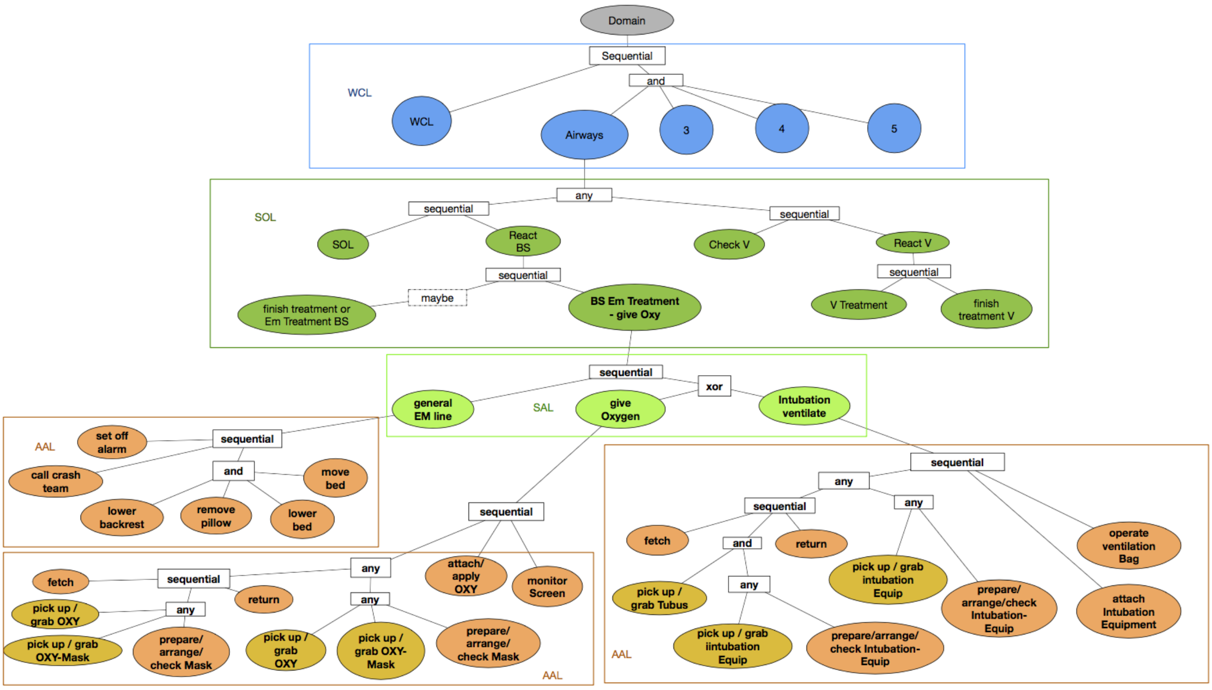


Figure 6.14: The 5-Layer Hierarchical Logic Goal Oriented Semantic-Tree-Plan Model with exemplary nodes taken from the nurse model.

sons (e.g. collaboration of nurse 1 and nurse 2 = N12; or nurse 1, nurse 2 and nurse 3 = N123) overall data-sets in relation to the error of recognizing the atomic activities.

Specifically, the 3-person collaboration shows a high precision of almost 90% with correct recognized atomic activities and also shows a quite robust behavior even with increasing recognition error. At the same time, the recall is around 60%. Meaning that the detection of collaboration misses some collaborations but the ones that are detected are detected correctly. Precision of 2-person collaboration is less accurate than precision of 3-person collaboration. This effect can be explained by the fact that 3-person collaboration has twice as many instances than 2-person collaboration.

	2 Persons		3 Persons	
Errorrate	Recal	Precision	Recal	Precision
0.0%	70.0%	65.5%	58.5%	87.7%
5.0%	64.5%	60.9%	52.5%	85.5%
10.0%	59.8%	56.8%	47.2%	83.2%
15.0%	55.0%	52.7%	41.7%	80.4%
20.0%	50.8%	49.0%	37.6%	77.9%
25.0%	47.1%	45.8%	33.5%	74.9%
30.0%	43.4%	42.8%	30.0%	72.3%
35.0%	40.7%	40.2%	27.5%	69.8%
50.0%	33.8%	34.2%	21.8%	62.7%
# instances	51		100	

Table 6.4: Recall and Precision of 2- and 3-Person collaboration summarized over all 4 data sets with increasing recognition error. (Source: Gernot Bahle in [175])

6.8 Transferring the 5l-HLGSTP Model into other Domains

Although the method introduced above was designed and evaluated for the nurse emergency training scenario, it is not necessarily scenario specific and it should be possible to re-use this method for various other domains. The following section attempts to prove that the method of modeling a domain in the 5l-HLGSTP is transferable to other domains, with different complexities and different preconditions for inter-person interactions. To evaluate this transfer-ability, two other domains were selected:

- Assembling of furniture and
- Constructing of a video wall

A reason for choosing these domains is that both have the four factors described in the contributions (hierarchical compound activities, a high degree of freedom in the actual performance of activities, activities can mostly be performed individually or collaboratively, and there is no predefined role-structure) as a common denominator. Equal as for the nurse emergency training, for both additional scenarios following data was available:

1. Video footage of the entire experiment was recorded. Three cameras were placed at relevant corners of the "field of action" to record all activities done by the participants.
2. For each experiment, a description of the task was available (e.g., IKEA manual) that could be broken down into hierarchical compound activities and eventually basic actions. For the video wall model, some activities as taking the monitors out of storage and carrying them to the assembly area had to be added manually.

6.8.1 Assembling Furniture

The first additional scenario chosen is assembling a piece of IKEA furniture. This scenario is clearly much more straightforward than the nurse emergency scenario and allows the subjects to take as much time as needed to figure out what to do and how to collaborate. However, as anyone who has tried to assemble an IKEA cupboard before knows, this is not a trivial task. It involves a broad framework of what needs to be done when including many sequential constraints. At the same time, it leaves much freedom for the actual execution. There is no prescribed collaboration as most furniture can be built alone as well as in a team.

The Assembly Model Not surprisingly, the furniture assembly model is simpler and smaller in size in comparison to the nurse emergency model. It is comprised of 4 WCL nodes (“Unpack and Check Content”, “Assemble Frame”, “Assemble Compartments”, “Other”). Each of the WCL nodes has 2-4 SOL children who themselves branch into 1-3 SAL nodes. Each SAL Node is a parent of 8-100 AAL leaf-children. All nodes were connected with logic nodes, explicitly ordering the AAL nodes. The specific challenge in this scenario is that there is much flexibility in the order of assembly. For example, first apply pegs to every board and then put them together; or apply pegs to one board, then attach it to the frame; and so on. Therefore, up to 100 AAL leaves were necessary to describe a task with, only 10 actual atomic actions. For this second scenario, two persons were filmed while assembling an IKEA bookshelf (see Figure §6.15). The entire scenario lasted approximately 25 minutes in which both persons first worked individually and later started to collaborate.

Scenarios	WCL Ns	# SOL Ns /WCL N	# SAL Ns /SOL N	# AAL Ls /SAL N
Nurse Scenario	7	3-12	1-29	1-10
Furniture	4	2-4	1-3	8-100
VideoWall	7	2-5	1-3	2-16

Table 6.5: Model Overview: Number of nodes (Ns) or leafs (Ls) = children per node (N) = parent in each level for different domains.

6.8.2 Constructing a Video-wall

The second additional scenario describes building a large video wall. This video-wall consists of a total of 9 Monitors (3x3) that are attached to each other in the manner of a tower of three monitors on top of each other on the right side, in the middle and on the left (see [205] and [206] for a more detailed description). This essentially means a repetition of very similar tasks for each tower with slight differences in many steps. Moreover, even if this scenario seems to be straightforward, with some very strict sequences of activities (e.g., the base has to be installed first; otherwise the monitors cannot be mounted), there are still many options for flexibility (e.g., first install all floor parts then all retainers, etc, or install left tower first then ..., or install a bit here, then a bit there, and so on). The most significant difference to both other scenarios is that for many steps of this scenario two or more people are required, as the monitors are too heavy and bulky for one person to lift or carry.

Video-wall Model Like the nurse training model, the Video Wall Model incorporates 7 WCL nodes (“Transport Items”, “Assemble Left Base”, “Assemble Left Monitor Tower”, “Assemble Right Base”, “Assemble Right Monitor Tower”, “Assemble Middle Base”, “Assemble Middle Monitor Tower”). Each WCL node has 2-5 SOL children, which in turn have 1-3 SAL children. Every SAL node is a parent of 2-16 AAL leaf-children. For this scenario, two sessions were recorded on videotape (see Figure §6.16). All sessions included four persons, two male, and two female. The subjects were given general instructions on how the wall needs to be put together, where the parts and tools are stored, and where to put the wall up. All execution details were left to the subjects. No roles or collaboration instructions were given.

6.8.3 Evaluation of Scenarios

The same way as in the evaluation of the 3-layer model, the evaluation of the 5-layer model started with the baseline of a perfect recognition rate of 100% for each atomic level activity, with labels derived from video footage. To simulate the behavior during real-world activity recognition, which is always prone to noise and errors, the model and algorithm were evaluated with increasingly erroneous data (in the same way as for the nurse emergency scenario).

Since this work focuses on the recognition of collaboration, a uniform distribution of substitution errors could be assumed. In practice, this may make results worse than they would be otherwise since, in real-world activity recognition, similar activities tend to get confused a lot more than dissimilar ones. Similar ones, however, are often close to one another in the model, lessening the impact an erroneous recognition has. Since the simulation of errors is no longer deterministic, we performed 500 runs per error rate. For the error rate, we simulated a range of 5% to 75%, increasing in 5% steps.

6.8.4 Results

The following section provides a summary of the results of the evaluation of applying the 5-layer model onto the different domains. Since the evaluation of this part has been performed in collaboration with other persons, the results are briefly summarized, and picture and table

[1] Equally to the evaluation of the 5l-HLGSTP model of the nurse emergency scenario, the development, and design of the 5l-HLGSTP model of the furniture assembly and the 5l-HLGSTP model of the video wall construction was done by the author of this Dissertation. The evaluation of both 5l-HLGSTP models was done in collaboration with the first author of [175] thus the evaluation results again only are summarized briefly and meant to show that the method is transferable to other domains, which not necessarily have the same restrictions as the nurse emergency scenario, and still provides reasonable results.

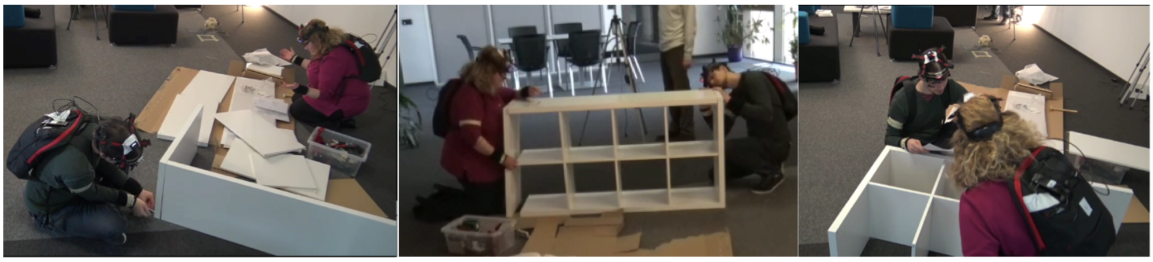


Figure 6.15: Two people assembling a shelf-board, not collaborating and collaborating

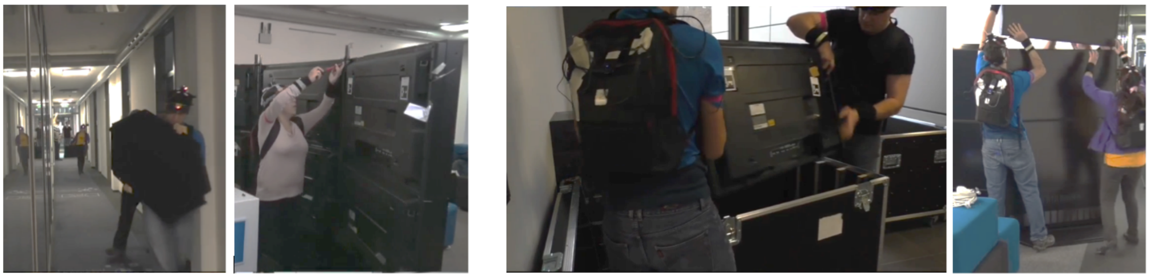


Figure 6.16: Four people building a video wall. At some points they work alone, for some tasks they have to collaborate.

sources are stated in the respective captions. [1]

Assembling Furniture

The analysis of the furniture data-set shows an average accuracy of 77% with precision of approximately 70% and a recall of 80% for correctly recognized atomic activities. See Figure §6.17. With these numbers, the furniture assembling scenario is roughly in the range of the 2-Person collaboration results of the nurse emergency scenario. Different to the other scenarios and because of the limited tasks to perform on this scenario, only two people were recorded assembling the shelf. Thus only collaboration of a maximum of two people could be calculated. As has been pointed out, with the nature of IKEA shelfs being relatively easy to built, initially both actors worked alone and only started to collaborate in the second half of the built. This obviously limits the amount of collaboration incidents.

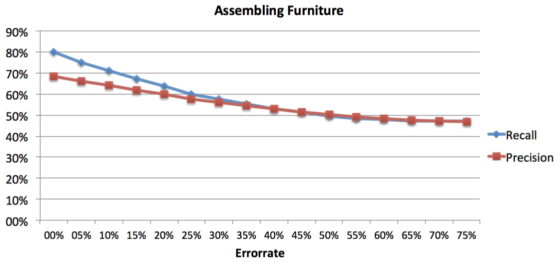


Figure 6.17: Assembling Furniture: Recall and Precision of detecting collaboration (7 possible classes) in relation to recognition error

Again, similar to precision of the nurse dataset, the results of the furniture dataset also shows robust behavior with only slowly decreasing accuracy with increasing recognition error. Overall, these results mirror the results of the nurse emergency scenario and thus strongly indicate that the proposed method works for at least one other domain.

Video-Wall

The analysis of the video-wall data-sets also provides similar results. Precision is highest for multi-person collaboration or collaboration with a higher number of data instances. In total as many as 9 combinations of interactions were possible (e.g. person 1 and person 2 and person 3 = P123, ..). Nevertheless, only 5 combinations happened often enough to produce a sufficient amount of data instances (combinations with less than 25 data instances were ignored). Meaning for example in the 25 Minutes of building the video wall and within 7 possible classes, person 2 and person 3 collaborated less than 125 seconds, which just does not provide enough data to reliably detect collaboration.

Figure §6.18 displays the 2-Person, 3-Person, and 4-Person collaboration average results. Different to the other scenarios, the video-wall construction was done by 4 persons, thus providing instances of 4-Person collaborations. These 4-Person collaborations provide explicitly well results with an average of both Precision and Recall of over 85% for correctly recognized atomic actions. This once more confirms the validity of the proposed method to be used in other domains than the one it was initially developed for. Like in both other scenarios, Precision also only decreases slowly with increasing recognition error, which indicates a robust detection algorithm.

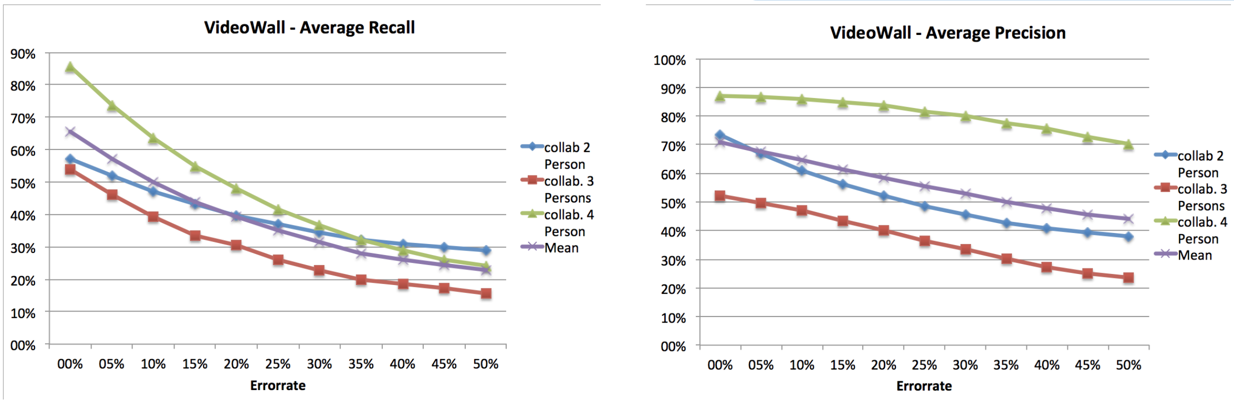


Figure 6.18: Building a video-wall: recall and precision of detecting collaboration (7 classes) in relation to recognition error.

6.9 Discussion, Outlook, Conclusion

The work presented in this chapter should introduce a new mindset in plan recognition, and as mentioned earlier, only a few initial steps should really be addressed. In this regard, the work in this chapter has some limitations that are bound to be resolved in the future.

6.9.1 Limitations

Model Creation: The main limitation of the method proposed lies, at first sight, in the necessity to "handcraft" the model. Nevertheless, as has been described earlier, the initial baseline model was derived from (partially existing) work-flows. Meaning, that if (a) workflow(s) of a domain is/are available, this new method shows how to turn work-flows into semantic trees. Trees themselves though are a common form of representation of information. For plans, they often can be generated from a variety of semi-formal descriptions, which includes texts. Moreover, in the text-parsing, trees are a common target structure. In this regard, the proposed method introduces a way to generate trees out of workflows semi-automatically. In the future, a (semi) automatic generation of models from, for example, textual descriptions of procedures should be possible. Some current student thesis works are addressing this problem right now ([207]). However, the work presented in this chapter, as was stated in the beginning, has to be seen as the groundwork, the first necessary steps towards research on automatically generated models. Models can only be generated automatically when it is known how these models are supposed to look like and which form of model suites to the intended purpose.

Leveraging Ground Truth: An issue that needs to be considered is the ambiguity of the ground truth in some situations. As an example: during building the video wall p1 is doing "install right base" (part of SOL "build right video wall") while p2 is doing "install left monitor" (part of SOL "build left video wall"). Thus, no collaboration happens between p1 and p2 regarding the definition of collaboration (collaboration within the semantic activity level - SAL). Then p1 stops the activity and starts talking to p2 (who continues working on their task in SOL "build left video wall," while talking to p1). After a short time, they stop talking, and p1 starts with "install right monitor" (part of SOL "build right video wall"). The main question in this sequence is, whether "p1 talking to p2" is a collaboration (they are doing the same activity "talking," yet does this imply they are collaborating on a task that is part of assembling the video wall)?
Furthermore, if this action is to be considered as a collaboration, which class should it be assigned to (p2 still doing a task, p1 was in between tasks and thus not contributing to the task p2 was working on). This example mostly shows the problems of putting complex real-life settings into formal models. In the evaluation, such cases were ignored and not counted as collaboration. Given their character as an exception, not a regular occurrence, this does not significantly impact the results.

Model Scale-Ability: In terms of scale-ability, the tree-based method comes to its limitations, as it is to some extend scale-able only on a high level. This tree-based method provides a mechanism to depict a scenario or a restricted domain to its very detail. Even-though it is generally possible to model more substantial and growing domains, due to the attention to detail, specifically in the lower layers (AAL, SAL), such models would quickly become very complex, requiring exponential space and time to process. However, the tree-based methodology itself was first developed with regards to the training scenario for nurses, and later transferred to the video wall and furniture scenarios. This shows that, despite scale-ability limitations, this method is generalize-able to domains of any size.
Apparently, the complexity of the approach increases exponentially with the number of people participating. However, this is only a limitation if the number of persons that really work together on atomic level activities becomes too large. The number of trees for a given number of persons n is 2n-1, so 7 or 8 participants are still tractable. If multiple groups collaborate both intra- and intergroup, then a hierarchical approach can be employed to avoid a combinatorial explosion. Thus, this chapter has developed a new way of "organizing plans" in a way flexible enough to be able to detect collaboration of multiple random agents in unstructured ad-hoc groups. Despite the limitation of the need to handcraft

the respective models yet, limited scale-ability and the evaluation via video retrieved labels, it could be proven that it is possible to detect collaboration in such ad-hoc groups with reasonable accuracy. This was the goal of this chapter! Mission accomplished!

6.9.2 Outlook

Nevertheless, research will never be done and finished. Thus, there is plenty of things that can be enhanced, adapted, improved in order to optimize the proposed methods in the future. For a starter, working with a patient (e.g., measuring blood pressure) does not necessarily mean that another nurse also measuring blood pressure collaborates with the first nurse if the other nurse measures blood pressure for another patient! The current model, as enhanced as it is, lacks the possibility to model specific devices or to point at specific persons. The same is true for location. The construction site for worker A does not necessarily mean the same construction site as of worker B. Furthermore, regarding detecting and understanding what a person is doing or thinking, the target of gaze can have a very distinct meaning, even when the person is not even close.
All these aspects - location, used devices, and specific persons and also the target of view of a person - could all be included in the model in the future. A possible way to do this could be in form of leave-attributes. For example, leaves or other nodes could specify which atomic action, or even entire semantic compound activities, have to take place at which specific location.(e.g., CPR has to be performed at the patient). This could even go as far as introducing a third dimension into the model to provide a measure to detail that particular actions have to be performed in the same location (e.g., chest compressions and operating AED have to be performed at the same patient). Figure §6.19 shows how this could potentially look like. Any kind of AAL-node could include a number of attributes, like location, view target, body posture. These in turn, can be organized with logic connections (from the logic meta layer). The manner of being an attribute to an AAL-leave and not leave on its own is indicated via the dash-dotted line between AAL-leave and attribute. These attributes, if it makes sense, can be linked together as a third dimension layer spanning all the other semantic activities.

6.9.3 Conclusion

This chapter has introduced a hierarchical tree plan based method to detect collaboration in random ad-hoc groups with a basic low-level action of compound activities that can be identified from noisy recognition of individual actions.

The results of the “postmortem” analysis of the proposed method using video footage, provide detection of collaboration on a high-level (world context level) in a range of 70-90% precision and recall with up to 20% possible error in recognized activities. On a lower level (semantic compound activities), results still are in the 70-80% range of precision and recall, with up to 15% possible error in recognition of the individual low-level actions. Deriving low-level atomic actions from video footage instead of real sensor-data and introducing randomly distributed errors is a limitation of this work. However, evaluating the proposed method goes beyond the boundaries of this dissertation and thus will be a part of future work.

Nevertheless, despite any limitations (discussed above), the results of this chapter are a strong indication that the proposed methods of modeling and detecting collaboration, are a practical approach for determining unscripted dynamic collaboration in complex tasks. The work in this chapter was meant to kick-off a shift and further developments in research from today’s standard of context-aware assistive systems that support individuals, to a new generation of “group aware” computing that supports groups of people working together more efficiently.

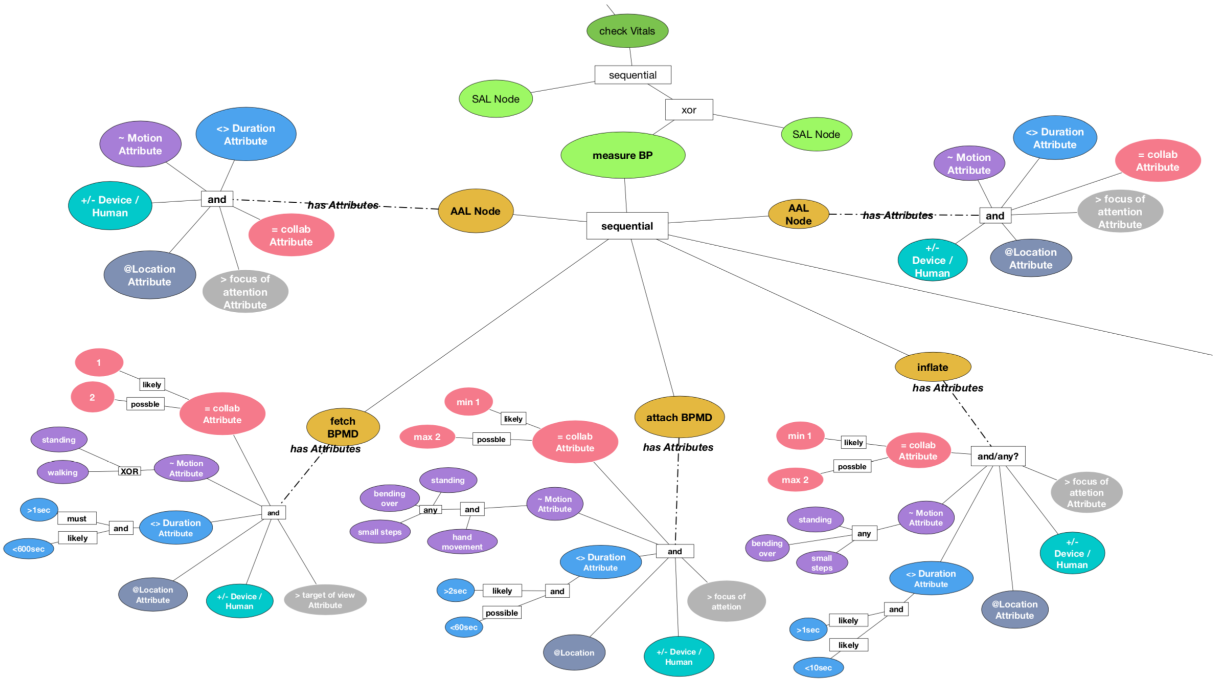


Figure 6.19: Possible adaptations to tree model

Conclusion and Outlook
What has been Achieved in this Work
What still Goes Beyond

Between the "Lightproject" from which the data sets for the detection of the condition and well-being of dementia patients were derived, and "iGroups" and "SmartNurse," the sponsor projects of the CPR-watch and the recognition of cooperation within groups almost a decade has passed. In this decade a lot has changed in the world of pervasive computing, and many devices have been developed. Things that seemed impossible ten years ago, such as equipping patients in a psychiatric clinic with motion sensors, suddenly became possible with the development of smart devices. Especially with psychiatric patients, the development of smart-phones has radically changed the situation.
The idea of analyzing movement patterns of bipolar patients with sensors to determine their condition is more than 10 years old. The idea itself was already well received at that time. Back then, however, nobody wanted to burden psychiatric patients with carrying a self-made sensor. With the introduction of the smart-phone, which already contains the necessary sensors, this was no longer a problem, since most patients would already own a smart-phone anyway.

However, since the beginning of the development of smart-phones and other smart-devices, new issues have formed. Even before "Snowden," many people argued that owning a smart-phone would make us "monitorable" and smartphones would allow respective bodies to spy on us. This direction is a disadvantage of recent developments that actually should simplify the needs of our "mobile life". Certainly, there is the possibility of being monitored without knowing it, particularly for those who do not understand the mechanisms of said technologies. Ensuring people's privacy though requires policymakers and legislation to impose appropriate and modest mobile communication rules.
This is beyond question. But it also lies in the responsibility of the researcher who follows new ideas and has an urge to develop new methods, to understand how these new methods could be abused and how this abuse could be prevented. Thus, while I was working on several studies with pervasive sensors, I was always aware that much of my work could be seen in the light of "espionage of humans" rather than an opportunity to gain "valuable insight." Therefore, at all stages of my work and in all the studies I have done in the health sector, I have, if possible, always tried to eliminate those aspects that could be abused (e.g., by anonymizing sensor readings) and to make people aware of why particular sensor readings were recorded and for what reason they were necessary. In addition, I have always tried to explain to the participants what measures have been taken to ensure that sensitive data is protected and handled with care. Nevertheless, it cannot be stressed enough that politics have to find a balance between guaranteeing their people's privacy and protection, while at the same time not by over-regulating, restrict and limit the potential, all these new devices could bring with them.

Modern smart-devices though, were not developed to spy on people in the first place, but to make our lives easier. The smart-phone (and all other pervasive smart-device) with its features of being "worn" on our body or really close to it, and of being connected to the world of the Internet, brings endless features that can make our lives more comfortable. To go back to the example this thesis started with; it has never been so easy to track our fitness! Moreover, it has never been so easy to learn or retrieve information and check if statements of people were correct. And this is due to our smart-phone.
Furthermore, it has never been so easy to change bad habits as smart-devices can prompt us based on context to get up and move (like fitness-tracker or equivalent apps on the smart-watches do), or provide the safety net we need to feel comfortable enough to perform CPR even though we have never done it before. The list of examples can go on an on.

Thus, researching in these days and particularly in this field is immensely exciting. Nearly every year, with new devices and new functionalities, new possibilities open up, which allow addressing new and old requirements and questions of our lives. Many of these technological developments mentioned above happened over the last decade. Even though there have been Palms and Handhelds in the early 2000s, the success-story of smart-devices started with the development of the iPhone and Google's decision to jump in on this trend. With the possibilities, these new technologies brought with them, also this dissertation evolved over the years with the new arising opportunities and with new gadget the overall picture of this thesis was shaped.

7.1 What has been Achieved - Discussion

This dissertation aimed to develop methods for using sensor systems to support cognitive status assessments and to support people in psychologically stressful situations. The underlying research in this dissertation, therefore, began by asking whether sensor data from our ubiquitous environment would contain relevant information about our cognitive state and behavior.

Therefore, the work began by **analyzing the indoor location data of dementia patients** and **retrieving information from those data-sets to derive the patient's condition**. In this way, methods for monitoring the general well-being and progression of dementia could be developed, which relies solely on simple, unobtrusive indoor location sensors. The **accuracy achieved was over 90% correct classifications**. In the following chapters, this work aimed to **derive the mental state** in patients with bipolar disorder (**70-80% accuracy**) and **determine the onset of state changes (95% accuracy)** by only evaluating the readings from smart-phone sensors. This analysis went so far that the results of this newly developed condition assessment based on sensors could be compared with the standards currently used in mental health care. Besides, this analysis could **verify known but never-evaluated deficiencies of the standard method** currently used in mental care. Furthermore, it was possible to demonstrate the clear superiority of the newly developed () sensor-based method.

In the further course of this work, methods for teaching skills and supporting cognitive learning were presented and evaluated. Two studies demonstrated that **people with the right instant assistants can acquire instant skills** - the first study showed improved effective CPR performance by more than 50% when using instant feedback devices. In the second study, people **could achieve significant improvements** in the same way with the use of appropriate assistant devices **in the training of CPR**. On average, when training with the assistant device, the ability to correctly perform CPR without the aid was improved by more than 20%.

The final part of this work then focused on recognizing human interaction on an invisible level. In addition, a new method was introduced that can be used to deduce when people work together in different ad-hoc environments, without resorting to well-known interaction mechanisms (eg, talking, etc.). This type of research is just at its beginning, and the corresponding chapter in this thesis gives only a first glimpse of what possibly is there to come. The "post-mortem" evaluation of the first and simplest method using data extracted from real-world video footage **showed high accuracy (over 90% precision and recall) for various combinations of collaboration**, and yet, with an introduced recognition error of **up to 20 % a recognition inaccuracy** of the collaboration recognition still works (with precision of 70 % and 90% recall).

This summary overview makes clear that even though this thesis started with an initial research-question, the work over the years increased in the complexity of its objectives from chapter to chapter. Furthermore, also the underlying fundamental research question evolved with the progression of this work.

7.1.1 Level of Complexity

In the context of the presented work, the first and oldest chapter proposes the most simple complexity, by also using the most simple sensor system throughout the thesis. Only by using one stationary sensor-system deployed in a static and well-defined space, state of (dementia) patients has been determined. Thus, the general strong point and the particular value of this first chapter neither can be found in its complex set-up or even in the surprisingly high recognition accuracy. The actual strong point is the **real-live data set collected in a long-term (almost one year long) daily data recording and the strategies developed to leverage minimally available ground-truth**, which both are difficult to obtain.

In comparison to the first chapter, the second chapter means a notable increase in the complexity of sensors used and the study set-up in general. It deals not only with a variety of sensor-modalities, but it also was deployed in an open area in the everyday life of people. Thus, it, even more, comes with minimal possibilities to gather ground-truth. In this regard, the general strong point and the particular value of the work in this chapter are to have been able to **achieve reasonable results with an application deployed in the real everyday life of mentally ill people, regardless of all possible challenges, drawbacks, and data losses imposed by the real-life deployment**.
However, the particular highlight of this work is that, despite all these complexities, it has been possible to develop a high-performance algorithm to detect actual

changes in mental states of patients, while meeting the limitations and challenges of real-world set-up. This, in general means that the change detection for bipolar patients is actually suitable for everyday use and thus goes beyond mere research!

From chapters 2 until 4, the complexity of this work increased regarding sensor-setups and space. In chapter 5, on the other hand, the complexity of the topic morphed its angle. It changed from **retrieving information** from (a varying number of) sensors into **using sensors (or the respective smart-device) for providing information**. The specific complexity here was, and this might sound contradictory at first, to keep it simple. Because an assistant system that can work anytime and instantly, ideally has to rely only on one sensor (device) that is present unobtrusively at all times. Furthermore, since the goal was to provide an instant CPR assistant, the way information is provided, has to be tangible at one glance without long learning periods. Even though the studies with the dementia patients and the bipolar patients were more complex regarding time and effort to gather the required data, the sensors-systems to collect the data themselves were straightforward. For the CPR-application **extensively more thoughts had to go into the design and the way a rather complex motion as performing CPR with all its relevant parameters** (a correct combination of speed and depth) including instructions could be displayed on one small screen (watch) in a way the user has a chance to grasp it instantly.

In chapter 6, the complexity once again changed its direction. As until then, the goal was to detect the cognitive state of or support one person, in chapter 6 the goal became to understand the cognitive behavior of a group of people and not of a single person. Thus the complexity is more subtle. After a series of successful attempts to understand what a single person does, it seemed much more interesting **to investigate what a group of people is doing**. Building on the results of all previous chapters, it was feasible to assume that this is possible. The real challenge in this chapter was to develop methods, relying on the assumption to know what every single person is doing, **to understand when and how multiple persons interact and collaborate**, meaning working together to reach a common goal and do this without relying on the standard interaction cues. Moreover, the base scenario predetermined that the group of people was ad-hoc without having anything pre-defined. Thus the question here was: "How can we, by knowing what every single person is doing, determine if multiple persons are working together as a group or if single persons are working alone?" In comparison to all other chapters before, this means going a step further and turn the angle, because first and foremost it was essential to understand and define **how collaboration of two or more persons looks like while they are performing a task**. At this point, it is not about to know what a person is doing (which has been dealt with in the previous chapters) but to **use the information of what a person is doing to determine which parts of this person's activities exactly are parts of a joint-collaboration**. Of course, this is not a trivial question, and the answer is not simple or such that it could be satisfactorily addressed in a single chapter of a paper. On the contrary, this thesis should mark a starting point in this regard, which in the future will stimulate comprehensive developments.

7.1.2 Research Question

In the course of this doctoral thesis, not only the complexity of the work has increased and developed, but in particular, the research question. As mentioned earlier, the fact that the research question could evolve was primarily due to the results achieved, especially at the beginning of this work. However, also the development of penetrating devices in the last years had a distinct effect on what was ultimately possible:
Particularly at the beginning and over the first half of this work, the initial research question remained **whether it would be possible to retrieve information about relevant aspects of our cognitive state from everyday activity sensor data.** This question in itself, but especially in the period in which I tried to answer it, could have filled a whole dissertation. In today's society, this question is of particular relevance. Once the necessary steps were taken to evaluate these aspects, **the hypothesis on this question (Hypothesis 1 of this dissertation) could be confirmed**.

Based on the positive outcome of this first question, the second hypothesis went so far as to state that it is even possible to derive complex cognitive states and to predict or recognize pathological cognitive changes from information extracted from pervasive sensor data. Again, several chapters of this work have shown that **this hypothesis (Hypothesis 2) can be approved**.
However, the development of new devices (e.g., smartwatches) and the confirmation of Hypotheses 1 and 2 have further fueled the research question. Thus, the third hypothesis changes the point of view on this topic and raises the question of whether given sensor systems can determine not only the cognitive state but also whether **sensors (or suitable devices including such sensors) are capable of being a positive support for cognitive behavior.**

A study by Lally et al. [208] has shown that it takes between 18 and 254 days (avg. 66 days) to form or change a habit. This casts a light on how difficult it can be to change bad habits. Thus the possibilities that have opened with the usage of smart-devices that accompany us every day, to help to influence our cognitive behavior to the better is promising, and it was possible to show that smart-devices actually can have a positive influence. Thus, at the end of this dissertation, also its **third hypothesis could be affirmed**.

7.2 What is Missing, What will Come Next - Limitations and Outlook

The beauty of our world is that it should change and evolve. Likewise, research will never be over. As long as we live, there will always be something to improve, there will always be a new research question, and a dissertation will never be able to fully and satisfactorily exploit a research topic. So, where does this thesis have limitations in terms of our cognitive state and our mental well-being, what could not be addressed, and what will - what should - come next?

The theme of this work is not self-contained. On the contrary, it covers a wide range of cognitive aspects and sometimes only scratches the surface. All the different parts (chapters) of this dissertation form an independent topic in their own right and could fill their own dissertation with further work. Thus it would be possible in all chapters to go deeper and to answer further questions. Many of these have been listed in the “Discussion” or “Outlook” section of the corresponding chapters. To summarize this dissertation nevertheless, an overview should provide where the limits of this dissertation are and what can be expected in the future.

In the indoor-location based assessment of the well-being of dementia, the main limitation of the study was the rough level of available ground-truth. Therefore, in terms of what would be possible with a more detailed ground-truth, there are still some questions left. For example, with a more sufficient ground-truth would it be possible to determine state and progress on a daily basis and develop algorithms that would allow us to determine when the condition of a dementia patient is about to begin to change? In addition, the on-ward approach to maieutic care is becoming increasingly popular. It’s living conditions resemble a typical family home in which an older person would live with relatives. Therefore, the results of the study are expected to be transferable to private living environments. This means that such systems could already be installed in private homes of (early state) dementia patients, leaving them relatively independent and able to remain in their familiar environment. However, this still needs to be confirmed.

Concerning the patients with affective disorders, methods were developed and evaluated in this work to extract information relevant for the determination of the condition and the change of state. The results of the "postmortem" assessment of state change detection are impressive. Thus the very plausible theory is that early detection of a state change could help treat affective disorders and keep (negative and sometimes devastating) effects in check. However, this theory has not yet been evaluated, since this dissertation could only provide a postmortem evaluation of data. Such an assessment would require a multi-year, blinded, controlled study that clearly goes beyond the limits of any chapter in a dissertation. Since the topic of affective disorders (depression is part of affective disorder) is of particular importance in our society, such studies will be conducted in one way or another in the near future.

Although Chapters 2 to 4 contain relatively lengthy studies to collect the data sets, all these topics generally lack long-term studies spanning several years. Such long-term studies would probably be able to provide more comprehensive cognitive knowledge about disorders. Nevertheless, after the ground-work of these chapters (mainly Chapters 3 and 4) had been completed and published, research into depression and mood disorders began in various parts of the world. This is reflected by the number of citations work in this chapters have received to-date [91] (181 and counting - 05/25/19), [92] (78 - 05/25/19), [209] (115 - 05/25/19). Thus, even after completing this dissertation, many open research questions will be dealt with in these working groups.

In the same way as the application to detect the early onset of state change, the CPR-assistant application has practical relevance and has the potential of becoming a real-world product in a variety of settings. Possible uses are ranging from a training-assistant to an actual CPR-support device for any paramedic unity. These, of course, are more business related applications. Research-wise, the CPR-assistant offers umpteen other options. Two of them quickly come to mind.
First of all, there is “training-support systems”. In this dissertation, it was demonstrated that the CPR-assistant could help to quick and effective learning. The question remaining is, how long the effects last, or how many training sessions with the assistant would be required to make the training effect permanent. This dissertation could only evaluate the immediate effect of training with the assistant-devices in comparison to standard teaching but did not evaluate how long the effects would last. This limitation does not take away from the value of the results, because it entails information that actually is required to understand how these assistant devices could be deployed in the future.

Furthermore, even leaving the area around CPR, many other scenarios would benefit from instant feedback assistants. In the future, teaching/training assistants will become a whole branch of research dealing with questions such as: which scenarios could benefit from instant training assistants? In which areas could smart devices help people learn more effectively? What should training assistants look like? These questions are already being worked on by different researchers (e.g., [150, 151], etc).
Additionally, as was mentioned in Chapter 5, the fact that the CPR-watch app can provide people with instant skills should be exploited. A possible application that comes to mind is the co-operation with other emergency devices. For example, together with the publicly available AED devices, an emergency system for inexperienced people could be developed. Such a sys-

tem could leverage the ability of the CPR-watch to understand whether CPR is performed effectively, together with the AED's ability to understand if the patient lives. This would contribute to making sure all necessary emergency activities are being performed as they should. Even though again this seems to be a professional product, even for such a combined CPR-Watch/AED system, research would be challenged to understand how such systems would need to be built, in order to be effective and how such a system could change the emergency behavior of laypeople. Very positively, this could seriously contribute to many more survivors of OHCA in the future.

The outcome of the final chapter could be summarized as a new way of "organizing plans" flexible enough to be able to detect collaboration of multiple random agents in unstructured ad-hoc groups. As has been pointed out often enough, this work was the first step, and there is plenty of room for improvement. The limitations of the proposed method are discussed in detail in Chapter 6. In addition, the requirement to hand-craft the model for each domain or the limited scale-ability should be considered. Since the results of the method are promising enough, it seems to be worth to invest some time to develop mechanisms that will allow crafting the model automatically, or in a first step at least semi-automatically. For example, this could be done by extracting and logically structuring relevant text passages or sentences out of domain descriptions. The other major limitation, however, is that the method has not yet been evaluated with sensor data, but with data retrieved from video material. Since various chapters in this dissertation have demonstrated (and in many other activity recognition publications [199], [59], [210], [22], just to name a few) that data based on atomic actions can be expected to be available in one way or another, the focus has been on the evaluation of the model and the respective algorithms for the recognition of the collaboration itself, irrespective of where the data used would come from.
Another step towards systems that can assess the subtle aspects of human cognitive participation in group activities and cognitive collaboration among individuals is to achieve similar results using basic atomic-action-data obtained from real-world sensor readings. However, this implies another aspect, since the activity detection (even if it works) is continually looking for ways to increase the recognition accuracy. Here, the tree-based model of a domain might come into play, since from a detected action, the model could tell the classifier which actions would be the most likely next actions to take. Thus, this model would help to increase the recognition accuracy. However, if such collaboration of the model and recognition classifier is to work in a loop of sensor data rather than video captions, the model itself must be further improved. For example, the use of object identifiers, location information, voice interaction information, and/or posture data that are essential to specific atomic actions could be included.

7.3 On a Final Note

Pervasive and Wearable Computing in health care is a vivid and exciting field. The opportunities, technological developments bring us today (2019), will allow us to develop many more applications in the future that will help people to manage their health and allow personalized health care. Particular, in the last five years, the deployment of wearables has even reached mental health. This research will open up new unprecedented opportunities in mental care and management of cognitive health. With this dissertation, in particular topics, I hope I have contributed to lay a foundation for much future research.

Nevertheless, as has been mentioned earlier, with all these new opportunities and developments, as researchers, we have also a responsibility. Today we are not only living in a world filled with technology but also in a world where people feel uneasy because they are unable to keep track of developments and what they might mean for them personally. Science and research are called to take these fears seriously. In all excitement to follow a new idea and to develop a new application, we must not forget that all technology is supposed to serve humanity and not vice versa. We should not allow respective bodies to abuse our developments for espionage on people or allow to misinterpret our findings for lying to people. On the other hand, we are supposed to help people learn how to understand science and interpret statistics. Because, since many people lack the understanding of what science and statistics really mean, they often are either very skeptical or blindly following. Science or research, nevertheless, are neither the enemies nor the holy grail. Essentially, science paired with prudence and responsibility is, what it is - an opportunity for humanity to evolve!

In this regard, I have done some groundwork for developing systems that should **help people with dementia to lead an autonomous life** as long as they possibly can. This is an issue that engages many people.

With the help of colleagues and partners from mental care and the trust of psychiatric patients, I developed methods for mentally affected people **to understand if and when their conditions change** for the worse and thus **provided them with the possibility to act in time.** With the rise of mental disorders, this denotes an essential and required support.

Thanks to the invaluable support of nurse teachers and nursing students, I was able to develop mechanisms that **will allow untrained laypeople** in the future to provide adequate emergency support and thus save lives. Moreover, these mechanisms **will help emergency care stu-**

dents to gain the necessary skills faster and more straight forward. Emergency care and the lack of skills in many laypeople have always been a relevant topic.

Eventually, I was also able, in this dissertation, to kick off a new direction of work in analyzing group behavior. Developing this initial methodology based on an emergency care training scenario was not a random pick. Particularly in emergency training but also in other training environments, teachers are called to understand which of their students can proceed and who needs more support. Thus, nurse teachers have often told us that they would love to have objective support in evaluating how and if the students collaborate and interact during a training session. The method to detect unscripted collaboration in ad-hoc groups is the very first step on the road, which eventually **should help trainers to understand better** their students' needs.

The variety of projects covered in this dissertation shows that supporting cognitive state and behavioral analysis today is a wide field of many needs. I hope that in this dissertation I have never lost sight of the goal of serving humanity and improving our lives, but that I have also helped to solve many of the problems that still complicate health and especially mental care.

Part

Appendix and Directories

Appendix

Appendix to Chapter II - Indoor-Location:

Patient description

- **1010**: Inhabitant is female with an MMS value of 12 (maximum 30). The overall performance differs. Inhabitant often seems to be tired. Gets up rather late. Walking abilities are slower than normal but existent. This inhabitant normally goes directly from her room to a table in the kitchen area, sits there and goes straight back to the room again when she wants. Zone in which she moves is small. In some phases she likes to sit in the company of inhabitant 1031. It is not clear if this preference is returned by 1031. Partially participates in activities.
- **1031**: Female inhabitant with an MMS value of 15.5. 1031 is handicapped in her walking abilities and uses a walking frame. Inhabitant has very steady habits. Sits the whole day at the same place in the kitchen area (or on the terrace in summer). The movement zone is very small. Does not want to return back to room before going to bed. Partially participates in activities (especially painting and handicraft work) but normally prefers to sit and watch independent of personal state. Sometimes inhabitant seems to sleep but normally responds immediately. Does not express state by activity and movement.
- **1032**: Female with an MMS value of 0, high degree of dementia. Inhabitant likes to walk around and talk but normally without sense or coherence. Her movement zone is the whole ward, including the rooms of the other inhabitants. Likes to sing and participate in activities. Inhabitant is very curious and disassembles things. Seems to express state by activity and movement.
- **1041**: Female. The MMS value is 11. Normally feels the necessity to wander around and touch things but does not want to be touched or held. Often lays Tag aside. Often seems to be helpless like a little child. It is not clearly distinguishable whether she expresses her state by movement. State seems to be best when she is comparatively quiet and does not feel the necessity to walk around.
- **1060**: Female with an MMS value of 16.5. Seems to be partially insecure and seems to realize her disabilities. Inhabitant seems to notice her surroundings and is oriented. Uses a walking stick but is mobile. Inhabitant undergoes phases where she withdraws to her room (which seem to be connected to her realization of her disabilities). In other phases she is social and participates in activities. It is hard to distinguish whether there is a connection of movement and state, except for time of stay in the social areas like living room and kitchen.
- **1090**: Male inhabitant with an MMS value of 13. Often disoriented without sense of time. Likes to eat. The inhabitant is mobile and walks without aids. His performance is alternating. On some days he likes to stay in his room. On other days he remains in the social areas. Partially wears Tag and partially opposed to.

Questionnaire

The following table summarizes the questions ask to the nurses at the end of the one year study, and provides the summarized averages for all possible reply-categories and the positive and negative replies. The values in this table are in % except for questions where explicitly was asked to state concrete numbers, which are averaged in the Mean column. Note that the questions listed are in abbreviated form to save space.

Possible Answer:	*No*	*Hardly*	**Sum**	*A little/*	*Yes*	**Sum**	
Questions:	%	%	**negative %**	*mostly* %	%	**positive %**	*Mean*
About the study deployment							
Additional burden for the work/extra work							
Was there any?	22	44	**66**	11	22	**33**	
Min/day							**15**
Visible installations (wires, wall-mounted sensors, . . .)							
Annoying?	12.5	25	**37.5**	25	37.5	**62.5**	
Got used to it?	0	0	**0**	22	78	**100**	
Disturbed work?	22	78	**100**	0	0	**0**	
Sensor tags							

Had impact on residents?	12.5	37.5	**50**	50	0	**50**	
Initially or long-term	(Initially)		**40**	(Long-term)		**60**	
Equally for all?	62.5	12.5	**75**	25	0	**25**	
Had impact on personnel?	12.5	12.5	**25**	62.5	12.5	**75**	
Impact personally?	22	22	**44**	22	33	**56**	
Convincing residents to wear a sensor tag							
Problems to convince?	11	0	**11**	78	11	**89**	
Did they get used to it?	0	22	**22**	33	55	**78**	
How may did wear tags?							**9.38**
How may did not wear?							2.2
Did sensor maintaining personnel							
Disturb your work?	33	56	**89**	11	0	**11**	
Disturb the residents?	44.5	44.5	**89**	0	11	**11**	
Which mode was more annoying additional observations (1) or sensors (2)?							
Personally?	*Observation*		**57**	*Sensors*		**43**	
For residents	*Observation*		**0**	*Sensors*		**100**	
For visitors?	*Observation*		**100**	*Sensors*		**0**	
About Using Technology and study in general							
Is it possible that the study contributes to enhance quality of life of dementia patients?							
	0	6.25	**6.25**	93.75	0	**93.75**	
Do you think, it is possible to draw useful conclusions out of sensor-based monitoring							
	0	0	**0**	78	22	**100**	
Is it possible that deployment of technology can relieve your work?							
	11	11	22	67	11	78	
Possible Answer:	***No***	***Hardly***	**Sum**	***A little/***	***Yes***	**Sum**	
Questions:	%	%	**negative %**	***mostly %***	%	**positive %**	***Mean***

Table 1: Questionnaire handed out to the nurses working in the ward during the study period

Appendix to Chapter II - Bipolar Patients:

Study Participants

Overall a total of 12 bipolar patients participated in the data collection trial over the period from November 2012 to August 2013. Some patients dropped out early (p0202 and p0602), some (p0502 and p0802) even extended the trial. The evolution of the state of the individual patients during the trial is shown in figure §3.5 in the respective chapter. Note though, that patients p0202 and p0402 show no change of state during the entire trial period (with patient 0202 dropping out of the data trial early). As a consequence they were not considered in the evaluations. Thus eventually only 10 participants are considered in this study. Key information about state and progression of condition of these 10 patients is summarized below:

- **P0101:** the patient is female, age around 50 and was diagnosed as manic in different degrees almost until the end of the study. Despite, the patient herself believed to be either depressive.
- **P0201:** the patient is female, age around 40 and was diagnosed slightly depressive, changed to normal state during the first third of the study. The patient stayed in normal state until the last month and dropped to very severely depressive at the end of the study.
- **P0102:** the patient is female with age around 40. She was diagnosed depressive at the beginning of the study, improved to normal state during the first month, yet dropped back to severely depressive after a few weeks.
- **P0302:** the patient, again female, age around 50 was diagnosed with very severe depression at the beginning of the study. Over the course of 1.5 months this patient improved step-wise to depression, slight depression and eventually to normal state at the end of the first half of the study. The patient stayed in normal state throughout the entire second half of the study.
- **P0502:** The patient is female with age around 45. The patient entered the study, almost normal, right after a longer closed-ward stay due to mania. The patient was discharged 2 weeks after the beginning of the study and, being back home after a several-month stay at the hospital, dropped instantly to a very severe depressive state. This made it impossible to perform the t2-measurement-point, as the patient was unable to come to the hospital. During the second half of the study the patient stepwise improved to depressive, slight depressive and finally normal state at the end of the study.
- **P0602:** The patient was female at an age around 55.

She was diagnosed almost normal (very slight depressive) at the beginning, worsened slightly after two weeks, yet improved to normal state at the end of the first half of the study, where the patient quited her participation.

- **P0702:** the patient was female at an age around 25. Mainly, the patient was diagnosed with very severe depression, yet the scale tests show partially manic drive in her behavior. During the study the patient underwent changes mainly within the depressive state, almost reaching normal state at the middle of the study, yet dropping back again to depression.
- **P0802:** This patient was the only male participant. He was at an age around 45 and was diagnosed with very severe depression. Within the first 3 weeks the patient improved to slight depressive state and eventually normal state right after. He stayed in normal state until the end of the study.
- **P0902:** Another female patient, at an age around 30. When this patient entered the study she was already in normal state where she stayed throughout the first third of the study, then dropped to depressive, improved again to slight depression and bounced back to depressive at the end of the study.
- **P1002:** this patient was female, age around 25. She was diagnosed manic at the beginning of the study, yet dropped to slight depression and further to depression at the end of the study.

Classification Results

The following table provides precision and recall results for each patient within each class for Location, Acceleration and the Fusion of both. A close look at the individual precision/recall values in reveals the reason why location performs best (see also results in respective chapter). Since there is not enough data the location classifier does not consider medium depression for patient p0502 for example, which is very poorly recognized by the other classifiers. Overall the fused approach has the advantage of considering more data points than either acceleration or location alone since, it considers data points covered by either modality.

Patient		**Recall**			**Precision**	
(N per state)	Fusion	Location	Acceler.	Fusion	Location	Acceler.
p0101		%			%	
normal (32/3/32)	84	2	74	75	1	88
slightly manic (23/20/27)	55	80	75	65	90	59
med.manic (12/3/12)	67	67	80	67	63	79
p0102		%			%	
depressive (12/10/12)	94	84	81	86	86	83
normal (26/26/26)	62	77	55	80	73	51
p0201		%			%	
depressive (33/21/33)	25	75	40	50	48	41
normal (13/13/13)	89	81	76	72	86	75
p0302		%			%	
depressive (18/11/18)	39	94	42	100	64	32
normal (42/40/42)	100	92	73	80	99	80
p0502		%			%	
sever depressive (14/5/14)	50	50	55	58	74	62
depressive (14/0/14)	35	0	56	71	0	39
normal (30/23/30)	69	82	81	85	57	87
p0602		%			%	
slightly depressive (13/12/11)	85	58	57	65	63	69
normal (22/19/10)	73	78	61	89	74	47
p0702		%			%	
sever depressive (34/24/24)	91	82	82	80	94	85
slightly depressive (8/7/8	0	21	22	0	7	18
p0802		%			%	
depressive (16/7/16)	44	70	58	64	57	49
normal (46/30/46)	91	93	84	82	96	88
p0902		%			%	
depressive (26/26/26)	92	83	73	83	82	76
normal (15/15/15)	67	70	57	83	71	53
p1002		%			%	
depressive (29/11/29)	86	68	75	74	91	83
slightly manic (11/9/11)	18	87	32	33	57	22

Table 2: Precision / recall values for the different states. Most patients experienced 2, some 3 different states during the trials.

Appendix to Chapter IV - SensorVSSelf:

Correlation of Features

In Chapter 5 only an overview of features was given. The following table shows more features for Location, Accleration, and Phone-calls with correlation and t-value for each patient.

	Location			Acceleration			Phonecalls		
patients	correlation	t-value	doF	correlation	t-value	doF	correlation	t-value	doF
Feature	distance traveled			High/low			number of phone calls		
p0101	-0.604	-2.393	12	-0.05	-0.26	12	-	-	-
p0201	0.061	0.312	28	0.496	2.609	28	0.012	-0.164	27
p0102	0.737	5.118	24	0.428	2.341	25	0.367	1.978	32
p0302	-0.339	-2.247	41	0.272	2.006	45	0.248	1.642	47
p0502	0.599	3.349	22	-0.732	-7.198	31	-	-	-
p0602	0.602	3.369	22	-0.923	-7.147	14	-0.717	-5.034	25
p0702	-0.505	-2.026	14	-0.215	-0.692	21	-0.505	-2.05	21
p0802	-0.662	-4.324	26	0.249	1.792	47	-	-	-
p0902	0.123	0.679	32	0.468	2.778	28	-0.423	-2.202	29
p1002	-0.821	-2.492	5	-0.126	-0.891	15	0.765	3.353	14
Feature	number of clusters			frequency variance			number of unique numbers		
p0101	0.778	3.912	12	0.308	-3.075	12	-	-	-
p0201	0.662	4.509	28	0.554	7.291	28	0.106	0.408	27
p0102	0.878	8.6	24	0.248	-2.268	25	0.476	2.815	32
p0302	0.289	1.887	41	0.349	1.675	45	0.561	4.431	47
p0502	-0.201	-0.919	22	0.131	-0.129	31	-	-	-
p0602	0.145	0.653	22	-0.85	-5.046	14	-0.635	-4.342	25
p0702	-0.802	-4.656	14	0.764	3.586	21	-0.282	-0.708	21
p0802	0.338	1.761	26	0.26	-4.917	47	-	-	-
p0902	-0.227	-1.274	32	0.307	2.83	28	-0.362	-1.666	29
p1002	-0.913	-3.864	5	-0.805	-6.337	15	0.754	3.179	14
Feature	percentage of stay outdoors			rms variance			average length		
p0101	-0.579	-2.248	12	-0.302	-1.744	12	-	-	-
p0201	0.505	2.987	28	0.534	2.817	28	-0.375	-1.347	27
p0102	0.887	9.003	24	-0.848	-7.083	25	-0.47	-2.694	32
p0302	0.412	2.82	41	-0.721	-6.089	45	-0.102	-0.186	47
p0502	-0.187	-0.852	22	0.668	5.469	31	-	-	-
p0602	-0.259	-1.199	22	-0.374	-1.097	14	-0.442	-2.572	25
p0702	-0.748	-3.898	14	-0.068	-0.483	21	0.687	3.998	21
p0802	-0.354	-1.852	26	-0.755	-7.483	47	-	-	-
p0902	-0.545	-3.558	32	0.073	0.34	28	0.225	0.993	29
p1002	0.117	0.203	5	0.416	1.754	15	0.672	3.482	14
90% C	|t| >= 1.782			|t| >= 1.782			|t| >= 1.76		

Table 3: Correlation results for different features per participant.

Appendix to Chapter V - CPR Training:

The following two tables summarize the performance of each participant (student, novice) at the beginning, after training/teaching and at the end after teaching/training.

			teaching first						training first			
	start		after teaching		after training		start		after training		after teaching	
	average		average		average		average		average		average	
N	29.7		29		27		28.7		29		29	
depth	981.7		932.3		532.1		883		977.1		944.2	
speed	126	cpm	125.4	cpm	101.7	cpm	135.3	cpm	142.2	cpm	134.8	cpm
too shallow	2.2	%	48.3	%	90.7	%	65.5	%	3.4	%	64.4	%
too fast	61.5	%	55.2	%	2.7	%	100	%	100	%	100	%
too slow	3.4	%	0	%	54.8	%	0	%	0	%	0	%
ideal	32.9	%	12.1	%	8	%	0	%	0	%	0	%
N	28.7		29.3		29.3		24.7		27.7		30.3	
depth	737.2		555.2		975.2		971		960.9		970.1	
speed	114	cpm	94.5	cpm	107.7	cpm	133.4	cpm	143.7	cpm	133.4	cpm
too shallow	100	%	100	%	0	%	11.8	%	32.9	%	6.6	%
too fast	17.6	%	0	%	0	%	97.7	%	100	%	96.7	%
too slow	2.3	%	75.9	%	1.2	%	2.3	%	0	%	0	%
ideal	0	%	0	%	98.8	%	0	%	0	%	2.2	%
N	25		29		29		30		28.7		27.7	
depth	953.1		981.9		986		332.4		934.3		953.3	
speed	124.5	cpm	120.9	cpm	118.6	cpm	120.7	cpm	104.5	cpm	110.3	cpm
too shallow	3.4	%	0	%	0	%	100	%	54	%	27.7	%
too fast	52.2	%	47.1	%	20.7	%	12.3	%	0	%	0	%
too slow	2.8	%	0	%	0	%	86.7	%	0	%	3.6	%
ideal	43.9	%	52.9	%	79.3	%	0	%	46	%	68.6	%

N	28.3		28.3		21		29.3		29.3		29.3	
depth	965.3		750.4		646		967.3		978.2		974.8	
speed	96.1	cpm	113.9	cpm	102.3	cpm	126.4	cpm	128.6	cpm	113.4	cpm
too shallow	19	%	100	%	100	%	1.1	%	0	%	0	%
too fast	0	%	23.1	%	0	%	65.6	%	81.6	%	3.2	%
too slow	66.3	%	10.9	%	19.4	%	1.1	%	1.1	%	1.1	%
ideal	33.7	%	0	%	0	%	31	%	17.2	%	95.7	%
N	27.7		27.7		30		29		28.7		29.3	
depth	943.8		971.5		976.3		960.2		951		928.1	
speed	140.1	cpm	136.5	cpm	120.6	cpm	124.3	cpm	110.6	cpm	102.1	cpm
too shallow	66.7	%	0	%	1.6	%	0	%	0	%	76.7	%
too fast	96.4	%	100	%	1.1	%	64.4	%	1.2	%	0	%
too slow	3.6	%	0	%	2.2	%	1.1	%	0	%	13.3	%
ideal	0	%	0	%	96.7	%	34.5	%	98.8	%	21.1	%
N	29.3		30		28.7		27.3		29		30	
depth	958.5		978.9		957.4		942		984.8		967.3	
speed	122.7	cpm	124	cpm	127.5	cpm	128.2	cpm	122.7	cpm	115.1	cpm
too shallow	6.9	%	13.3	%	10.3	%	46.7	%	0	%	0	%
too fast	30.7	%	50	%	73.3	%	90.5	%	67.8	%	2.2	%
too slow	0	%	0	%	0	%	0	%	0	%	1.2	%
ideal	64.7	%	42.2	%	22.1	%	3.4	%	32.2	%	96.6	%
N	30.7		30.3		30.7		30.3		29		29	
depth	969.3		954.4		972.1		970.3		977		971.5	
speed	121.2	cpm	107.5	cpm	103.3	cpm	126.1	cpm	117.4	cpm	107.7	cpm
too shallow	0	%	29.3	%	0	%	0	%	0	%	0	%
too fast	14.4	%	0	%	0	%	64.3	%	8	%	1.1	%
too slow	1.1	%	2.2	%	5.4	%	1	%	0	%	1.2	%
ideal	83.5	%	68.5	%	94.6	%	34.6	%	92	%	97.7	%
N	29		29		27.7		29.3		28.7		28	
depth	938.9		979.2		950.8		959.5		964.4		946	
speed	114.5	cpm	125.8	cpm	110.6	cpm	108	cpm	115.6	cpm	123.3	cpm
too shallow	41.4	%	0	%	13.2	%	29.5	%	1.1	%	69.4	%
too fast	3.4	%	46.8	%	1.2	%	8.9	%	4.7	%	67.2	%
too slow	2.3	%	0	%	1.2	%	2.2	%	0	%	1.3	%
ideal	52.9	%	53.2	%	84.4	%	60.5	%	94.2	%	7	%
N	29		29.3		55.3		28		28.3		30.7	
depth	711.4		950.3		935.6		912.8		936.9		967.1	
speed	129	cpm	105.5	cpm	106.4	cpm	112.5	cpm	103	cpm	101.7	cpm
too shallow	100	%	1.1	%	17.2	%	66.7	%	42	%	8.6	%
too fast	57.5	%	1.1	%	0	%	2.4	%	0	%	2.2	%
too slow	0	%	19.4	%	4.6	%	2.4	%	12.8	%	18.6	%
ideal	0	%	79.5	%	79.3	%	32.1	%	55.6	%	73.9	%
N	29		29.7		29.3		28.7		28.3		28.3	
depth	1001.1		998.7		984.1		979.9		987.8		989	
speed	89.3	cpm	124.2	cpm	121.7	cpm	114.3	cpm	117	cpm	113.7	cpm
too shallow	46	%	14.1	%	0	%	0	%	2.3	%	1.1	%
too fast	0	%	15.9	%	23.7	%	23.4	%	24.7	%	8.5	%
too slow	85.1	%	0	%	0	%	3.4	%	0	%	0	%
ideal	13.8	%	71.1	%	76.3	%	73.2	%	73	%	90.4	%
N							29		27.7		29.3	
depth							985.4		989.9		986.8	
speed							125	cpm	125.4	cpm	126.5	cpm
too shallow							4.6	%	1.1	%	0	%
too fast							54	%	49.7	%	52	%
too slow							0	%	0	%	0	%
ideal							42.5	%	49.2	%	48	%
N							29		28.7		29.7	
depth							963.2		962.6		950.8	
speed							102	cpm	106.6	cpm	106.5	cpm
too shallow							2.3	%	1.2	%	31.1	%
too fast							0	%	1.2	%	1.1	%
too slow							23	%	8.1	%	7.8	%
ideal							74.7	%	89.5	%	65.6	%
N							28.3		29.7		56.3	
depth							954		993.9		991.4	
speed							100.3	cpm	130.2	cpm	112.6	cpm
too shallow							58	%	0	%	8.9	%
too fast							1.1	%	75	%	3.8	%
too slow							44.8	%	0	%	6	%
ideal							34.8	%	25	%	90	%

Table 4: Results of nurse students in the two groups teaching first and training first

	teaching first						training first					
	start		after teaching		after training		start		after training		after teaching	
	average		average		average		average		average		average	
N	30.3		27.3		29		29		29		29	
depth	813		973.4		992.3		979.8		984.4		971.6	
speed	121.1	cpm	98.9	cpm	104.4	cpm	117	cpm	113	cpm	113.8	cpm
too shallow	67	%	0	%	0	%	4.6	%	23	%	0	%
too fast	16.3	%	0	%	0	%	12.6	%	0	%	0	%
too slow	12.2	%	40.1	%	4.6	%	0	%	0	%	1.1	%
ideal	28.8	%	58.7	%	95.4	%	82.8	%	100	%	98.9	%

N	27		28		28.7		30		27.7		30	
depth	301.6		945.4		948.5		990.7		992.7		990.8	
speed	127.1	cpm	80.2	cpm	100.6	cpm	124	cpm	110.2	cpm	105.9	cpm
too shallow	100	%	50	%	12.9	%	27	%	0	%	49.8	%
too fast	74.1	%	0	%	0	%	43.4	%	0	%	0	%
too slow	5.6	%	98.8	%	20.7	%	0	%	1.2	%	1.1	%
ideal	0	%	0	%	66.4	%	44.8	%	98.8	%	50.2	%
N	28.3		28.7		28.7		28		29.3		29	
depth	837.9		971.3		977.4		553.7		408.8		907.6	
speed	120.1	cpm	124.8	cpm	131.9	cpm	127.1	cpm	171	cpm	112.2	cpm
too shallow	98.9	%	1.1	%	0	%	100	%	100	%	98.8	%
too fast	28.7	%	72.1	%	96.6	%	12.2	%	31.8	%	10.8	%
too slow	2.5	%	0	%	0	%	82.9	%	2.5	%	0	%
ideal	1.1	%	27.9	%	3.4	%	0	%	0	%	1.2	%
N	26.3		29		29.3		28.7		28.7		28.7	
depth	871.5		926.8		932.1		926.3		913.7		971.3	
speed	94.5	cpm	108.8	cpm	108.2	cpm	125.3	cpm	107.2	cpm	124.8	cpm
too shallow	100	%	74.7	%	63.3	%	66.7	%	83.4	%	0	%
too fast	0	%	0	%	0	%	46.8	%	0	%	51.3	%
too slow	76.3	%	0	%	0	%	0	%	1.1	%	0	%
ideal	0	%	25.3	%	36.7	%	8.3	%	16.6	%	48.7	%
N	28.7		28.7		27		25		28.3		25.3	
depth	858.7		925.6		933.7		916.3		919.8		925.8	
speed	76.9	cpm	106.1	cpm	105.6	cpm	108.7	cpm	108.9	cpm	118.1	cpm
too shallow	100	%	68.8	%	33.3	%	97.6	%	95.3	%	95.2	%
too fast	0	%	0	%	0	%	11.1	%	0	%	2.3	%
too slow	100	%	2.3	%	3.8	%	10.2	%	0	%	1.2	%
ideal	0	%	31.2	%	64.2	%	2.4	%	4.7	%	4.8	%
N	26.7		29		28		27.7		30.3		29.7	
depth	891.6		908.8		931.4		527.5		883.3		918.5	
speed	103.8	cpm	111.2	cpm	109.8	cpm	47.1	cpm	107.3	cpm	102.7	cpm
too shallow	64.6	%	98.9	%	59.5	%	100	%	71	%	24.3	%
too fast	2.5	%	1.1	%	0	%	0	%	2.2	%	0	%
too slow	13.9	%	2.3	%	0	%	100	%	2.2	%	11.7	%
ideal	29.1	%	1.1	%	40.5	%	0	%	29	%	72.6	%
N	29		29		30		30		30.3		25	
depth	853		961		968.4		861.6		940		988.3	
speed	117.4	cpm	115.9	cpm	101.8	cpm	84.7	cpm	93.8	cpm	94.6	cpm
too shallow	77.6	%	1.1	%	0	%	77.3	%	0	%	0	%
too fast	25.3	%	18.1	%	0	%	1.1	%	0	%	0	%
too slow	5.6	%	0	%	4.5	%	87.7	%	47.6	%	30	%
ideal	9	%	80.7	%	95.5	%	0	%	52.4	%	64.9	%
N	22.3		32.3		31		25.7		26.3		27.7	
depth	172.4		940.6		997.3		980.1		991		941.2	
speed	108	cpm	118.3	cpm	104.8	cpm	94.9	cpm	104.3	cpm	92.2	cpm
too shallow	100	%	5.4	%	0	%	44.9	%	0	%	14.9	%
too fast	35.4	%	34	%	2.3	%	0	%	0	%	0	%
too slow	63.6	%	6.1	%	12.5	%	55.6	%	5.1	%	71.2	%
ideal	0	%	58.8	%	85.2	%	35.7	%	94.9	%	25.3	%
N	27.3		30		32.7		31.3		29.3		29	
depth	555.3		895.3		950.5		617.7		945		957.1	
speed	151	cpm	126.7	cpm	113.6	cpm	134.3	cpm	123	cpm	132.6	cpm
too shallow	100	%	64.4	%	3.3	%	100	%	11.1	%	0	%
too fast	95	%	76.9	%	26.6	%	88.4	%	37.8	%	74.7	%
too slow	5	%	1	%	4.6	%	0	%	0	%	0	%
ideal	0	%	4.6	%	68.8	%	0	%	61	%	25.3	%
N	29.3		31.7		27.3		23		26.7		30	
depth	972.4		964.6		986.8		335.3		542.3		814.7	
speed	102.3	cpm	128.7	cpm	105.4	cpm	52.1	cpm	80.2	cpm	94.5	cpm
too shallow	0	%	2.1	%	0	%	100	%	100	%	100	%
too fast	1.1	%	69.1	%	0	%	0	%	0	%	0	%
too slow	34.5	%	0	%	8.6	%	100	%	98.9	%	55.8	%
ideal	64.4	%	29.8	%	91.4	%	0	%	0	%	0	%
N	28.3		28		27.3		29		28.3		29.3	
depth	906.4		963.7		958.6		853.1		778.4		927	
speed	107.6	cpm	121.1	cpm	123.4	cpm	83.3	cpm	70.2	cpm	105.1	cpm
too shallow	70.2	%	0	%	0	%	100	%	100	%	41.5	%
too fast	0	%	22.1	%	31.7	%	0	%	2.5	%	0	%
too slow	7	%	0	%	0	%	96.6	%	98.8	%	3.4	%
ideal	29.8	%	77.9	%	68.3	%	0	%	0	%	56.2	%
N	26.7		29		28.7		27		29		28.7	
depth	979		943.1		938.5		510.4		801.6		663.7	
speed	123.3	cpm	115.4	cpm	108.9	cpm	87.7	cpm	80.1	cpm	78.2	cpm
too shallow	0	%	38.1	%	23.9	%	100	%	100	%	100	%
too fast	37.7	%	5.5	%	0	%	0	%	0	%	0	%
too slow	1.3	%	2.3	%	6	%	96.3	%	100	%	100	%
ideal	61.1	%	55.3	%	72.6	%	0	%	0	%	0	%
N	29		28		28.3							
depth	962.4		983.1		967.8							
speed	93.7	cpm	94.8	cpm	118.6	cpm						
too shallow	13.8	%	0	%	0	%						
too fast	0	%	0	%	1.2	%						
too slow	58.6	%	51.4	%	0	%						
ideal	29.9	%	48.6	%	98.8	%						
N	27		28.7		28.3							
depth	788.7		948.8		963.3							
speed	82.8	cpm	98.2	cpm	95.3	cpm						
too shallow	89.7	%	12.8	%	3.6	%						
too fast	1.4	%	0	%	0	%						
too slow	96	%	32.7	%	41.3	%						
ideal	0	%	57.1	%	57.6	%						

Table 5: Results of novices in the two groups teaching first and training first

Appendix to Chapter VI - Collaboration:

Nurse Emergency 5l-HGSTP Model

The following pages show the entire 5l-HGSTP Model of the nurse A2E algorithm and parts of the Video-Wall model. Due to readability the model is divided into parts. Note that the different parts are connected via the blue WCL nodes as in Figure §1.

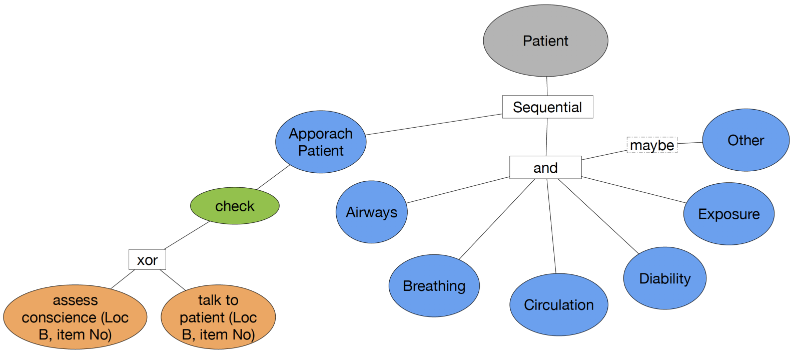


Figure 1: WCL-level of A2E algorithms

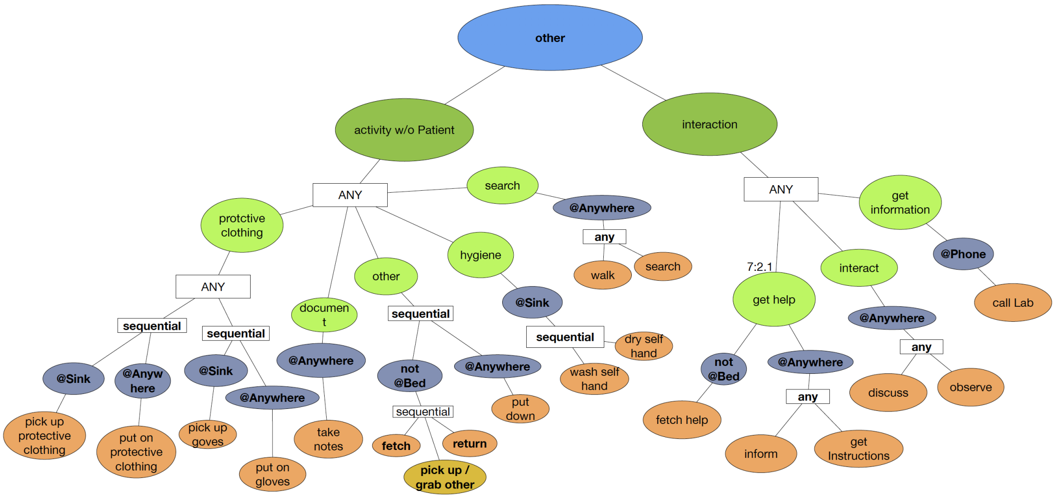


Figure 2: other modeled in the 5l HGSTP

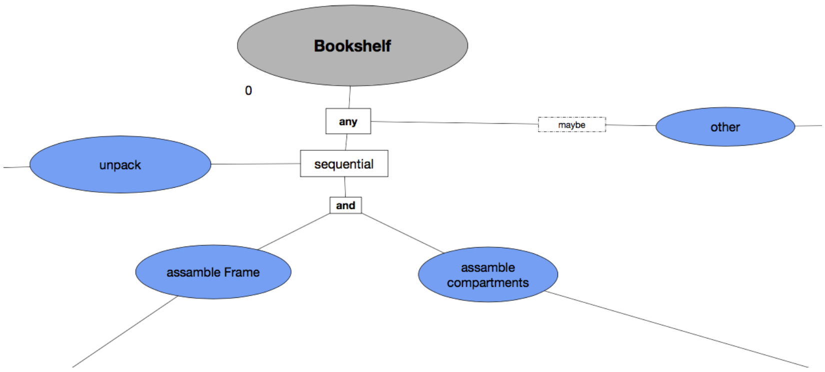


Figure 3: WCL of the furniture model

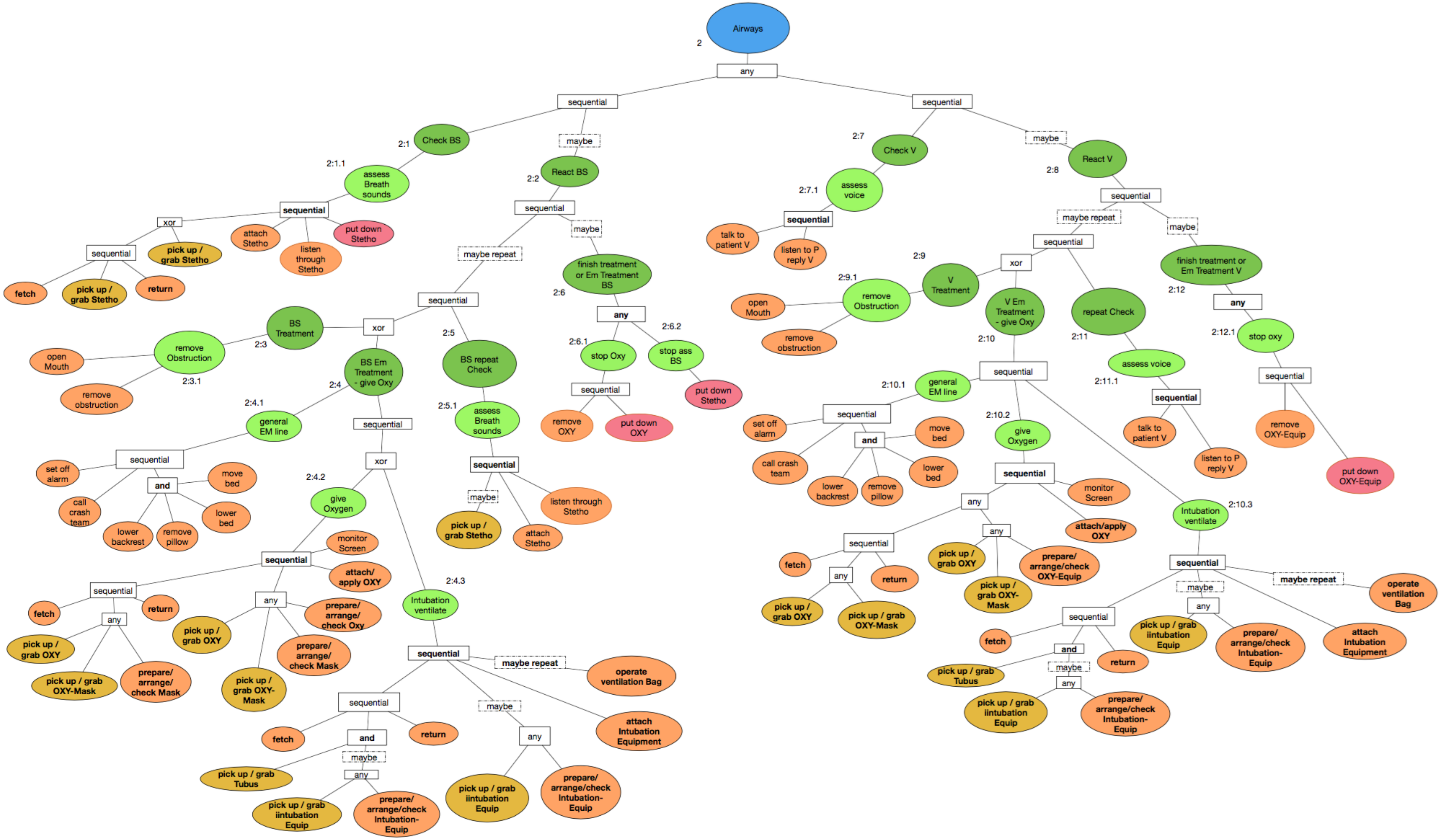


Figure 4: Airways modeled in the 5l HGSTP

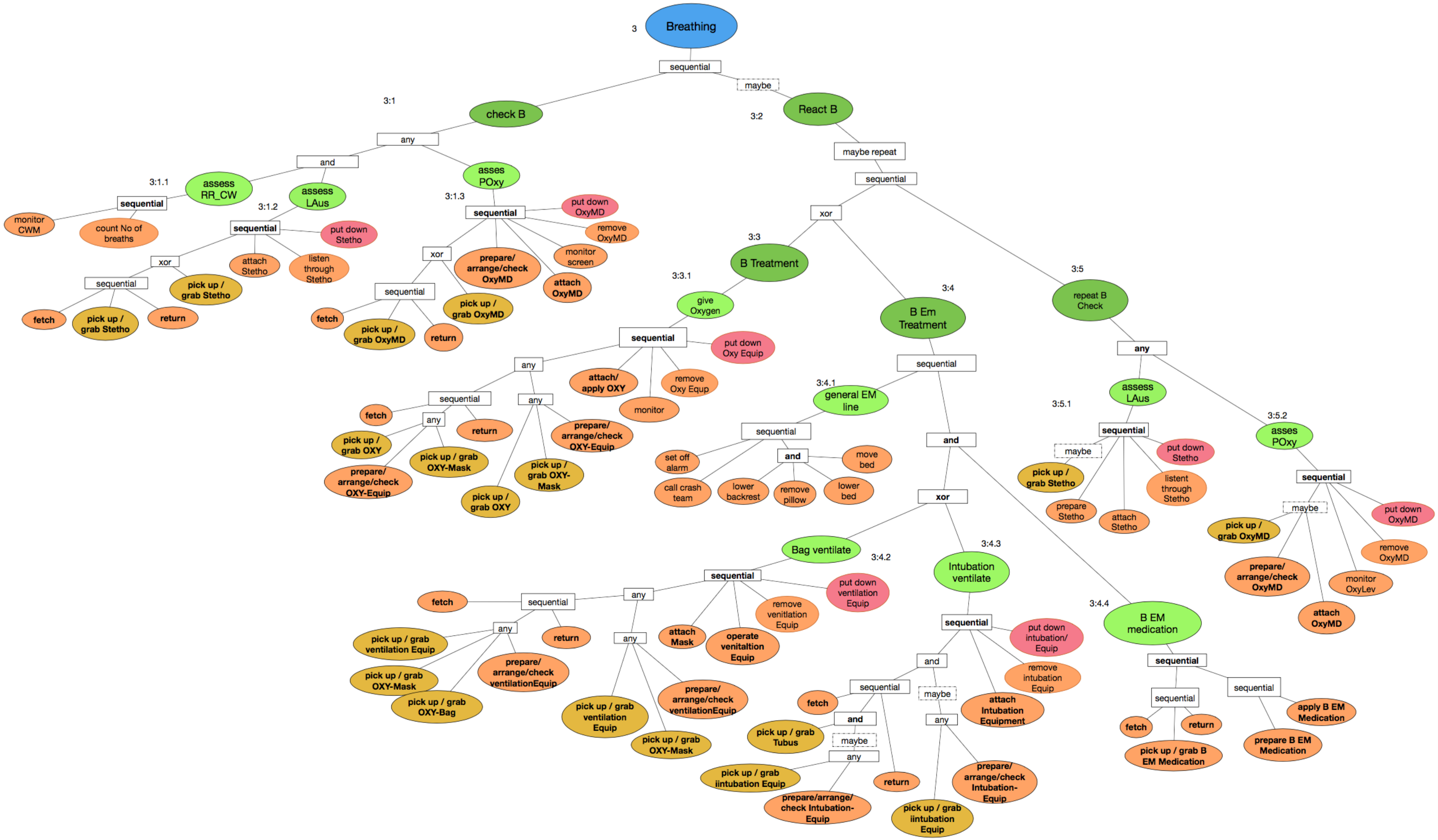


Figure 5: Breathing modeled in the 5l HGSTP

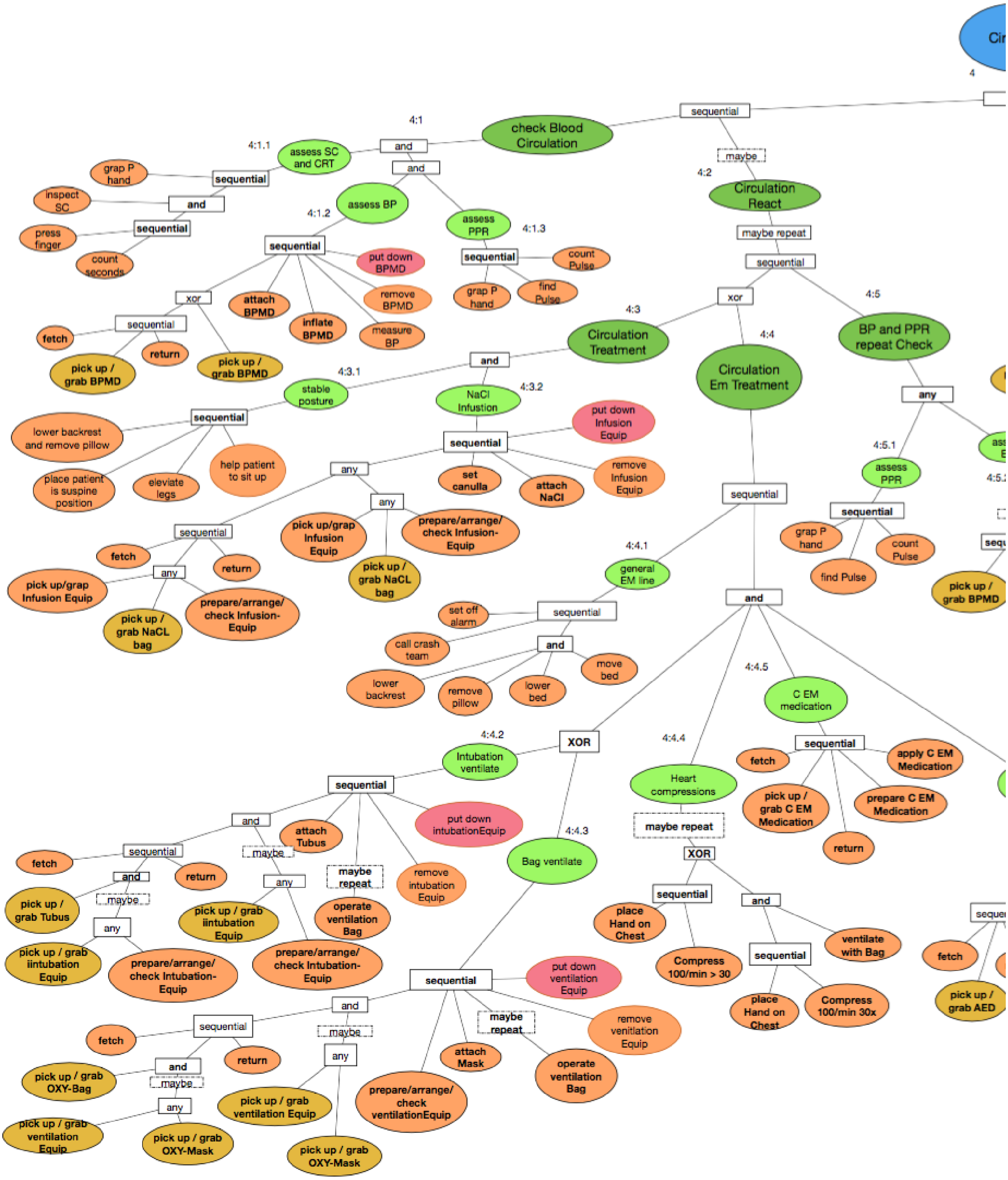


Figure 6: Circulation modeled in the 5l HGSTP part 1

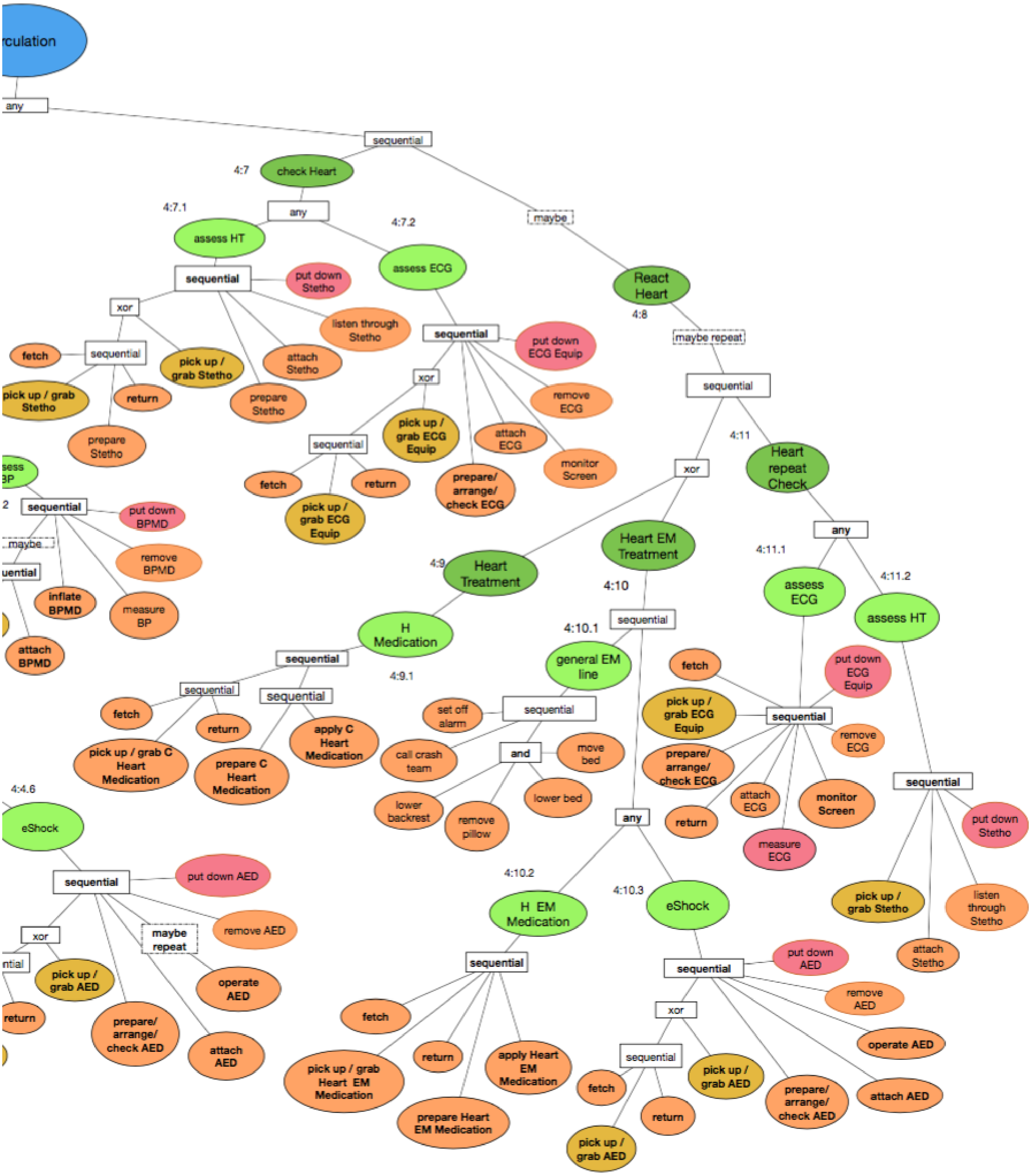


Figure 7: Circulation modeled in the 5l HGSTP part 2

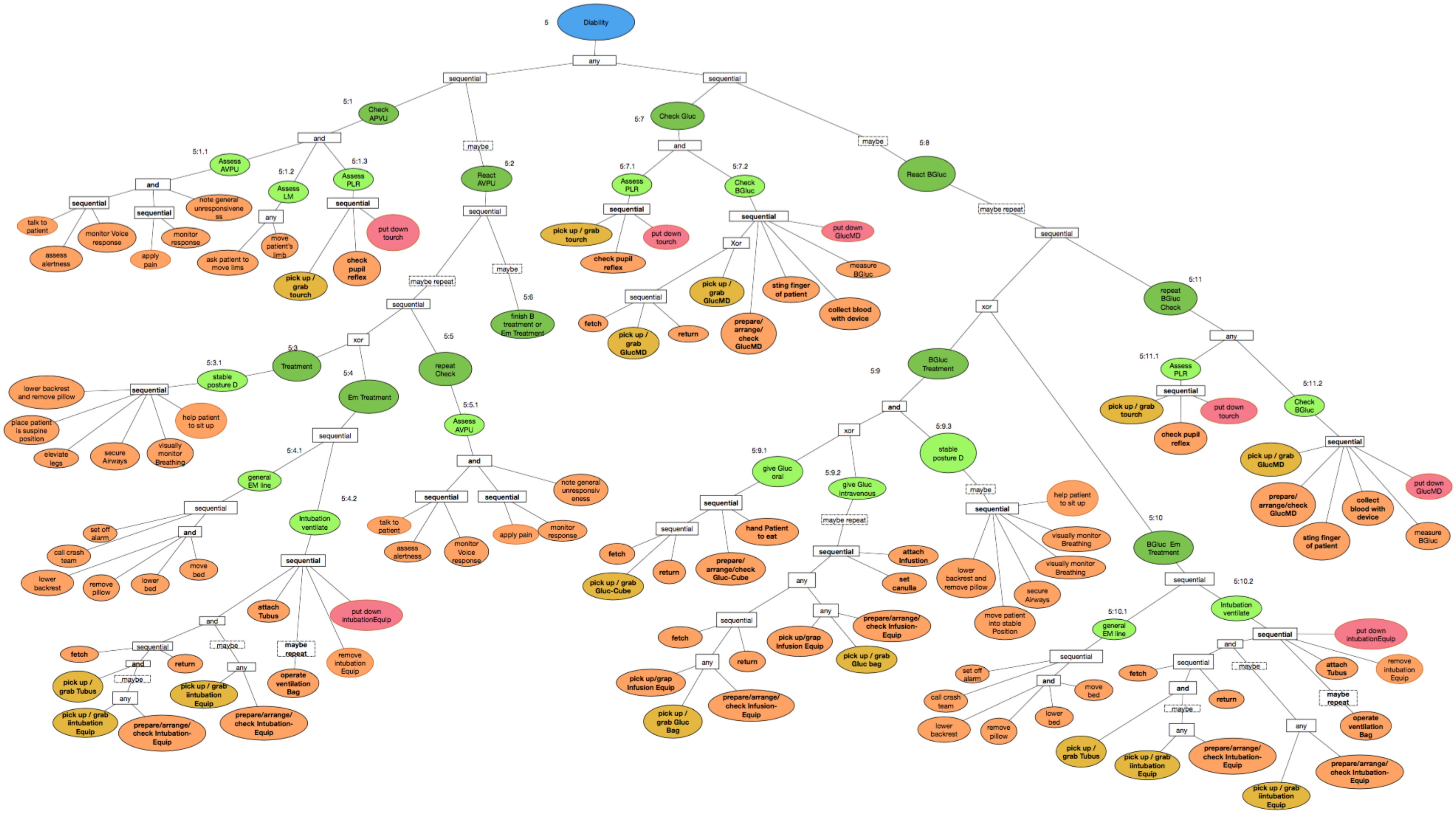


Figure 8: Disability modeled in the 5l HGSTP

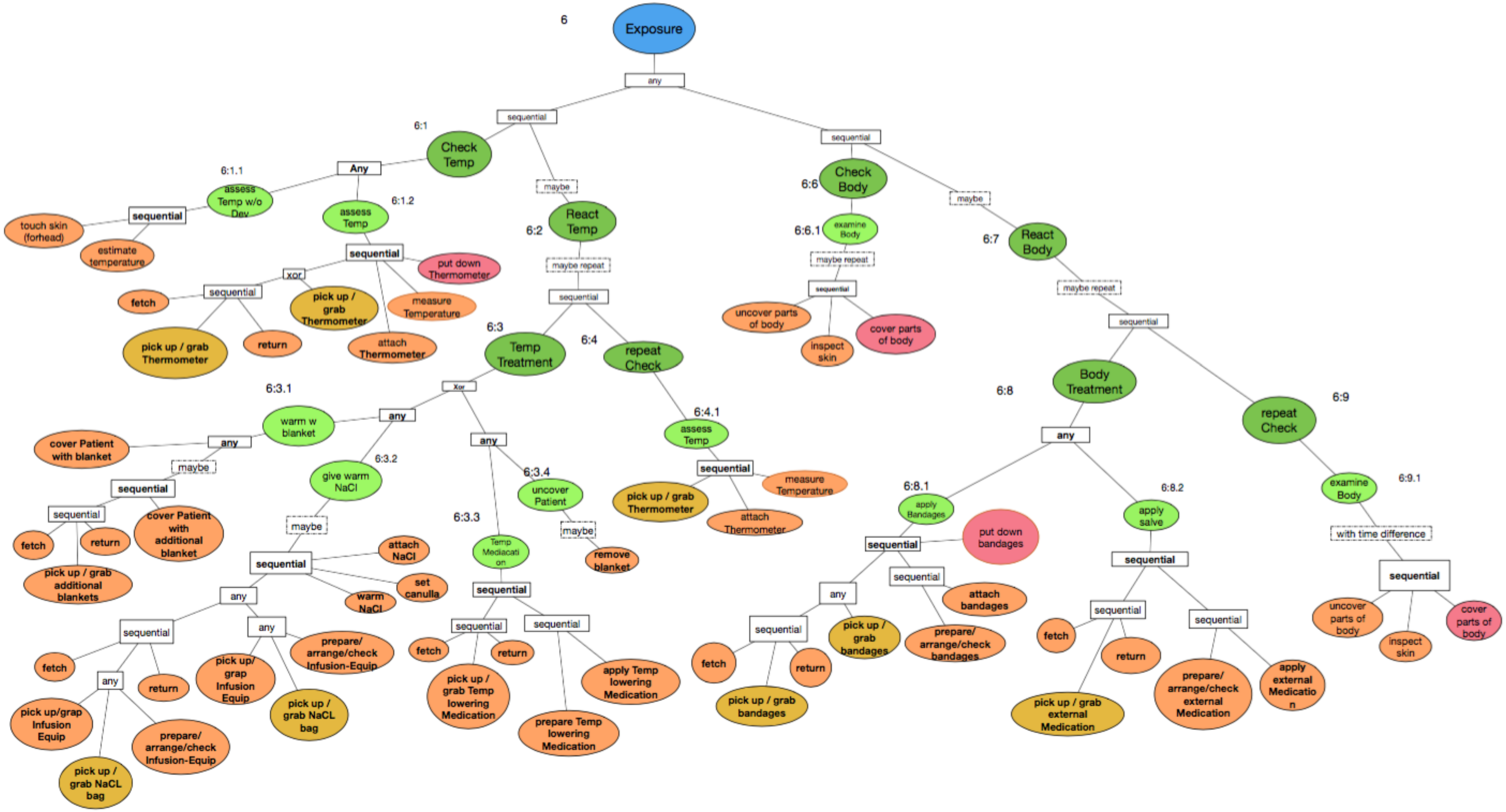


Figure 9: Exposure modeled in the 5l HGSTP

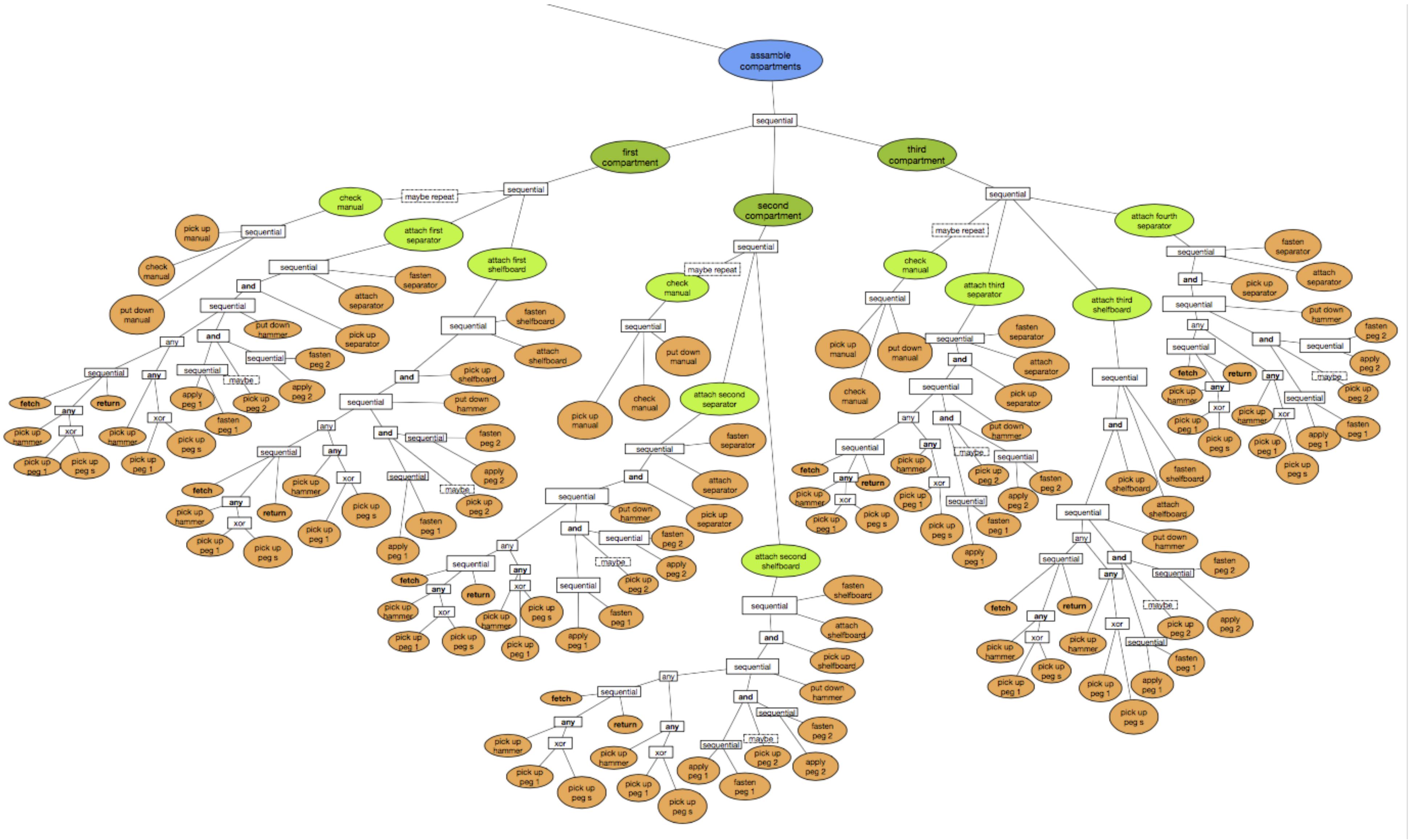


Figure 10: Exploratory for the furniture model: assemble compartments

List of Figures

List of Tables

Bibliography

[1] H. A. Whiteford, L. Degenhardt, J. Rehm, A. J. Baxter, A. J. Ferrari, H. E. Erskine, F. J. Charlson, R. E. Norman, A. D. Flaxman, N. Johns *et al.*, "Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010," *The Lancet*, vol. 382, no. 9904, pp. 1575–1586, 2013. (pages 1 and 31).

[2] WHO, "Depression," *http://www.who.int/mediacentre/factsheets/fs369/en*, 2018. (page 1).

[3] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni *et al.*, "Cost of disorders of the brain in europe 2010," *European neuropsychopharmacology*, vol. 21, no. 10, pp. 718–779, 2011. (page 1).

[4] J. Olesen, A. Gustavsson, M. Svensson, H.-U. Wittchen, B. Jönsson, C. S. Group, and E. B. Council, "The economic cost of brain disorders in europe," *European journal of neurology*, vol. 19, no. 1, pp. 155–162, 2012. (page 1).

[5] A. Grünerbl, G. Pirkl, M. Weal, M. Gobbi, and P. Lukowicz, "Monitoring and enhancing nurse emergency training with wearable devices," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 1261–1267. (page 1).

[6] G. Gargiulo, P. Bifulco, R. A. Calvo, M. Cesarelli, C. Jin, and A. van Schaik, "A mobile eeg system with dry electrodes," in *Biomedical Circuits and Systems Conference, 2008. BioCAS 2008. IEEE*. IEEE, 2008, pp. 273–276. (page 2).

[7] S. Debener, F. Minow, R. Emkes, K. Gandras, and M. De Vos, "How about taking a low-cost, small, and wireless eeg for a walk?" *Psychophysiology*, vol. 49, no. 11, pp. 1617–1621, 2012. (page 2).

[8] N. Kosmyna, U. Sarawgi, and P. Maes, "Attentivu: Evaluating the feasibility of biofeedback glasses to monitor and improve attention," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 2018, pp. 999–1005. (page 2).

[9] J. Hernandez, D. McDuff, K. S. Quigley, P. Maes, and R. W. Picard, "Wearable motion-based heart-rate at rest: A workplace evaluation," *IEEE journal of biomedical and health informatics*, 2018. (page 2).

[10] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition," in *Advances in Neural Information Processing Systems*, 2006, pp. 787–794. (page 3).

[11] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013. (pages 3 and 87).

[12] J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 747–752. (page 3).

[13] T. Choudhury, S. Consolvo, B. Harrison, J. Hightower, A. LaMarca, L. LeGrand, A. Rahimi, A. Rea, G. Bordello, B. Hemingway *et al.*, "The mobile sensing platform: An embedded activity recognition system," *IEEE Pervasive Computing*, vol. 7, no. 2, 2008. (page 3).

[14] G. Bieber, J. Voskamp, and B. Urban, "Activity recognition for everyday life on mobile phones," in *Proceedings of the 5th International on ConferenceUniversal Access in Human-Computer Interaction*. Springer-Verlag, 2009, pp. 289–296. (page 3).

[15] M. Muehlbauer, G. Bahle, and P. Lukowicz, "What can an arm holster worn smart phone do for activity recognition?" in *ISWC*, 2011, pp. 79–82. (page 3).

[16] C. Seeger, A. Buchmann, and K. Van Laerhoven, "myhealthassistant: A phone-based body sensor network that captures the wearer's exercises throughout the day," in *The 6th International Conference on Body Area Networks*. Beijing, China: ACM Press, 11 2011. (page 3).

[17] G. Ogris, T. Stiefmeier, P. Lukowicz, and G. Tröster, "Using a complex multi-modal on-body sensor system for activity spotting." in *ISWC*. IEEE, 2008, pp. 55–62. (page 3).

[18] A. Zinnen, C. Wojek, and B. Schiele, "Multi activity recognition based on bodymodel-derived primitives." in *LoCA*, ser. Lecture Notes in Computer Science, vol. 5561. Springer, 2009, pp. 1–18. (page 3).

[19] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 20, no. 12, pp. 1371–1375, 1998. (page 3).

[20] D. H. Hu, S. J. Pan, V. W. Zheng, N. N. Liu, and Q. Y. 0001, "Real world activity recognition with multiple goals." in *UbiComp*, ser. ACM International Conference Proceeding Series, vol. 344. ACM, 2008, pp. 30–39. (page 3).

[21] C. W. Geib and R. P. Goldman, "A probabilistic plan recognition algorithm based on plan tree grammars," *Artificial Intelligence*, vol. 173, no. 11, pp. 1101–1132, 2009. (page 3).

[22] U. Blanke and B. Schiele, "Remember and transfer what you have learned-recognizing composite activities based on activity spotting," in *Wearable Computers (ISWC), 2010 Int. Symp. on*. IEEE, 2010, pp. 1–8. (pages 3 and 113).

[23] P. Lukowicz, "Wearable computing and artificial intelligence for healthcare applications." *Artificial Intelligence in Medicine*, vol. 42, no. 2, pp. 95–98, 2008. (pages 3, 13, and 33).

[24] C. Orwat, A. Graefe, and T. Faulwasser, “Towards pervasive computing in health care - a literature review,” *BMC Medical Informatics and Decision Making*, vol. 8, no. 26, 2008. (pages 3, 13, and 33).

[25] S. Agarwal, A. Joshi, T. Finin, Y. Yesha, and T. Ganous, “A pervasive computing system for the operating room of the future,” *Mob. Netw. Appl.*, vol. 12, no. 2-3, pp. 215–228, Mar. 2007. (page 4).

[26] M. Cheng, M. Kanai–Pak, N. Kuwahara, H. Ozaku, K. Kogure, and J. Ota, “Dynamic scheduling–based inpatient nursing support: applicability evaluation by laboratory experiments,” *Journal International Journal of Autonomous and Adaptive Communications Systems*, vol. 5, no. 1, pp. 39 – 57, 2012. (page 4).

[27] F. Naya, R. Ohmura, M. Miyamae, H. Noma, K. Kogure, and M. Imai, “Wireless sensor network system for supporting nursing context-awareness,” *Journal International Journal of Autonomous and Adaptive Communications Systems*, vol. 4, no. 4, pp. 361 – 382, 2011. (page 4).

[28] M. E. Pollack, “Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment,” *AI magazine*, vol. 26, no. 2, p. 9, 2005. (page 4).

[29] T. L. Hayes, F. Abendroth, A. Adami, M. Pavel, T. A. Zitzelberger, and J. A. Kaye, “Unobtrusive assessment of activity patterns associated with mild cognitive impairment,” *Alzheimer’s & Dementia*, vol. 4, no. 6, pp. 395–405, 2008. (page 4).

[30] J. A. Kaye, S. A. Maxwell, N. Mattek, T. L. Hayes, H. Dodge, M. Pavel, H. B. Jimison, K. Wild, L. Boise, and T. A. Zitzelberger, “Intelligent systems for assessing aging changes: home-based, unobtrusive, and continuous assessment of aging,” *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 66, no. suppl_1, pp. i180–i190, 2011. (page 4).

[31] P. N. Dawadi, D. J. Cook, M. Schmitter-Edgecombe, and C. Parsey, “Automated assessment of cognitive health using smart home technologies,” *Technology and health care*, vol. 21, no. 4, pp. 323–343, 2013. (page 4).

[32] R. Brunken, J. L. Plass, and D. Leutner, “Direct measurement of cognitive load in multimedia learning,” *Educational psychologist*, vol. 38, no. 1, pp. 53–61, 2003. (page 4).

[33] S. P. Marshall, “The index of cognitive activity: Measuring cognitive workload,” in *Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on*. IEEE, 2002, pp. 7–7. (page 4).

[34] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252. (page 4).

[35] B. Kort, R. Reilly, and R. W. Picard, “An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion,” in *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*. IEEE, 2001, pp. 43–46. (page 4).

[36] R. W. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence: Analysis of affective physiological state,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001. (page 4).

[37] J. A. Healey and R. W. Picard, “Detecting stress during real-world driving tasks using physiological sensors,” *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005. (page 4).

[38] R. Matthews, N. J. McDonald, P. Hervieux, P. J. Turner, and M. A. Steindorf, “A wearable physiological sensor suite for unobtrusive monitoring of physiological and cognitive state,” in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 2007, pp. 5276–5281. (page 4).

[39] M. B. Russo, M. C. Stetz, and M. L. Thomas, “Monitoring and predicting cognitive state and performance via physiological correlates of neuronal signals,” *Aviation, space, and environmental medicine*, vol. 76, no. 7, pp. C59–C63, 2005. (page 4).

[40] 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*, vol. 14, no. 2, pp. 410–417, 2010. (page 4).

[41] M. Barry, J. Gutknecht, I. Kulka, P. Lukowicz, and T. Stricker, *Multimedial enhancement of a butoh dance performance-mapping motion to emotion with a wearable computer system*. na, 2004. (page 4).

[42] ——, “From motion to emotion: a wearable system for the multimedia enrichment of a butoh dace performance,” *Journal of Mobile Multimedia*, pp. 112–132, 2005. (page 4).

[43] E. Crane and M. Gross, “Motion capture and emotion: Affect detection in whole body movement,” *Affective computing and intelligent interaction*, pp. 95–101, 2007. (page 4).

[44] R. McDonnell, S. Jörg, J. McHugh, F. Newell, and C. O’Sullivan, “Evaluating the emotional content of human motions on real and virtual characters,” in *Proceedings of the 5th symposium on Applied perception in graphics and visualization*. ACM, 2008, pp. 67–74. (page 4).

[45] J. Amores and P. Maes, “Essence: Olfactory interfaces for unconscious influence of mood and cognitive performance,” in *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 2017, pp. 28–34. (page 4).

[46] K. Kunze, M. Iwamura, K. Kise, S. Uchida, and S. Omachi, “Activity recognition for the mind: Toward a cognitive" quantified self",” *Computer*, vol. 46, no. 10, pp. 105–108, 2013. (page 5).

[47] A. Samal and P. A. Iyengar, “Automatic recognition and analysis of human faces and facial expressions: A survey,” *Pattern recognition*, vol. 25, no. 1, pp. 65–77, 1992. (page 5).

[48] M. Pantic and L. J. M. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000. (page 5).

[49] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211. (page 5).

[50] K. Masai, K. Kunze, Y. Sugiura, M. Ogata, M. Inami, and M. Sugimoto, “Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 4, p. 15, 2017. (page 5).

[51] K. Masai, K. Kunze, Y. Sugiura, and M. Sugimoto, “Mapping natural facial expressions using unsupervised learning and optical sensors on smart eyewear,” in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 2018, pp. 158–161. (page 5).

[52] A. Gruenerbl, G. Bahle, F. Hanser, and P. Lukowicz, “Using indoor location to assess the state of dementia patients: Results and experience report from a long term, real world study,” in *7th International Conference on Intelligent Environments (IE 2011)*. Notthingham, UK: IEEE Xplore, 2011. (pages 12, 13, and 24).

[53] ——, “Uwb indoor location for monitoring dementia patients: The challenges and perception of a real-life deployment,” *International Journal of Ambient Computing and Intelligence*, no. 4, 2013. (pages 12, 13, 15, 17, and 28).

[54] A. Gruenerbl, G. Bahle, J. Weppner, and P. Lukowicz, “Ubiquitous context aware monitoring systems in psychiatric and mental care: challenges and issues of real life deployments,” in *Proceedings of the 3rd International Conference on Context-Aware Systems and Applications*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, pp. 105–109. (pages 12, 17, 32, 33, and 52).

[55] F. H. and S. H. D., *Demenz. Grundlagen, Diagnostik, Formen (Dementia. Fundamentals, Diagnostics, Shapes)*. Schriftenreihe der Bayerischen Landesapothekerkammer (Publication of Bavarian Regional State Pharmacy), H. 74. Eschborn: GOVI Pharmazeutischer Verlag, 2003. (page 13).

[56] U. Kastner and R. Löbach, *Handbuch Demenz (Handbook of Dementia)*. Urban & Fischer, München Jena, 2007. (page 13).

[57] S. G Schröder, *Psychopathologie der Demenz (Psychopathology of Dementia)*. Schattauer, Stuttgart, 2006. (page 13).

[58] M. D. Hurd, P. Martorell, A. Delavande, K. J. Mullen, and K. M. Langa, “Monetary costs of dementia in the united states,” *N Engl J Med*, vol. 2013, no. 368, pp. 1326–1334, 2013. (page 13).

[59] E. Tapia, S. Intille, and K. Larson, “Activity recognition in the home using simple and ubiquitous sensors,” *Pervasive Computing*, pp. 158–175, 2004. (pages 13 and 113).

[60] J. Chen, A. Kam, J. Zhang, N. Liu, and L. Shue, “Bathroom activity monitoring based on sound,” *Pervasive Computing*, pp. 47–61, 2005. (page 13).

[61] D. Wilson, S. Consolvo, K. Fishkin, and M. Philipose, “In-home assessment of the activities of daily living of the elderly,” in *Extended Abstracts of CHI 2005: Workshops-HCI Challenges in Health Assessment*, 2005. (page 13).

[62] T. Huynh, U. Blanke, and B. Schiele, “Scalable recognition of daily activities with wearable sensors,” in *Proceedings of the 3rd international conference on Location-and context-awareness*. Springer-Verlag, 2007, pp. 50–67. (page 13).

[63] J. Lester, T. Choudhury, and G. Borriello, “A practical approach to recognizing physical activities,” *Pervasive Computing*, pp. 1–16, 2006. (page 13).

[64] L. M. Ni, Y. Liul, Y. C. Lau, and A. P. Patil, “LANDMARC: Indoor location sensing using active RFID,” *Wireless Networks*, vol. 10, no. 6, pp. 701–710, 2004. (page 13).

[65] Ubisense. (2017, May) Ubisense, precise real-time location systems. [Online]. Available: http://www.ubisense.net/en/ (pages 13 and 16).

[66] X.-F. Teng, Y.-T. Zhang, C. C. Y. Poon, and P. Bonato, “Wearable medical systems for p-health,” pp. 62–74, 2008. (pages 13 and 33).

[67] P. Bonato, “Clinical applications of wearable technology.” *Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, pp. 6580–6583, 2009. (pages 13 and 33).

[68] M. E. Pollack, “Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment,” *AI Magazine*, vol. 26, no. 2, pp. 9–24, 2005. (page 13).

[69] J. Yang, W. Mann, S. Nochajski, and M. Tomita, “Use of assistive devices among elders with cognitive impairment: a follow-up study.” *Topics in Geriatric Rehabilitation*, vol. 13, no. 2, pp. 13–31, 1997. (page 13).

[70] R. Megret, D. Szolgay, J. Benois-Pineau, P. Joly, J. Pinquier, J. F. Dartigues, and C. Helmer, “Wearable video monitoring of people with age dementia: Video indexing at the service of healthcare,” in *CBMI, International Workshop on Content-Based Multimedia Indexing*, June 2008, pp. 101–108. (page 13).

[71] A. König, C. F. Crispim Junior, A. Derreumaux, G. Bensadoun, P.-D. Petit, F. Bremond, R. David, F. Verhey, P. Aalten, and P. Robert, “Validation of an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients,” *Journal of Alzheimer’s Disease*, vol. 44, no. 2, pp. 675–685, 2015. (page 13).

[72] L. Rebenitsch, C. B. Owen, R. Ferrydiansyah, C. Bohil, and F. Biocca, “An exploration of real-time environmental interventions for care of dementia patients in assistive living,” in *PETRA ’10: Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*. New York, NY, USA: ACM, 2010, pp. 1–8. (page 13).

[73] S. J. Allin, A. Bharucha, J. Zimmerman, D. Wilson, M. J. Roberson, S. Stevens, H. Wactlar, and C. Atkeson, “Toward the automatic assessment of behavioral disturbances of dementia,” in *In UbiHealth 2003: The 2nd International Workshop on Ubiquitous*, 2003, pp. 12–15. (page 13).

[74] D. Chen, R. Malkin, and J. Yang, “Multimodal detection of human interaction events in a nursing home environment,” in *Sixth International Conference on Multimodal Interfaces (ICMI’04*. ACM, 2004, pp. 14–15. (page 13).

[75] C. C. Lin, P. Lin, P. Lu, G. Hsieh, W. L. Lee, and R. Lee, “A healthcare integration system for disease assessment and safety monitoring of dementia patients,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 5, pp. 579 –586, 2008. (page 14).

[76] A. Lotfi, C. Langensiepen, S. M. Mahmoud, and M. J. Akhlaghinia, “Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour,” *Journal of ambient intelligence and humanized computing*, vol. 3, no. 3, pp. 205–218, 2012. (page 14).

[77] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, “Landmarc: indoor location sensing using active rfid,” *Wireless networks*, vol. 10, no. 6, pp. 701–710, 2004. (page 15).

[78] S.-C. Kim, Y.-S. Jeong, and S.-O. Park, “Rfid-based indoor location tracking to ensure the safety of the elderly in smart home environments,” *Personal and ubiquitous computing*, vol. 17, no. 8, pp. 1699–1707, 2013. (page 15).

[79] G. Ogris, T. Stiefmeier, H. Junker, P. Lukowicz, and G. Troster, “Using ultrasonic hand tracking to augment motion analysis based recognition of manipulative gestures,” in *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*. IEEE, 2005, pp. 152–159. (page 15).

[80] R. Megret, D. Szolgay, J. Benois-Pineau, P. Joly, J. Pinquier, J.-F. Dartigues, and C. Helmer, “Wearable video monitoring of people with age dementia: Video indexing at the service of helthcare,” in *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*. IEEE, 2008, pp. 101–108. (page 15).

[81] F. Hanser, A. Gruenerbl, C. Rodegast, and P. Lukowicz, “Design and real life deployment of a pervasive monitoring system for dementia patients,” in *Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008. Second International Conference on*. IEEE, 2008, pp. 279–280. (pages 16 and 18).

[82] E. Kasten, C. Utecht, and M. Waselewski, *Den Alltag demenzerkrankter Menschen neu gestalten (Redesign the Daily Routine of People suffering from Dementia)*. Schlütersche, Hannover, 2004. (page 16).

[83] U. Schindler, *Die Pflege demenziell Erkrankter neu erleben: Mäeutik im Praxisalltag*. Vincentz Network GmbH & Co KG, 2003. (page 16).

[84] M. Broeker. (2015, May) Infomappe mÄeutik – erlebensorientierte pflege und betreuung (maiutics, experienceoriented care). [Online]. Available: http://www.imoz.de/fileadmin/dl/downloads/Infomappe_Maeeutik.pdf (page 16).

[85] 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,” *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975. (page 16).

[86] M. Kurz, G. Hölzl, A. Ferscha, A. Calatroni, D. Roggen, G. Tröster, H. Sagha, R. Chavarriaga, J. Millán, D. Bannach, K. Kunze, and P. Lukowicz, “The opportunity framework and data processing ecosystem for opportunistic activity and context recognition,” *International Journal of Sensors, Wireless Communications and Control, Special Issue on Autonomic and Opportunistic Communications*, vol. 1, no. 2, pp. 102–125, 2011. (page 17).

[87] H. Stefan, F. Allmer, and J. Eberls, *Praxis der Pflegediagnosen (Experience in Care Diagnosis)*. Springer, Wien, 2007. (pages 19 and 20).

[88] J. Cerejeira, L. Lagarto, and E. Mukaetova-Ladinska, “Behavioral and psychological symptoms of dementia,” *Frontiers in neurology*, vol. 3, p. 73, 2012. (page 21).

[89] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” vol. 7, no. 7, pp. 179–188, 1936. (pages 24 and 47).

[90] U. of Waikato. (2017, May) Data mining with open source mashine lerning software. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/ (pages 24 and 47).

[91] A. Gruenerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, “Smartphone-based recognition of states and state changes in bipolar disorder patients,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2015. (pages 32, 33, and 112).

[92] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Currasco, S. Oehler, O. Mayora, C. Haring, and Lukowicz, “Using smartphone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients.” in *5th ACM Augmented Human International Conferenc*. ACM, 2014. (pages 32, 33, 36, and 112).

[93] A. Gruenerbl, G. Bahle, J. Weppner, P. Oleksy, C. Haring, and P. Lukowicz, “Towards smart phone based monitoring of bipolar disorder,” in *Proc. of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*. ACM, 2012, pp. 3:1–3:6. (pages 32, 33, and 36).

[94] A. Muaremi, F. Gravenhorst, A. Gruenerbl, B. Anrich, and G. Troester, “Assessing bipolar episodes using speech cues derived from phone calls.” in *4th International Symposium on Pervasive Computing Paradigms for Mental Health (MindCare)*, 2014. (page 32).

[95] N. I. of Metal Health, “Bipolar disorder,” 2009, http://www.nimh.nih.gov/health/publications/bipolar-disorder. (page 33).

[96] F. Lobban, “Enhanced relaps prevention for bipolar disorder,” 2007, bMC Psychiatry. (page 33).

[97] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. American Psychiatric Association, 2013, available: dsm.psychiatryonline.org. [June - 2013]. (page 33).

[98] J. E. Bardram, “Pervasive healthcare as a scientific discipline,” *Methods of information in medicine*, vol. 47, no. 3, pp. 178–185, 2008. (page 33).

[99] M. E. Pollack, “The use of ai to assist elders with cognitive impairment for an aging population,” *AI Magazine*, vol. 26, no. 2, pp. 9–24, 2005. (page 33).

[100] T. L. Westeyn, G. D. Abowd, T. E. Starner, J. M. Johnson, P. W. Presti, and K. A. Weaver, “Monitoring children’s developmental progress using augmented toys and activity recognition,” *Personal Ubiquitous Comput.*, vol. 16, no. 2, pp. 169–191, February 2012. (page 33).

[101] T. de Jongh, I. Gurol-Urganci, V. Vodepivec-Jamsek, J. Car, and R. Atun, “Mobile phone messaging for facilitating self-management of long-term illnesses,” *Cochrane Database of Systematic Reviews*, vol. 12, no. CD007459, 2012. (page 33).

[102] J. Bopp, D. Miklowitz, G. Goodwin, W. Stevens, J. Rendell, and G. JR., “The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study,” *Bipolar Disord*, vol. 12, no. 3, pp. 327–34, 2010. (pages 33, 57, and 59).

[103] T.-J. Yun, H. Y. Jeong, T. D. Hill, B. Lesnick, R. Brown, G. D. Abowd, and R. I. Arriaga, “Using sms to provide continuous assessment and improve health outcomes for children with asthma,” in *Proc. of the 2nd International Health Informatics Symposium*. ACM, 2012, pp. 621–630. (pages 33, 57, and 59).

[104] T. L. Simpson, D. R. Kivlahan, K. R. Bush, and M. E. McFall, “Telephone self-monitoring among alcohol use disorder patients in early recovery: a randomized study of feasibility and measurement reactivity,” *Drug and alcohol dependence*, vol. 79, no. 2, p. 241–50, 2005. (page 33).

[105] E. McAdams, A. Krupaviciute, C. Gehin, A. Dittmar, G. Delhomme, P. Rubel, J. Fayn, and J. McLaughlin, "Wearable electronic systems: Applications to medical diagnostics/ monitoring," *Wearable Monitoring Systems*, p. 179, 2010. (page 34).

[106] G. Technologies, "emoods bipolar mood tracker (version 1.0) [mobile software application]," 2013, https://play.google.com/store/apps. (page 34).

[107] K. Heron and J. Smyth, "Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments." *Br J Health Psychol*, vol. 15, pp. 1–39, 2010. (page 34).

[108] J. Treasure, C. Macare, I. Mentxaka, and A. Harrison, "The use of a vodcast to support eating and reduce anxiety in people with eating disorder: a case series." *Eur Eat Disorder Rev.*, vol. 18, pp. 512–521, 2010. (page 34).

[109] C. Depp, B. Mausbach, E. Granholm, V. Cardenas, D. Ben-Zeev, T. Patterson, B. Lebowitz, and D. Jeste, "Mobile interventions for severe mental illness: design and preliminary data from three approaches." *J Nerv Ment Dis.*, vol. 198, no. 10, pp. 712–721, 2010. (page 34).

[110] E. Granholm, D. Ben-Zeev, P. Link, K. Bradshaw, and H. JL., "Mobile assessment and treatment for schizophrenia (mats): a pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations." *Schizophrenia Bulletin.*, vol. 38, no. 3, 2012. (page 34).

[111] S. Reid, S. Kauer, S. Hearps, A. Crooke, A. Khor, L. Sanci, and G. Patton, "A mobile phone application for the assessment and management of youth mental health problems in primary care: a randomised controlled trial." *IEEE Commun Mag*, vol. 12, no. 1, 2011. (page 34).

[112] N. Lane, , E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, "A survey of mobile phone sensing." *BMC Fam Pract*, vol. 48, no. 9, pp. 140–150, 2010. (page 34).

[113] M. Matthews, S. Abdullah, G. Gay, and C. T., "Tracking mental well-being: Balancing rich sensing and patient needs," *Computer*, vol. 47, no. 4, pp. 36–43, 2014. (page 34).

[114] A. Honka, K. Kaipainen, H. Hietala, and N. Saranummi, "Rethinking health: Ict enabled services to empower people to manage their health," *IEEE Reviews in Biomedical Engineering*, vol. 4, 2011. (page 34).

[115] M. Morris and F. Guilak, "Mobile heart health: Project highlight." *IEEE Pervasive Computing*, vol. 8, no. 2, 2009. (page 34).

[116] T. Massey, G. Marfia, M. Potkonjak, and M. Sarrafzadeh, "Experimental analysis of a mobile health system for mood disorders." *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 241–247, 2010. (pages 34 and 57).

[117] R. Paradiso, A. M. Bianchi, K. Lau, and E. P. Scilingo, "Psyche: personalised monitoring systems for care in mental health." *Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, vol. 2010, no. 1983, pp. 3602–5. (page 34).

[118] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr, "Harnessing context sensing to develop a mobile intervention for depression," *Journal of Medical Internet Research*, vol. 13, no. 3, 2011. (pages 34, 57, and 59).

[119] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Can your smartphone infer your mood?" in *9th ACM Conference on Embedded Networked Sensor Systems (SenSys 2011)*. Seattle, WA, USA: ACM, 2011. (pages 34, 57, and 59).

[120] Oxtext, "Truecolours - improved management for people with biplar disorder," 2014, available: http://oxtext.psych.ox.ac.uk/. [July-2014]. (page 34).

[121] Optimism, "Optimism apps," 2013, available: www.findingoptimism.com. [Accessed: 09-Jul-2013]. (pages 34 and 57).

[122] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Supporting disease insight through data analysis: refinements of the monarca self-assessment system." in *UbiComp*. ACM, 2013, pp. 133–142. (pages 34, 58, and 59).

[123] A. Muaremi, F. Gravenhorst, A. Gruenerbl, B. Anrich, and G. Troester, "Assessing bipolar episodes using speech cues derived from phone calls." in *4th International Symposium on Pervasive Computing Paradigms for Mental Health (MindCare)*, 2014. (page 46).

[124] A. Muaremi, "Wearable sensing of mental health and human behavior," Ph.D. dissertation, 2014. (pages 46 and 52).

[125] S. Feese, A. Muaremi, B. Arnrich, G. Tröster, B. Meyer, and K. Jonas, "Discriminating individually considerate and authoritarian leaders by speech activity cues," in *Workshop on Social Behavioral Analysis and Behavioral Change (SBABC)*, 2011. (page 46).

[126] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*, 2010. (page 47).

[127] O. Mayora, M. Frost, B. Arnrich, F. Gravenhorst, A. Grunerbl, A. Muaremi, V. Osmani, A. Puiatti, N. Reichwaldt, C. Scharnweber *et al.*, "Mobile health systems for bipolar disorder: the relevance of non-functional requirements in monarca project," *International Journal of Handheld Computing Research (IJHCR)*, vol. 5, no. 1, pp. 1–12, 2014. (page 52).

[128] ——, "Mobile health systems for bipolar disorder: the relevance of non-functional requirements in monarca project," in *E-health and telemedicine: Concepts, methodologies, tools, and applications*. IGI Global, 2016, pp. 1395–1405. (page 52).

[129] A. Gruenerbl, G. Bahle, S. Oehler, R. Banzer, C. Haring, and P. Lukowicz, "Sensors vs. human: comparing sensor based state monitoring with questionnaire based self-assessment in bipolar disorder patients," in *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. ACM, 2014, pp. 133–134. (page 56).

[130] E. J. Austin, I. J. Deary, G. J. Gibson, M. J. McGregor, and J. B. Dent, "Individual response spread in self-report scales: Personality correlations and consequences," *Personality and Individual Differences*, vol. 24, no. 3, pp. 421–438, 1998. (page 57).

[131] D. L. Schacter, "The seven sins of memory: Insights from psychology and cognitive neuroscience." *American psychologist*, vol. 54, no. 3, p. 182, 1999. (page 57).

[132] R. Elgie and P. L. Morselli, "Social functioning in bipolar patients: the perception and perspective of patients, relatives and advocacy organizations–a review," *Bipolar disorders*, vol. 9, no. 1-2, pp. 144–157, 2007. (page 57).

[133] I. E. Allen and C. A. Seaman, "Likert scales and data analyses," *Quality progress*, vol. 40, no. 7, p. 64, 2007. (page 59).

[134] M. Müller, *Information retrieval for music and motion*. Springer, 2007, vol. 2. (page 59).

[135] ——, *Information retrieval for music and motion*. Springer, 2007, vol. 2. (page 62).

[136] A. Gruenerbl, G. Pirkl, E. Monger, M. Gobbi, and P. Lukowicz, "Smart-watch life saver: smart-watch interactive-feedback system for improving bystander cpr," in *Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 19–26. (pages 70, 71, and 80).

[137] A. Gruenerbl, H. Javaheri, E. Monger, M. Gobbi, and P. Lukowicz, "Training cpr with a wearable real time feedback system," in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. (pages 70 and 71).

[138] H. R. Society, "Sudden cardiac arrest (sca)," *http://www.hrsonline.org/*, April 14 2015. (page 71).

[139] M. Wissenberg, F. K. Lippert, F. Folke, P. Weeke, C. M. Hansen, E. F. Christensen, H. Jans, P. A. Hansen, T. Lang-Jensen, J. B. Olesen *et al.*, "Association of national initiatives to improve cardiac arrest management with rates of bystander intervention and patient survival after out-of-hospital cardiac arrest," *JAMA*, vol. 310, no. 13, pp. 1377–1384, 2013. (page 71).

[140] J.-T. Gräsner, J. Wnent, I. Gräsner, S. Seewald, M. Fischer, and T. Jantzen, "Einfluss der basisreanimationsmaßnahmen durch laien auf das überleben nach plötzlichem herztod," *Notfall+ Rettungsmedizin*, vol. 15, no. 7, pp. 593–599, 2012. (pages 71 and 87).

[141] M. Sasaki, T. Iwami, T. Kitamura, S. Nomoto, C. Nishiyama, T. Sakai, K. Tanigawa, K. Kajino, T. Irisawa, T. Nishiuchi *et al.*, "Incidence and outcome of out-of-hospital cardiac arrest with public-access defibrillation-a descriptive epidemiological study in a large urban community," *Circulation Journal*, vol. 75, no. 12, pp. 2821–2826, 2011. (page 71).

[142] J. Yeung, R. Meeks, D. Edelson, F. Gao, J. Soar, and G. D. Perkins, "The use of cpr feedback/prompt devices during training and cpr performance: a systematic review," *Resuscitation*, vol. 80, no. 7, pp. 743–751, 2009. (pages 71 and 72).

[143] C. Buléon, J.-J. Parienti, L. Halbout, X. Arrot, H. D. F. Régent, D. Chelarescu, J.-L. Fellahi, J.-L. Gérard, and J.-L. Hanouz, "Improvement in chest compression quality using a feedback device (cprmeter): a simulation randomized crossover study," *The American journal of emergency medicine*, vol. 31, no. 10, pp. 1457–1461, 2013. (page 71).

[144] D. M. González-Otero, J. Ruiz, S. Ruiz de Gauna, U. Irusta, U. Ayala, and E. Alonso, "A new method for feedback on the quality of chest compressions during cardiopulmonary resuscitation," *BioMed research international*, vol. 2014, 2014. (page 72).

[145] Y. Song, J. Oh, and Y. Chee, "A new chest compression depth feedback algorithm for high-quality cpr based on smartphone," *Telemedicine and e-Health*, 2014. (page 72).

[146] T. Chan, K. Wan, J. Chan, H. Lam, Y. Wong, P. Kan *et al.*, "New era of cpr: application of i-technology in resuscitation," *Hong Kong Journal of Emergency Medicine*, vol. 19, no. 5, p. 305, 2012. (page 72).

[147] L. Herzler, "Medical emergency app takes top prize at weekend hackathon," *Philadelphia Business Journal*, *http://www.bizjournals.com/*, April 14 2015. [Online]. Available: http://www.bizjournals.com/ (page 72).

[148] R. Sigrist, G. Rauter, R. Riener, and P. Wolf, "Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review," *Psychonomic bulletin & review*, vol. 20, no. 1, pp. 21–53, 2013. (page 72).

[149] K. Huang, E. Y.-L. Do, and T. Starner, "Pianotouch: A wearable haptic piano instruction system for passive learning of piano skills," in *Proceedings of ISWC 2008*. IEEE, 2008, pp. 41–44. (page 72).

[150] J. Kuhn, P. Lukowicz, M. Hirth, A. Poxrucker, J. Weppner, and J. Younas, "gphysics-using smart glasses for head-centered, context-aware learning in physics experiments." in *TLT*, vol. 9, no. 4, 2016, pp. 304–317. (pages 72 and 112).

[151] P. Lukowicz, A. Poxrucker, J. Weppner, B. Bischke, J. Kuhn, and M. Hirth, "Glass-physics: using google glass to support high school physics experiments," in *Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 151–154. (pages 72 and 112).

[152] M. Hassan, F. Daiber, F. Wiehr, F. Kosmalla, and A. Krüger, "Footstriker: An ems-based foot strike assistant for running," *Proceedings of the ACM on IMWUT*, vol. 1, no. 1, p. 2, 2017. (page 72).

[153] S. E. Rasmussen, M. A. Nebsbjerg, L. Q. Krogh, K. Bjørnshave, K. Krogh, J. A. Povlsen, I. S. Riddervold, T. Grøfte, H. Kirkegaard, and B. Løfgren, "A novel protocol for dispatcher assisted cpr improves cpr quality and motivation among rescuers—a randomized controlled simulation study," *Resuscitation*, vol. 110, pp. 74–80, 2017. (page 72).

[154] A. Nord, L. Svensson, H. Hult, S. Kreitz-Sandberg, and L. Nilsson, "Effect of mobile app-based vs. dvd- based cpr train. on students' practical cpr skills and will. to act: a clust. rand. study," *BMJ open*, vol. 6, no. 4, 2016. (page 72).

[155] J. S. You, H. S. Chung, S. P. Chung, J. W. Park, and D. G. Son, "Qr code: use of a novel mobile application to improve performance and perception of cpr in public," *Resuscitation*, vol. 84, no. 9, pp. e129–e130, 2013. (page 72).

[156] F. Semeraro, F. Taggi, G. Tammaro, G. Imbriaco, L. Marchetti, and E. L. Cerchiari, "icpr: a new application of high-quality cardiopulmonary resuscitation training," *Resuscitation*, vol. 82, no. 4, pp. 436–441, 2011. (page 72).

[157] N. P. Alonso, M. P. Rios, et al., and J. L. Velasco, "Randomised clinical simulation designed to evaluate the effect of telemed. using google glass on cardiop. resusci. (cpr)," *Emerg Med J*, vol. 34, no. 11, pp. 734–738, 2017. (page 72).

[158] J.-C. Wang, S.-H. Tsai, et al., and W.-I. Liao, "Kinect-based real-time audiovisual feedback device improves cardiop. resusci. quality of lower-body-weight rescuers," *American journal of emerg. med.*, 2017. (page 72).

[159] F. Salvetti, B. Bertagni, P. Ingrassia, and G. Pratticò, "Hololens, augmented reality and teamwork: Merging virtual and real workplaces," *International Journal of Advanced Corporate Learning (iJAC)*, vol. 11, no. 1, pp. 44–47, 2018. (page 72).

[160] A. H. Association, "The history of cpr," 2017, available: http://cpr.heart.org/ About CPR First Aid, History of CPR, [December-2017]. (page 73).

[161] J. P. Nolan, J. Soar, D. A. Zideman, D. Biarent, L. L. Bossaert, C. Deakin, R. W. Koster, J. Wyllie, and B. Böttiger, "European resuscitation council guidelines for resuscitation 2010 section 1. executive summary," *Resuscitation*, vol. 81, no. 10, pp. 1219–1276, 2010. (pages 73 and 94).

[162] R. A. Berg, R. Hemphill, B. S. Abella, T. P. Aufderheide, D. M. Cave, M. F. Hazinski, E. B. Lerner, T. D. Rea, M. R. Sayre, and R. A. Swor, "Part 5: Adult basic life support 2010 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care," *Circulation*, vol. 122, no. 18 suppl 3, pp. S685–S705, 2010. (page 73).

[163] R. W. Koster, M. A. Baubin, L. L. Bossaert, A. Caballero, P. Cassan, M. Castrén, C. Granja, A. J. Handley, K. G. Monsieurs, G. D. Perkins *et al.*, "European resuscitation council guidelines for resuscitation 2010 section 2. adult basic life support and use of automated external defibrillators," *Resuscitation*, vol. 81, no. 10, pp. 1277–1292, 2010. (page 73).

[164] A. H. Idris, "The sweet spot: Chest compressions between 100–120/minute optimize successful resuscitation from cardiac rest," *JEMS: a journal of emergency medical services*, vol. 37, no. 9, p. 4, 2012. (page 73).

[165] L. Wik, J. Kramer-Johansen, H. Myklebust, H. Sørebø, L. Svensson, B. Fellows, and P. A. Steen, "Quality of cardiopulmonary resuscitation during out-of-hospital cardiac arrest," *Jama*, vol. 293, no. 3, pp. 299–304, 2005. (page 73).

[166] D. P. Edelson, B. S. Abella, J. Kramer-Johansen, L. Wik, H. Myklebust, A. M. Barry, R. M. Merchant, T. L. V. Hoek, P. A. Steen, and L. B. Becker, "Effects of compression depth and pre-shock pauses predict defibrillation failure during cardiac arrest," *Resuscitation*, vol. 71, no. 2, pp. 137–145, 2006. (page 73).

[167] I. G. Stiell, S. P. Brown, J. Christenson, S. Cheskes, G. Nichol, J. Powell, B. Bigham, L. J. Morrison, J. Larsen, E. Hess *et al.*, "What is the role of chest compression depth during out-of-hospital cardiac arrest resuscitation?" *Critical care medicine*, vol. 40, no. 4, p. 1192, 2012. (page 73).

[168] I. G. Stiell, S. P. Brown, G. Nichol, S. Cheskes, C. Vaillancourt, C. W. Callaway, L. J. Morrison, J. Christenson, T. P. Aufderheide, D. P. Davis *et al.*, "What is the optimal chest compression depth during out-of-hospital cardiac arrest resuscitation of adult patients?" *Circulation*, pp. CIRCULATIONAHA–114, 2014. (page 73).

[169] D. Gordon, J.-H. Hanne, M. Berchtold, T. Miyaki, and M. Beigl, "Recognizing group activities using wearable sensors," in *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 2011, pp. 350–361. (page 87).

[170] G. Singla, D. J. Cook, and M. Schmitter-Edgecombe, "Recognizing independent and joint activities among multiple residents in smart environments," *Journal of ambient intelligence and humanized computing*, vol. 1, no. 1, pp. 57–63, 2010. (page 87).

[171] A. Gruenerbl, G. Bahle, and P. Lukowicz, "Detecting spontaneous collaboration in dynamic group activities from noisy individual activity data," in *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 2017, pp. 279–284. (pages 88 and 90).

[172] S. Cameron, I. Rutherford, and K. Mountain, "Debating the use of work-based learning and interprofessional education in promoting collaborative practice in primary care: a discussion paper," *Quality in primary care*, 2012. (page 89).

[173] H. Barr, "Competent to collaborate: towards a competency-based model for interprofessional education," *Journal of interprofessional care*, vol. 12, no. 2, pp. 181–187, 1998. (page 89).

[174] J. I. Westbrook, C. Duffield, L. Li, and N. J. Creswick, "How much time do nurses have for patients? a longitudinal study quantifying hospital nurses' patterns of task time distribution and interactions with health professionals," *BMC Health Services Research*, vol. 11, no. 1, p. 319, 2011. (page 89).

[175] G. Bahle, A. Gruenerbl, and L. Paul, "From individual activity recognition to unscripted collaboration analysis." *submitted at: ACM Transactions on Intelligent Systems and Technology (TIST)*, p. 24, 2018. (pages 90, 101, 102, 103, and 134).

[176] J. Mayfield, "Controlling inference in plan recognition," *User Modeling and User-Adapted Interaction*, vol. 2, no. 1-2, pp. 55–82, 1992. (page 90).

[177] R. Wilensky, "Planning and understanding: A computational approach to human reasoning," 1983. (page 90).

[178] C. F. Schmidt, N. Sridharan, and J. L. Goodson, "The plan recognition problem: An intersection of psychology and artificial intelligence," *Artificial Intelligence*, vol. 11, no. 1-2, pp. 45–83, 1978. (page 90).

[179] D. J. Litman and J. F. Allen, "A plan recognition model for subdialogues in conversations," *Cognitive science*, vol. 11, no. 2, pp. 163–200, 1987. (page 90).

[180] H. A. Kautz, "A formal theory of plan recognition," Ph.D. dissertation, Bell Laboratories, 1987. (page 90).

[181] A. Sadilek and H. Kautz, "Modeling and reasoning about success, failure, and intent of multi-agent activities," in *Proc. of UbiComp*, 2010. (page 90).

[182] ——, "Location-based reasoning about complex multi-agent behavior," *Journal of Artificial Intelligence Research*, vol. 43, pp. 87–133, 2012. (page 90).

[183] H. H. Zhuo and L. Li, "Multi-agent plan recognition with partial team traces and plan libraries," in *IJCAI*, vol. 22, 2011, p. 484. (page 90).

[184] H. H. Zhuo, Q. Yang, and S. Kambhampati, "Action-model based multi-agent plan recognition," in *Advances in Neural Information Processing Systems*, 2012, pp. 368–376. (page 90).

[185] S. Saria and S. Mahadevan, "Probabilistic plan recognition in multiagent systems." in *ICAPS*, 2004, pp. 287–296. (page 90).

[186] H. H. Bui, "A general model for online probabilistic plan recognition," in *IJCAI*, vol. 3, 2003, pp. 1309–1315. (page 90).

[187] M. Tambe, "Towards flexible teamwork," *Journal of artificial intelligence research*, vol. 7, pp. 83–124, 1997. (page 90).

[188] B. Banerjee, L. Kraemer, and J. Lyle, "Multi-agent plan recognition: Formalization and algorithms." in *AAAI*, 2010. (pages 90 and 95).

[189] P. Stone and M. Veloso, "Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork," *Artificial Intelligence*, vol. 110, no. 2, pp. 241–273, 1999. (page 90).

[190] S. S. Intille and A. F. Bobick, "A framework for recognizing multi-agent action from visual evidence," *AAAI/IAAI*, vol. 99, pp. 518–525, 1999. (pages 90 and 95).

[191] D. Avrahami-Zilberbrand and G. Kaminka, "Towards dynamic tracking of multi-agents teams: An initial report," in *Proceedings of the AAAI Workshop on Plan, Activity, and Intent Recognition (PAIR)*, 2007, pp. 17–22. (page 91).

[192] R. E. Wray and J. E. Laird, "Variability in human behavior modeling for military simulations," in *In Proceedings of Behavior Representation in Modeling and Simulation Conference (BRIMS*. Citeseer, 2003. (page 91).

[193] E. G. Jones, B. Browning, M. B. Dias, B. Argall, M. Veloso, and A. Stentz, "Dynamically formed heterogeneous robot teams performing tightly-coordinated tasks," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 570–575. (page 91).

[194] K. L. Genter, N. Agmon, and P. Stone, "Role-based ad hoc teamwork," in *Plan, Activity, and Intent Recognition*, 2011. (pages 91 and 95).

[195] M. Bowling and P. McCracken, "Coordination and adaptation in impromptu teams," in *AAAI*, vol. 5, 2005, pp. 53–58. (page 91).

[196] P. Stone, G. A. Kaminka, S. Kraus, J. S. Rosenschein *et al.*, "Ad hoc autonomous agent teams: Collaboration without pre-coordination." in *AAAI*, 2010. (page 91).

[197] P. Stone, G. A. Kaminka, S. Kraus, J. S. Rosenschein, and N. Agmon, "Teaching and leading an ad hoc teammate: Collaboration without pre-coordination," *Artificial Intelligence*, vol. 203, pp. 35–65, 2013. (page 91).

[198] G. Sukthankar and K. Sycara, "Simultaneous team assignment and behavior recognition from spatio-temporal agent traces," in *AAAI*, vol. 6, 2006, pp. 716–721. (page 92).

[199] G. Bahle, A. Gruenerbl, P. Lukowicz, E. Bignotti, M. Zeni, and F. Giunchiglia, "Recognizing hospital care activities with a coat pocket worn," in *Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on*, 2014, pp. 175–181. (pages 92, 98, and 113).

[200] N. Kern, B. Schiele, H. Junker, P. Lukowicz, and G. Tröster, "Wearable sensing to annotate meeting recordings," *Personal and Ubiquitous Computing*, vol. 7, no. 5, pp. 263–274, 2003. (page 93).

[201] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, 2011. (page 93).

[202] M. Davis, S. King, N. Good, and R. Sarvas, "From context to content: leveraging context to infer media metadata," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 188–195. (page 93).

[203] T. Thim, N. Krarup, E. L. Grove, C. V. Rohde, and B. Løfgren, "Initial assessment and treatment with the airway, breathing, circulation, disability, exposure (abcde) approach," *Int J Gen Med*, vol. 5, pp. 117–121, 2012. (pages 94 and 95).

[204] G. Sukthankar and K. P. Sycara, "Hypothesis pruning and ranking for large plan recognition problems." in *AAAI*, vol. 8, 2008, pp. 998–1003. (page 95).

[205] J. A. Ward, G. Pirkl, P. Hevesi, and P. Lukowicz, "Towards recognising collaborative activities using multiple on-body sensors," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 221–224. (page 103).

[206] J. Ward, G. Pirkl, P. Hevesi, and P. Lukowicz, "Detecting physical collaborations in a group task using body-worn microphones and accelerometers," in *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 2017, pp. 268–273. (page 103).

[207] L. Krupp, A. Gruenerbl, G. Bahle, and L. Paul, "Towards automatic semantic models by extraction of relevant information from online text," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*. (page 105).

[208] P. Lally, C. H. Van Jaarsveld, H. W. Potts, and J. Wardle, "How are habits formed: Modelling habit formation in the real world," *European journal of social psychology*, vol. 40, no. 6, pp. 998–1009, 2010. (page 111).

[209] F. Gravenhorst, A. Muaremi, J. Bardram, A. Grünerbl, O. Mayora, G. Wurzer, M. Frost, V. Osmani, B. Arnrich, P. Lukowicz *et al.*, "Mobile phones as medical devices in mental disorder treatment: an overview," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 335–353, 2015. (page 112).

[210] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 42–50, 2008. (page 113).

Curriculum Vitae

Agnes Johanna Grünerbl

Nationality: **Austria**
Email: **agnes.gruenerbl@dfki.de**

Education

1988 - 1996 Compulsory School: Gries am Brenner, Austria
1996 - 2000 Gymnasium: KORG der Barmherzigen Schwestern an der Kettenbrücke, Innsbruck, Austria
May/June 2000 Matura (Abitur)
2000 - 2001 Study of Anglistics and Romanistics at the University of Innsbruck
2001 - 2006 Study of Biomedical Informatics at the University of Health Science, Medical Informatics and Technology Tyrol (UMIT)
November 2004 Graduation (Bachelor of Science - B.Sc.) at UMIT
September 2006 Graduation (Diplom Ingenieur, DI) at UMIT

Theses

Bachelor Thesis Erfassung von Gehbewegungen mit Wearable Sensoren
Diploma Thesis A 3D Common-Shape-Model of the Proximal Femur to analyze the Stage of Osteoporosis in CT-Images

Research and Experience

2004-2006 Student auxiliary worker in research at the Institute for Computersystms and Networks, UMIT.

2005-2007 Research Assistant at the Institute of Biomedical Image Analysis, UMIT:
Building a statistical model of the hip bone and proximal femur via image analysis tools.

2006-2008 Research Assistant at the Research Division for Pervasive Health, UMIT:
* Deploying long term observation studies of dementia patients in elderly care (Lichtprojekt).

2008-2012 Research Assistant at the Embedded Systems Lab, University of Passau:
* Deploying several studies in health care facilities using modern technology (UWB location tracking, smart phones, smart carpets, capacitive sensing) for location tracking and gait analysis (SensFloor), activity monitoring and detection of behavioral patterns (Allow), swallowing disorders, etc.

since 2012 Research Assistant at the Embedded Intelligence Research Group
at the Technical University of Kaiserslautern (TUK) and
German Research Center for Artificial Intelligence (DFKI):
* Facilitating life and treatment of bipolar patients and their psychiatrists with smartphones (MONARCA).
* Bridging the semantic gap and humans learning from machine (SmartSociety)
* Supporting CPR performance and training (SmartSociety, SmartNurse, iGroups)
* Determining of collaboration in unscripted ad-hoc groups (iGroups)

Industrial

since 2016-05-14: Cofounder and CEO of Animys GmbH